

# **Lecture 3. Variational Bayes**

**Introduction to Bayesian Statistical learning**

**24.03.2025-28.03.2025 Instructors: Alina Bazarova, Jose Robledo**

# Analytic Variational Bayes (slightly heavier on the math)

Formula for posterior distribution (reminder)

$$p(\theta | X) = \frac{p(\theta)p(X | \theta)}{\int_{\mathbb{R}} p(\theta)p(X | \theta)d\theta} = \frac{p(X, \theta)}{\int_{\mathbb{R}} p(\theta)p(X | \theta)d\theta}$$

## Analytic Variational Bayes (slightly heavier on the math)

Formula for posterior distribution (reminder)

$$p(\theta | X) = \frac{p(\theta)p(X | \theta)}{\int_{\mathbb{R}} p(\theta)p(X | \theta)d\theta} = \frac{p(X, \theta)}{\int_{\mathbb{R}} p(\theta)p(X | \theta)d\theta}$$

We already know that evaluating posterior analytically can be rather challenging.

## Analytic Variational Bayes (slightly heavier on the math)

Formula for posterior distribution (reminder)

$$p(\theta | X) = \frac{p(\theta)p(X | \theta)}{\int_{\mathbb{R}} p(\theta)p(X | \theta)d\theta} = \frac{p(X, \theta)}{\int_{\mathbb{R}} p(\theta)p(X | \theta)d\theta}$$

We already know that evaluating posterior analytically can be rather challenging.

MCMC is a sampling technique which we have considered previously: **we construct a Markov chain which converges to posterior distribution**

# Analytic Variational Bayes (slightly heavier on the math)

Formula for posterior distribution (reminder)

$$p(\theta | X) = \frac{p(\theta)p(X | \theta)}{\int_{\mathbb{R}} p(\theta)p(X | \theta)d\theta} = \frac{p(X, \theta)}{\int_{\mathbb{R}} p(\theta)p(X | \theta)d\theta}$$

We already know that evaluating posterior analytically can be rather challenging.

MCMC is a sampling technique which we have considered previously: **we construct a Markov chain which converges to posterior distribution**

The goal of Variational Bayes: approximate  $p(\theta | X)$  with a simpler distribution  $q(\theta)$

## Analytic Variational Bayes (slightly heavier on the math)

Formula for posterior distribution (reminder)

$$p(\theta | X) = \frac{p(\theta)p(X | \theta)}{\int_{\mathbb{R}} p(\theta)p(X | \theta)d\theta} = \frac{p(X, \theta)}{\int_{\mathbb{R}} p(\theta)p(X | \theta)d\theta}$$

We already know that evaluating posterior analytically can be rather challenging.

MCMC is a sampling technique which we have considered previously: **we construct a Markov chain which converges to posterior distribution**

**The goal of Variational Bayes: approximate  $p(\theta | X)$  with a simpler distribution  $q(\theta)$**

Seen this in Laplace approximation already!

# Analytic Variational Bayes (slightly heavier on the math)

Formula for posterior distribution (reminder)

$$p(\theta | X) = \frac{p(\theta)p(X | \theta)}{\int_{\mathbb{R}} p(\theta)p(X | \theta)d\theta} = \frac{p(X, \theta)}{\int_{\mathbb{R}} p(\theta)p(X | \theta)d\theta}$$

We already know that evaluating posterior analytically can be rather challenging.

MCMC is a sampling technique which we have considered previously: **we construct a Markov chain which converges to posterior distribution**

**The goal of Variational Bayes: approximate  $p(\theta | X)$  with a simpler distribution  $q(\theta)$**

Seen this in Laplace approximation already!

**Assume there is a distribution density function  $q(\theta)$  which is in turn parametrised by a series of hyper-parameters.**

# Free energy and Kullback-Leibler divergence

Then  $\log p(X) = \log \frac{p(X, \theta)}{p(\theta | X)}$  and taking into account  $\int q(\theta) d\theta = 1$ :

# Free energy and Kullback-Leibler divergence

Then  $\log p(X) = \log \frac{p(X, \theta)}{p(\theta | X)}$  and taking into account  $\int q(\theta) d\theta = 1$ :

$$\log p(X) = \log p(X) \int q(\theta) d\theta$$

# Free energy and Kullback-Leibler divergence

Then  $\log p(X) = \log \frac{p(X, \theta)}{p(\theta | X)}$  and taking into account  $\int q(\theta) d\theta = 1$ :

$$\log p(X) = \log p(X) \int q(\theta) d\theta = \int q(\theta) \log \frac{p(X, \theta)}{p(\theta | X)} d\theta$$

# Free energy and Kullback-Leibler divergence

Then  $\log p(X) = \log \frac{p(X, \theta)}{p(\theta | X)}$  and taking into account  $\int q(\theta) d\theta = 1$ :

$$\log p(X) = \log(X) \int q(\theta) d\theta = \int q(\theta) \log \frac{p(X, \theta)}{p(\theta | X)} d\theta = \int q(\theta) \log \left[ \frac{p(X, \theta)}{p(\theta | X)} \cdot \frac{q(\theta)}{q(\theta)} \right] d\theta$$

# Free energy and Kullback-Leibler divergence

Then  $\log p(X) = \log \frac{p(X, \theta)}{p(\theta | X)}$  and taking into account  $\int q(\theta) d\theta = 1$ :

$$\log p(X) = \log(X) \int q(\theta) d\theta = \int q(\theta) \log \frac{p(X, \theta)}{p(\theta | X)} d\theta = \int q(\theta) \log \left[ \frac{p(X, \theta)}{p(\theta | X)} \cdot \frac{q(\theta)}{q(\theta)} \right] d\theta = \int q(\theta) \log \frac{p(X, \theta)}{q(\theta)} d\theta + \int q(\theta) \log \frac{q(\theta)}{p(\theta | X)} d\theta$$

# Free energy and Kullback-Leibler divergence

Then  $\log p(X) = \log \frac{p(X, \theta)}{p(\theta | X)}$  and taking into account  $\int q(\theta) d\theta = 1$ :

$$\log p(X) = \log(X) \int q(\theta) d\theta = \int q(\theta) \log \frac{p(X, \theta)}{p(\theta | X)} d\theta = \int q(\theta) \log \left[ \frac{p(X, \theta)}{p(\theta | X)} \cdot \frac{q(\theta)}{q(\theta)} \right] d\theta = \int q(\theta) \log \frac{p(X, \theta)}{q(\theta)} d\theta + \int q(\theta) \log \frac{q(\theta)}{p(\theta | X)} d\theta$$

$$\log p(X) = \int q(\theta) \log \frac{p(X, \theta)}{q(\theta)} d\theta + \int q(\theta) \log \frac{q(\theta)}{p(\theta | X)} d\theta$$

# Free energy and Kullback-Leibler divergence

Then  $\log p(X) = \log \frac{p(X, \theta)}{p(\theta | X)}$  and taking the expectation with respect to  $q(\theta)$ :

$$\log p(X) = \log(X) \int q(\theta) d\theta = \int q(\theta) \log \frac{p(X, \theta)}{p(\theta | X)} d\theta = \int q(\theta) \log \left[ \frac{p(X, \theta)}{p(\theta | X)} \cdot \frac{q(\theta)}{q(\theta)} \right] d\theta = \int q(\theta) \log \frac{p(X, \theta)}{q(\theta)} d\theta + \int q(\theta) \log \frac{q(\theta)}{p(\theta | X)} d\theta$$

$$\log p(X) = \int q(\theta) \log \frac{p(X, \theta)}{q(\theta)} d\theta + \int q(\theta) \log \frac{q(\theta)}{p(\theta | X)} d\theta$$

$F$ , depends only on  $\theta$  (free energy)

# Free energy and Kullback-Leibler divergence

Then  $\log p(X) = \log \frac{p(X, \theta)}{p(\theta | X)}$  and taking the expectation with respect to  $q(\theta)$ :

$$\log p(X) = \log(X) \int q(\theta) d\theta = \int q(\theta) \log \frac{p(X, \theta)}{p(\theta | X)} d\theta = \int q(\theta) \log \left[ \frac{p(X, \theta)}{p(\theta | X)} \cdot \frac{q(\theta)}{q(\theta)} \right] d\theta = \int q(\theta) \log \frac{p(X, \theta)}{q(\theta)} d\theta + \int q(\theta) \log \frac{q(\theta)}{p(\theta | X)} d\theta$$

$$\log p(X) = \int q(\theta) \log \frac{p(X, \theta)}{q(\theta)} d\theta + \int q(\theta) \log \frac{q(\theta)}{p(\theta | X)} d\theta$$

$F$ , depends only on  $\theta$  (free energy)

Kullback-Leibler ( $KL$ ) divergence

# Free energy and Kullback-Leibler divergence

Then  $\log p(X) = \log \frac{p(X, \theta)}{p(\theta | X)}$  and taking the expectation with respect to  $q(\theta)$ :

$$\log p(X) = \log(X) \int q(\theta) d\theta = \int q(\theta) \log \frac{p(X, \theta)}{p(\theta | X)} d\theta = \int q(\theta) \log \left[ \frac{p(X, \theta)}{p(\theta | X)} \cdot \frac{q(\theta)}{q(\theta)} \right] d\theta = \int q(\theta) \log \frac{p(X, \theta)}{q(\theta)} d\theta + \int q(\theta) \log \frac{q(\theta)}{p(\theta | X)} d\theta$$

**constant!**

$$\log p(X) = \int q(\theta) \log \frac{p(X, \theta)}{q(\theta)} d\theta + \int q(\theta) \log \frac{q(\theta)}{p(\theta | X)} d\theta$$

$F$ , depends only on  $\theta$  (free energy)

Kullback-Leibler ( $KL$ ) divergence

Note, that  $KL$  divergence is always  $\geq 0$  and hence  $\log p(X) \geq \int q(\theta) \log \frac{p(X, \theta)}{q(\theta)} d\theta$

# Free energy and Kullback-Leibler divergence

Then  $\log p(X) = \log \frac{p(X, \theta)}{p(\theta | X)}$  and taking the expectation with respect to  $q(\theta)$ :

$$\log p(X) = \log(X) \int q(\theta) d\theta = \int q(\theta) \log \frac{p(X, \theta)}{p(\theta | X)} d\theta = \int q(\theta) \log \left[ \frac{p(X, \theta)}{p(\theta | X)} \cdot \frac{q(\theta)}{q(\theta)} \right] d\theta = \int q(\theta) \log \frac{p(X, \theta)}{q(\theta)} d\theta + \int q(\theta) \log \frac{q(\theta)}{p(\theta | X)} d\theta$$

**constant!**

$$\log p(X) = \int q(\theta) \log \frac{p(X, \theta)}{q(\theta)} d\theta + \int q(\theta) \log \frac{q(\theta)}{p(\theta | X)} d\theta$$

$F$ , depends only on  $\theta$  (free energy)

Kullback-Leibler ( $KL$ ) divergence

Note, that  $KL$  divergence is always  $\geq 0$  and hence  $\log p(X) \geq \int q(\theta) \log \frac{p(X, \theta)}{q(\theta)} d\theta$

Moreover,  $\int q(\theta) \log \frac{q(\theta)}{p(\theta | X)} d\theta = \int q(\theta) \log q(\theta) d\theta - \int q(\theta) \log p(\theta | X) d\theta$  **measure of how close  $q(\theta)$**

# Free energy and Kullback-Leibler divergence

**constant!**

$$\log p(X) = \int q(\theta) \log \frac{p(X, \theta)}{q(\theta)} d\theta + \int q(\theta) \log \frac{q(\theta)}{p(\theta | X)} d\theta$$

$F$ , depends only on  $\theta$  (free energy)

Kullback-Leibler ( $KL$ ) divergence

Hence **maximising free energy** is equivalent to **minimising KL divergence**

# Mean-field approximation

We assume a mean-field approximation for  $q(\theta)$ , namely  $q(\theta) = \prod_i q_{\theta_i}(\theta_i)$

# Mean-field approximation

We assume a mean-field approximation for  $q(\theta)$ , namely  $q(\theta) = \prod_i q_{\theta_i}(\theta_i)$

where  $\theta_i$  are separate non-intersecting groups of parameters with the corresponding distribution density functions  $q_{\theta_i}$ .

# Mean-field approximation

We assume a mean-field approximation for  $q(\theta)$ , namely  $q(\theta) = \prod_i q_{\theta_i}(\theta_i)$

where  $\theta_i$  are separate non-intersecting groups of parameters with the corresponding distribution density functions  $q_{\theta_i}$ .

The key property of  $q_{\theta_i}$ :

$$\log q(\theta_i) \propto \int q_{\theta_{-i}}(\theta_{-i}) p(X, \theta) d\theta_{-i} \quad q_{\theta_{-i}}(\theta_{-i}) = \prod_{j \neq i} q_{\theta_j}(\theta_j)$$

where index  $-i$  means that  $i$ th group of parameters is excluded

# Mean-field approximation

We assume a mean-field approximation for  $q(\theta)$ , namely  $q(\theta) = \prod_i q_{\theta_i}(\theta_i)$

where  $\theta_i$  are separate non-intersecting groups of parameters with the corresponding distribution density functions  $q_{\theta_i}$ .

The key property of  $q_{\theta_i}$ :

$$\log q(\theta_i) \propto \int q_{\theta_{-i}}(\theta_{-i}) p(X, \theta) d\theta_{-i} \quad q_{\theta_{-i}}(\theta_{-i}) = \prod_{j \neq i} q_{\theta_j}(\theta_j)$$

where index  $-i$  means that  $i$ th group of parameters is excluded.

The proof of the above stems from the calculus of variations.

# Sketch of the proof

We need to maximise free energy  $F = \int q(\theta) \log \frac{p(X, \theta)}{q(\theta)} d\theta$  with respect to each factorised  $q_{\theta_i}(\theta_i)$

# Sketch of the proof

We need to maximise free energy  $F = \int q(\theta) \log \frac{p(X, \theta)}{q(\theta)} d\theta$  with respect to each factorised  $q_{\theta_i}(\theta_i)$   $F = \int f(\theta, q(\theta)) d\theta$  is a function of a function (functional) hence **calculus of variations**





# Sketch of the proof

$$\frac{\partial}{\partial q_{\theta_i}(\theta_i)} \int q(\theta) \log \frac{p(X, \theta)}{q(\theta)} d\theta_{-i} = 0, \text{ recall } q(\theta) = \prod_i q(\theta_i)$$

use differentiation by parts

$$\int q_{\theta_{-i}}(\theta_{-i}) \log p(X, \theta) d\theta_{-i} - \int q_{\theta_{-i}}(\theta_{-i}) \log q(\theta_{-i}) d\theta_{-i} - \int q_{\theta_{-i}}(\theta_{-i}) \log q(\theta_i) d\theta_{-i} + \text{const} = 0$$

# Sketch of the proof

$$\frac{\partial}{\partial q_{\theta_i}(\theta_i)} \int q(\theta) \log \frac{p(X, \theta)}{q(\theta)} d\theta_{-i} = 0$$

use differentiation by parts

**constant**



$$\int q_{\theta_{-i}}(\theta_{-i}) \log p(X, \theta) d\theta_{-i} - \int q_{\theta_{-i}}(\theta_{-i}) \log q(\theta_{-i}) d\theta_{-i} - \int q_{\theta_{-i}}(\theta_{-i}) \log q(\theta_i) d\theta_{-i} + \text{const} = 0$$

Hence, given that  $\int q_{\theta_{-i}}(\theta_{-i}) d\theta_{-i} = 1$  we get

# Sketch of the proof

$$\frac{\partial}{\partial q_{\theta_i}(\theta_i)} \int q(\theta) \log \frac{p(X, \theta)}{q(\theta)} d\theta_{-i} = 0$$

use differentiation by parts

**constant**



$$\int q_{\theta_{-i}}(\theta_{-i}) \log p(X, \theta) d\theta_{-i} - \int q_{\theta_{-i}}(\theta_{-i}) \log q(\theta_{-i}) d\theta_{-i} - \int q_{\theta_{-i}}(\theta_{-i}) \log q(\theta_i) d\theta_{-i} + \text{const} = 0$$

Hence, given that  $\int q_{\theta_{-i}}(\theta_{-i}) d\theta_{-i} = 1$  we get

$$\log q(\theta_i) = \int q_{\theta_{-i}}(\theta_{-i}) \log p(X, \theta) d\theta_{-i} + \text{const}$$

# Sketch of the proof

$$\frac{\partial}{\partial q_{\theta_i}(\theta_i)} \int q(\theta) \log \frac{p(X, \theta)}{q(\theta)} d\theta_{-i} = 0$$

use differentiation by parts

**constant**

$$\int q_{\theta_{-i}}(\theta_{-i}) \log p(X, \theta) d\theta_{-i} - \int q_{\theta_{-i}}(\theta_{-i}) \log q(\theta_{-i}) d\theta_{-i} - \int q_{\theta_{-i}}(\theta_{-i}) \log q(\theta_i) d\theta_{-i} + \text{const} = 0$$

Hence, given that  $\int q_{\theta_{-i}}(\theta_{-i}) d\theta_{-i} = 1$  we get

$$\log q(\theta_i) = \int q_{\theta_{-i}}(\theta_{-i}) \log p(X, \theta) d\theta_{-i} + \text{const}$$

$$\log q(\theta_i) \propto \int q_{\theta_{-i}} \log p(X, \theta) d\theta_{-i} \blacksquare$$

# Algorithm (Mean field variational Bayes for 2 parameters $\theta_1, \theta_2$ )

1. Initialise  $q(\theta_1)$
2. Given  $q(\theta_1)$  update  $q(\theta_2)$  using  $\log q(\theta_2) \propto \int \log p(X, \theta) q(\theta_1) d\theta_1$
3. Given  $q(\theta_2)$  update  $q(\theta_1)$  using  $\log q(\theta_1) \propto \int \log p(X, \theta) q(\theta_2) d\theta_2$
4. Iterate until stopping condition is met.

## Example: a single Gaussian

Assume we draw measurements  $y = (y_1, \dots, y_n)$  from a Gaussian distribution with

mean  $\mu$  and precision  $\beta$ : 
$$P(y_i | \mu, \beta) = \left( \frac{\beta}{2\pi} \right)^{\frac{1}{2}} e^{-\frac{\beta}{2}(y_i - \mu)^2}$$

## Example: a single Gaussian

Assume we draw measurements  $y = (y_1, \dots, y_n)$  from a Gaussian distribution with

mean  $\mu$  and precision  $\beta$ :  $P(y_i | \mu, \beta) = \left( \frac{\beta}{2\pi} \right)^{\frac{1}{2}} e^{-\frac{\beta}{2}(y_i - \mu)^2}$

$$P(y | \mu, \beta) = \prod_i P(y_i | \mu, \beta) = \left( \frac{\beta}{2\pi} \right)^{\frac{n}{2}} e^{-\frac{\beta}{2} \sum_i (y_i - \mu)^2}$$











**Example: single Gaussian. Priors and likelihood. Update on  $\mu$**

Similarly choose conjugate priors for  $\mu \sim N(m_0, \nu_0)$  and  $\beta \sim Ga(b_0, c_0)$ .

**Example: single Gaussian. Priors and likelihood. Update on  $\mu$**

Similarly choose conjugate priors for  $\mu \sim N(m_0, \nu_0)$  and  $\beta \sim Ga(b_0, c_0)$ .

Recall that  $P(\mu, \beta | Y) \propto P(Y | \mu, \beta)P(\mu)P(\beta)$  and

## Example: single Gaussian. Priors and likelihood. Update on $\mu$

Similarly choose conjugate priors for  $\mu \sim N(m_0, \nu_0)$  and  $\beta \sim Ga(b_0, c_0)$ .

Recall that  $P(\mu, \beta | Y) \propto P(Y | \mu, \beta)P(\mu)P(\beta)$  and

$$P(y | \mu, \beta) = \left( \frac{\beta}{2\pi} \right)^{\frac{n}{2}} e^{-\frac{\beta}{2} \sum_i (y_i - \mu)^2}, \log P(\mu) = -\frac{(\mu - m_0)^2}{2\nu_0} + \text{const}\{\mu\}, \log P(\beta) = (c - 1)\log \beta - \frac{\beta}{b} + \text{const}\{\beta\}$$





**Example: single Gaussian. Priors and likelihood. Update on  $\mu$**

Similarly choose conjugate priors for  $\mu \sim N(m_0, \nu_0)$  and  $\beta \sim Ga(b_0, c_0)$ .

Recall that  $P(\mu, \beta | Y) \propto P(Y | \mu, \beta)P(\mu)P(\beta)$ , hence

$$L = \log P(\mu, \beta | Y) = \frac{N}{2}\beta - \frac{\beta}{2} \sum_n (y_n - \mu)^2 - \frac{(\mu - m_0)^2}{2\nu_0} + (c_0 - 1)\log \beta_0 - \frac{\beta_0}{b_0} + \text{const}\{\mu, \beta\}$$

$$\log q(\mu) \propto \int Lq(\beta)d\beta$$

**Example: single Gaussian. Priors and likelihood. Update on  $\mu$**

Similarly choose conjugate priors for  $\mu \sim N(m_0, \nu_0)$  and  $\beta \sim Ga(b_0, c_0)$ .

Recall that  $P(\mu, \beta | Y) \propto P(Y | \mu, \beta)P(\mu)P(\beta)$ , hence

$$L = \log P(\mu, \beta | Y) = \frac{N}{2}\beta - \frac{\beta}{2} \sum_n (y_n - \mu)^2 - \frac{(\mu - m_0)^2}{2\nu_0} + (c_0 - 1)\log \beta_0 - \frac{\beta_0}{b_0} + \text{const}\{\mu, \beta\}$$

$$\log q(\mu) \propto \int L q(\beta) d\beta = \int L Ga(\beta, m, \nu) d\beta$$















# Non-linear models and convergence issues

Assume our model follows the equation  $y = g(\theta) + \varepsilon$ , where  $g(\theta)$  is a non-linear function and  $\varepsilon$  is additive Gaussian noise.

# Non-linear models and convergence issues

Assume our model follows the equation  $y = g(\theta) + \varepsilon$ , where  $g(\theta)$  is a non-linear function and  $\varepsilon$  is additive Gaussian noise.

In this case  $g(\theta)$  is approximated with Taylor expansion at the mode of posterior distribution  $m$ :  $g(\theta) \approx g(m) + J(\theta - m)$ , where  $J$  is the Jacobian matrix



# Stochastic Variational Bayes

Recall that the problem we discussed previously is maximising free energy

$$F = \int q(\theta) \log \frac{p(X, \theta)}{q(\theta)} d\theta.$$

# Stochastic Variational Bayes

Recall that the problem we discussed previously is maximising free energy

$$F = \int q(\theta) \log \frac{p(X, \theta)}{q(\theta)} d\theta.$$

Stochastic VB uses **gradient descent** algorithm to directly maximise  $F$

# Stochastic Variational Bayes

Recall that the problem we discussed previously is maximising free energy

$$F = \int q(\theta) \log \frac{p(X, \theta)}{q(\theta)} d\theta.$$

Stochastic VB uses **gradient descent** algorithm to directly maximise  $F$

This will require us to compute gradient  $\nabla_{\phi} F = \nabla_{\phi} \left( \int q(\theta) \log \frac{p(X, \theta)}{q(\theta)} d\theta \right)$



































