

Lecture 3. Variational Bayes

Introduction to Bayesian Statistical learning

18.03.2024-22.03.2024 Instructors: Alina Bazarova, Oleg Filatov. Technical issues: Alexandre Strube

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 θ (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 θ (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 θ (free energy)

Kullback-Leibler (KL) divergence

Note, that KL divergence is always ≥ 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 (free energy) Kullback-Leibler (KL) divergence

Note, that KL divergence is always ≥ 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 $q(\theta)$ and $p(\theta | X)$ are**

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 θ (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 θ_i are separate non-intersecting groups of parameters with the corresponding distribution density functions q_{θ_i} .

Mean-field approximation

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

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

The key property of q_{θ_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 θ_i are separate non-intersecting groups of parameters with the corresponding distribution density functions q_{θ_i} .

The key property of q_{θ_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

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**

We require a maximum of F with respect to each group of parameters θ_i , hence we re-write a functional in terms of these parameters alone

$$F = \int f(\theta, q(\theta)) d\theta_{-i} = \int q(\theta) \log \frac{p(X, \theta)}{q(\theta)} d\theta_{-i}$$

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**

We require a maximum of F with respect to each group of parameters θ_i , hence we re-write a functional in terms of these parameters alone

$$F = \int f(\theta, q(\theta)) d\theta_{-i} = \int q(\theta) \log \frac{p(X, \theta)}{q(\theta)} d\theta_{-i}$$

From variational calculus this is equivalent to solving: $\frac{\partial}{\partial q_{\theta_i}(\theta_i)} \int q(\theta) \log \frac{p(X, \theta)}{q(\theta)} d\theta_{-i} = 0$

Sketch of the proof

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

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 θ_1, θ_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 μ and precision β :
$$P(y_i | \mu, \beta) = \left(\frac{\beta}{2\pi} \right)^{\frac{1}{2}} e^{-\frac{\beta}{2}(y_i - \mu)^2}$$

Example: a single Gaussian

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

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

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

Example: a single Gaussian

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

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

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

We factorise our approximate posterior as $q(\mu, \beta) = q(\mu)q(\beta)$ and use the conjugate prior exponential family:

Example: a single Gaussian

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

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

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

We factorise our approximate posterior as $q(\mu, \beta) = q(\mu)q(\beta)$ and use the conjugate prior exponential family:

$$q(\mu | m, \nu) = \frac{1}{\sqrt{2\pi\nu}} e^{-\frac{1}{2\nu}(\mu - m)^2} \sim N(m, \nu)$$

Example: a single Gaussian

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

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

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

We factorise our approximate posterior as $q(\mu, \beta) = q(\mu)q(\beta)$ and use the conjugate prior exponential family:

$$q(\mu | m, \nu) = \frac{1}{\sqrt{2\pi\nu}} e^{-\frac{1}{2\nu}(\mu - m)^2} \sim N(m, \nu) \qquad q(\beta | b, c) = \frac{1}{\Gamma(c)} \frac{\beta^{c-1}}{b^c} e^{-\frac{\beta}{b}} \sim Ga(b, c)$$

Example: a single Gaussian

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

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

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

We factorise our approximate posterior as $q(\mu, \beta) = q(\mu)q(\beta)$ and use the conjugate prior exponential family:

$$q(\mu | m, \nu) = \frac{1}{\sqrt{2\pi\nu}} e^{-\frac{1}{2\nu}(\mu - m)^2} \sim N(m, \nu)$$

$$q(\beta | b, c) = \frac{1}{\Gamma(c)} \frac{\beta^{c-1}}{b^c} e^{-\frac{\beta}{b}} \sim Ga(b, c)$$

$$\log q(\mu) = -\frac{(\mu - m)^2}{2\nu} + \text{const}\{\mu\}$$

Example: a single Gaussian

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

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

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

We factorise our approximate posterior as $q(\mu, \beta) = q(\mu)q(\beta)$ and use the conjugate prior exponential family:

$$q(\mu | m, \nu) = \frac{1}{\sqrt{2\pi\nu}} e^{-\frac{1}{2\nu}(\mu - m)^2} \sim N(m, \nu)$$

$$q(\beta | b, c) = \frac{1}{\Gamma(c)} \frac{\beta^{c-1}}{b^c} e^{-\frac{\beta}{b}} \sim Ga(b, c)$$

$$\log q(\mu) = -\frac{(\mu - m)^2}{2\nu} + \text{const}\{\mu\}$$

$$\log q(\beta) = (c - 1)\log \beta - \frac{\beta}{b} + \text{const}\{\beta\}$$

Example: single Gaussian. Priors and likelihood. Update on μ

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 μ

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 μ

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

$$P(y | \mu, \beta) = \left(\frac{\beta}{2\pi} \right)^{\frac{n}{2}} e^{-\frac{\beta}{2} \sum_i (y_i - \mu)^2}, \log P(\mu) = -\frac{(\mu - m_0)^2}{2\nu_0} + \text{const}\{\mu\}, \log P(\beta) = (c_0 - 1)\log \beta - \frac{\beta}{b_0} + \text{const}\{\beta\}$$

Example: single Gaussian. Priors and likelihood. Update on μ

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

$$P(y | \mu, \beta) = \left(\frac{\beta}{2\pi}\right)^{\frac{n}{2}} e^{-\frac{\beta}{2} \sum_i (y_i - \mu)^2}, \log P(\mu) = -\frac{(\mu - m_0)^2}{2\nu_0} + \text{const}\{\mu\}, \log P(\beta) = (c_0 - 1)\log \beta - \frac{\beta}{b_0} + \text{const}\{\beta\}$$

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

Example: single Gaussian. Priors and likelihood. Update on μ

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

$$P(y | \mu, \beta) = \left(\frac{\beta}{2\pi}\right)^{\frac{n}{2}} e^{-\frac{\beta}{2} \sum_i (y_i - \mu)^2}, \log P(\mu) = -\frac{(\mu - m_0)^2}{2\nu_0} + \text{const}\{\mu\}, \log P(\beta) = (c_0 - 1)\log \beta - \frac{\beta}{b_0} + \text{const}\{\beta\}$$

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

Example: single Gaussian. Priors and likelihood. Update on μ

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

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

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

Example: single Gaussian. Priors and likelihood. Update on μ

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

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

$$\log q(\mu) \propto \int L q(\beta) d\beta = \int L Ga(\beta, b, c) d\beta$$

Example: single Gaussian. Priors and likelihood. Update on μ

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

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

$$\log q(\mu) \propto \int L q(\beta) d\beta = \int L Ga(\beta, b, c) d\beta = -\frac{(\mu - m_0)^2}{2\nu_0} \int Ga(\beta, b, c) d\beta -$$

Example: single Gaussian. Priors and likelihood. Update on μ

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

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

$$\log q(\mu) \propto \int Lq(\beta)d\beta = \int LGa(\beta, b, c)d\beta = -\frac{(\mu - m_0)^2}{2\nu_0} \int Ga(\beta, b, c)d\beta - \frac{1}{2} \sum_n (y_n - \mu)^2 \int \beta Ga(\beta, b, c)d\beta + \text{const}\{\mu\}$$

Example: single Gaussian. Priors and likelihood. Update on μ

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

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

$$\log q(\mu) \propto \int L q(\beta) d\beta = \int L Ga(\beta, b, c) d\beta = -\frac{(\mu - m_0)^2}{2\nu_0} \int \underset{\mathbf{1}}{Ga(\beta, b, c) d\beta} - \frac{1}{2} \sum_n (y_n - \mu)^2 \int \underset{\text{expectation}}{\beta Ga(\beta, b, c) d\beta} + \text{const}\{\mu\}$$

$$\log q(\mu) = -\frac{(\mu - m_0)^2}{2\nu_0} - \frac{bc}{2} \sum_n (y_n - \mu)^2 + \text{const}\{\mu\}, \text{ integrating out the terms}$$

Example: single Gaussian. Priors and likelihood. Update on μ

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

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

$$\log q(\mu) \propto \int L q(\beta) d\beta = \int L Ga(\beta, b, c) d\beta = -\frac{(\mu - m_0)^2}{2\nu_0} \int \underset{\mathbf{1}}{Ga(\beta, b, c) d\beta} - \frac{1}{2} \sum_n (y_n - \mu)^2 \int \underset{\text{expectation}}{\beta Ga(\beta, b, c) d\beta} + \text{const}\{\mu\}$$

$$\log q(\mu) = -\frac{(\mu - m_0)^2}{2\nu_0} - \frac{bc}{2} \sum_n (y_n - \mu)^2 + \text{const}\{\mu\}, \text{ integrating out the terms}$$

Example: single Gaussian. Priors and likelihood. Update on μ

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

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

$$\log q(\mu) \propto \int Lq(\beta)d\beta = \int LGa(\beta, b, c)d\beta = -\frac{(\mu - m_0)^2}{2\nu_0} \int Ga(\beta, b, c)d\beta - \frac{1}{2} \sum_n (y_n - \mu)^2 \int \beta Ga(\beta, b, c)d\beta + \text{const}\{\mu\}$$

$$\log q(\mu) = -\frac{(\mu - m_0)^2}{2\nu_0} - \frac{bc}{2} \sum_n (y_n - \mu)^2 + \text{const}\{\mu\}, \text{ integrating out the terms}$$

$$\text{Which can be re-written as } \log q(\mu) = -\frac{1 + N\nu_0bc}{2\nu_0} \left(\mu - \frac{m_0 + \nu_0bcs_1}{1 + N\nu_0bc} \right)^2 + \text{const}\{\mu\}$$

Example: single Gaussian. Priors and likelihood. Update on μ

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

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

$$\log q(\mu) \propto \int L q(\beta) d\beta = \int L Ga(\beta, b, c) d\beta = -\frac{(\mu - m_0)^2}{2\nu_0} \int Ga(\beta, b, c) d\beta - \frac{1}{2} \sum_n (y_n - \mu)^2 \int \beta Ga(\beta, b, c) d\beta + \text{const}\{\mu\}$$

$$\log q(\mu) = -\frac{(\mu - m_0)^2}{2\nu_0} - \frac{bc}{2} \sum_n (y_n - \mu)^2 + \text{const}\{\mu\}, \text{ integrating out the terms}$$

$$\text{Which can be re-written as } \log q(\mu) = -\frac{1 + N\nu_0 bc}{2\nu_0} \left(\mu - \frac{m_0 + \nu_0 bcs_1}{1 + N\nu_0 bc} \right)^2 + \text{const}\{\mu\}$$

$$\text{Recall } q(\mu) \sim N(m, \nu) \text{ and hence } m = \frac{m_0 + \nu_0 bcs_1}{1 + N\nu_0 bc} \text{ and } \nu = \frac{\nu_0}{1 + N\nu_0 bc} \text{ where } s_1 = \sum_n y_n$$

Update on β

We apply a similar procedure to derive an update on β .

$$\log q(\beta) = \int Lq(\mu)d\mu = \int LN(\mu, m, \nu)d\mu = \left(\frac{N}{2} + c_0 - 1\right) \log \beta + \frac{\beta}{b_0} - \frac{\beta}{2} \int \sum_n (y_n - \mu)^2 N(\mu, m, \nu) d\mu + \text{const}\{\beta\}$$

$$\log q(\beta) = \left(\frac{N}{2} + c_0 - 1\right) \log \beta - \left(\frac{1}{b_0} + \frac{X}{2}\right) \beta, \text{ where } X \text{ is the integral above:}$$

$$X = \frac{1}{2} \int (s_2 - 2\mu s_1 + \mu^2) N(\mu, m, \nu) d\mu = \frac{1}{2} s_2 - 2s_1 m + N(m + \nu^2), \text{ where } s_2 = \sum_n y_n^2$$

$$\text{Hence, } \frac{1}{b} = \frac{1}{b_0} + \frac{X}{2} \text{ and } c = \frac{N}{2} + c_0.$$

We can now proceed in an iterative procedure (fix β , update μ and the other way round until necessary)!

Jupyter notebook avb_gaussian

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 ε 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 ε 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

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 ε 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

Convergence.

- Convergence of VB is guaranteed since it is a generalisation of Expectation Maximisation algorithm
- As soon as we use Taylor approximation, the theory breaks down, and convergence becomes more empirical: e.g. monitoring free energy F , stop when it reaches maximum

Stochastic Variational Bayes

Recall that the problem we discussed previously is maximising free energy

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

Stochastic Variational Bayes

Recall that the problem we discussed previously is maximising free energy

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

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

Stochastic Variational Bayes

Recall that the problem we discussed previously is maximising free energy

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

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

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

Stochastic Variational Bayes

Recall that the problem we discussed previously is maximising free energy

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

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

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

Where ϕ are the hyper-parameters of q

Stochastic Variational Bayes

Recall that the problem we discussed previously is maximising free energy

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

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

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

Where ϕ are the hyper-parameters of q

But first we need to estimate the integral, which can be done using Monte-Carlo simulations

Stochastic Variational Bayes

Recall that the problem we discussed previously is maximising free energy

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

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

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

Where ϕ are the hyper-parameters of q

But first we need to estimate the integral, which can be done using Monte-Carlo simulations

$$F \approx \frac{1}{L} \sum_l \log p(X, \theta^l) - \log \frac{q(\theta^l)}{p(\theta^l)}, \text{ where } \theta^l \text{ are drawn from } q(\theta)$$

$$\text{Moreover, } \nabla_{\phi} F \approx \frac{1}{L} \sum_l \nabla_{\phi} \left(\log p(X, \theta^l) - \log \frac{q(\theta^l)}{p(\theta^l)} \right)$$

How to choose L ?

- Large L will give a better approximation, but can be computationally heavy
- Small L will make gradients noisy

How to choose L ?

- Large L will give a better approximation, but can be computationally heavy
- Small L will make gradients noisy

To reduce the noise in the gradient use **reparametrization trick** to draw θ^l :

How to choose L ?

- Large L will give a better approximation, but can be computationally heavy
- Small L will make gradients noisy

To reduce the noise in the gradient use **reparametrization trick** to draw θ^l :

Deterministically generate θ^l from an independent randomly generated parameter ε .

How to choose L ?

- Large L will give a better approximation, but can be computationally heavy
- Small L will make gradients noisy

To reduce the noise in the gradient use **reparametrization trick** to draw θ^l :

Deterministically generate θ^l from an independent randomly generated parameter ε .

E.g. use **probability integral transform**: if ξ is a random variable with cdf F_ξ , $F_\xi \sim U[0,1]$.

How to choose L ?

- Large L will give a better approximation, but can be computationally heavy
- Small L will make gradients noisy

To reduce the noise in the gradient use **reparametrization trick** to draw θ^l :

Deterministically generate θ^l from an independent randomly generated parameter ε .

E.g. use **probability integral transform**: if ξ is a random variable with cdf F_ξ ,
 $F_\xi \sim U[0,1]$.

Conversly, $F_\xi^{-1}(U[0,1]) \sim \xi$

How to choose L ?

- Large L will give a better approximation, but can be computationally heavy
- Small L will make gradients noisy

To reduce the noise in the gradient use **reparametrization trick** to draw θ^l :

Deterministically generate θ^l from an independent randomly generated parameter ε .

E.g. use **probability integral transform**: if ξ is a random variable with cdf F_ξ , $F_\xi(\xi) \sim U[0,1]$.

Conversely, $F_\xi^{-1}(U[0,1]) \sim \xi$

Hence we can generate any random variable from a uniform one.

How to choose L ?

- Large L will give a better approximation, but can be computationally heavy
- Small L will make gradients noisy

To reduce the noise in the gradient use **reparametrization trick** to draw θ^l :

Deterministically generate θ^l from an independent randomly generated parameter ε .

E.g. use **probability integral transform**: if ξ is a random variable with cdf F_ξ , $F_\xi(\xi) \sim U[0,1]$.

Conversely, $F_\xi^{-1}(U[0,1]) \sim \xi$

Hence we can generate any random variable from a uniform one.

Can be even simpler: $q(\theta) \sim N(\theta; \mu, \sigma)$. Generate $\varepsilon \sim N(0,1)$, then $\theta = \mu + \sigma\varepsilon$

How to choose L ?

In practice even $L = 1$ can be sufficient, however we need to choose **gradient descent algorithm** which deals with **stochastic optimisation**, e.g. **Adam**

To improve computational efficiency use **mini-batches**: divide data into subsets and performing optimisation on each batch in turn.

Very common technique in the machine learning!

Example: fitting a Gaussian distribution

Assume we draw measurements $y = (y_1, \dots, y_n)$ from a Gaussian distribution with mean μ and precision β : $P(y \mid \mu, \beta) = \left(\frac{\beta}{2\pi}\right)^{\frac{n}{2}} e^{-\frac{\beta}{2} \sum (y_i - \mu)^2}$.

Example: fitting a Gaussian distribution

Assume we draw measurements $y = (y_1, \dots, y_n)$ from a Gaussian distribution with mean μ and precision β : $P(y | \mu, \beta) = \left(\frac{\beta}{2\pi}\right)^{\frac{n}{2}} e^{-\frac{\beta}{2} \sum (y_i - \mu)^2}$.

Here we are not restricted to conjugate priors, hence a prior for μ and β can be

$\begin{bmatrix} \mu \\ -\log(\beta) \end{bmatrix} \sim MVN(m_0, C_0)$, where MVN stands for multivariate normal

Example: fitting a Gaussian distribution

Assume we draw measurements $y = (y_1, \dots, y_n)$ from a Gaussian distribution with mean μ and precision β : $P(y | \mu, \beta) = \left(\frac{\beta}{2\pi}\right)^{\frac{n}{2}} e^{-\frac{\beta}{2} \sum (y_i - \mu)^2}$.

Here we are not restricted to conjugate priors, hence a prior for μ and β can be

$\begin{bmatrix} \mu \\ -\log(\beta) \end{bmatrix} \sim MVN(m_0, C_0)$, where MVN stands for multivariate normal

Similarly for the approximate posterior $q(\theta) = q\left(\begin{bmatrix} \mu \\ -\log(\beta) \end{bmatrix}\right) \sim MVN(m, C)$

Example: fitting a Gaussian distribution

Assume we draw measurements $y = (y_1, \dots, y_n)$ from a Gaussian distribution with mean μ and precision β : $P(y | \mu, \beta) = \left(\frac{\beta}{2\pi}\right)^{\frac{n}{2}} e^{-\frac{\beta}{2} \sum (y_i - \mu)^2}$.

Here we are not restricted to conjugate priors, hence a prior for μ and β can be

$\begin{bmatrix} \mu \\ -\log(\beta) \end{bmatrix} \sim MVN(m_0, C_0)$, where MVN stands for multivariate normal

Similarly for the approximate posterior $q(\theta) = q \left(\begin{bmatrix} \mu \\ -\log(\beta) \end{bmatrix} \right) \sim MVN(m, C)$

Recall $MVN(m, C)$ has a pdf function

$$p(x, m, C) = (2\pi)^{-n/2} |C|^{-1/2} \exp \left((x - m)^T C^{-1} (x - m) \right)$$

Free energy

$$\int q(\theta) \log \frac{p(\theta)p(y|\theta)}{q(\theta)} d\theta \approx - \int q(\theta) \log \frac{q(\theta)}{p(\theta)} d\theta + \frac{1}{L} \sum_l \log p(y|\theta^l)$$

Free energy

$$\int q(\theta) \log \frac{p(\theta)p(y|\theta)}{q(\theta)} d\theta \approx - \int q(\theta) \log \frac{q(\theta)}{p(\theta)} d\theta + \frac{1}{L} \sum_l \log p(y|\theta^l)$$

$$\log \frac{q(\theta)}{p(\theta)} = -\frac{1}{2} \log \left(\frac{|C|}{|C_0|} \right) - \frac{1}{2} (\theta - m)^T C^{-1} (\theta - m) - \frac{1}{2} (\theta - m_0)^T C_0^{-1} (\theta - m_0)$$

Free energy

$$\int q(\theta) \log \frac{p(\theta)p(y|\theta)}{q(\theta)} d\theta \approx - \int q(\theta) \log \frac{q(\theta)}{p(\theta)} d\theta + \frac{1}{L} \sum_l \log p(y|\theta^l)$$

$$\log \frac{q(\theta)}{p(\theta)} = -\frac{1}{2} \log \left(\frac{|C|}{|C_0|} \right) - \frac{1}{2} (\theta - m)^T C^{-1} (\theta - m) - \frac{1}{2} (\theta - m_0)^T C_0^{-1} (\theta - m_0)$$

$$\int q(\theta) \log \frac{q(\theta)}{p(\theta)} d\theta = \frac{1}{2} \left(\text{Tr}(C_0^{-1} C) - \log \left(\frac{|C|}{|C_0|} \right) - N + (m - m_0)^T C_0^{-1} (m - m_0) \right)$$

Jupyter notebooks `svb_gaussian_tf2`, `svb_biexp_tf2`