Probabilistic Machine Learning

Variational Inference

Alexandre Graell i Amat
alexandre.graell@chalmers.se
https://sites.google.com/site/agraellamat

November 30, 2023

CHALMERS

Bayesian (probablistic) inference

In this course we consider problems of the form:

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D})}$$

- $\mathcal{D}$: Observed data
- $\boldsymbol{\theta}$: parameters of some model explaining the data

Goal: Find $p(\boldsymbol{\theta}|\mathcal{D})$.

- Can be found exactly in some cases (conjugate priors)
- Computation complexity can be alleviated when $p(\mathcal{D}, \boldsymbol{\theta})$ defined by specific classes of probabilistic graphical models (BNs, MRFs, FGs)

And when computing $p(\boldsymbol{\theta}|\mathcal{D})$ is intractable?

Approximate inference

Need to resort to approximations:

Stochastic methods:

- Monte Carlo approximation (numerical sampling)

Deterministic approximate inference methods:

- Variational inference
- Expectation propagation

Approximate inference

We will use Bishop's notation:

- $\boldsymbol{z}$: set of latent variables and parameters
- $\boldsymbol{x}$: set of observed variables

Given a probabilistic model that specifies $p(\boldsymbol{x}, \boldsymbol{z})$, we want to find an approximation of $p(\boldsymbol{z}|\boldsymbol{x})$ and $p(\boldsymbol{x})$.

Deterministic approximate inference

Idea: Approximate a complex posterior distribution $p(\boldsymbol{z}|\boldsymbol{x})$ by a tractable distribution $q(\boldsymbol{z}) \in \Omega$ that is close to $p(\boldsymbol{z}|\boldsymbol{x})$.

Ω: A tractable family of densities over latent variables $\boldsymbol{z}$

- Each $q(\boldsymbol{z}) \in \Omega$ is a candidate approximation to $p(\boldsymbol{z}|\boldsymbol{x})$

Goal: Find best candidate (closest to true posterior).

- Given definition of discrepancy between $q(\boldsymbol{z})$ and $p(\boldsymbol{z}|\boldsymbol{x})$, free parameters of $q(\boldsymbol{z})$ set by minimizing discrepancy

Kullback-Leibler divergence

Kullback-Leibler divergence:

$$\mathsf{KL}[p(\boldsymbol{x}) \parallel q(\boldsymbol{x})] = \int p(\boldsymbol{x}) \ln \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} \mathsf{d}\boldsymbol{x}$$
$$= - \int p(\boldsymbol{x}) \ln \frac{q(\boldsymbol{x})}{p(\boldsymbol{x})} \mathsf{d}\boldsymbol{x}$$

Properties:

1. $\mathsf{KL}[p(\boldsymbol{x}) \parallel q(\boldsymbol{x})] \geq 0$
2. $\mathsf{KL}[p(\boldsymbol{x}) \parallel q(\boldsymbol{x})] = 0$ if and only if $p(\boldsymbol{x}) = q(\boldsymbol{x})$
3. $\mathsf{KL}[p(\boldsymbol{x}) \parallel q(\boldsymbol{x})] \neq \mathsf{KL}[q(\boldsymbol{x}) \parallel p(\boldsymbol{x})]$

Idea: Find a tractable distribution $p(\boldsymbol{z}) \in \Omega$ that minimizes KL divergence.

Deterministic approximate inference

Two possibilities:

Variational inference: Minimize reverse KL divergence

$$q^*(\boldsymbol{z}) = \arg \min_{q(\boldsymbol{z}) \in \Omega} \mathsf{KL}[q(\boldsymbol{z}) \parallel p(\boldsymbol{z}|\boldsymbol{x})]$$

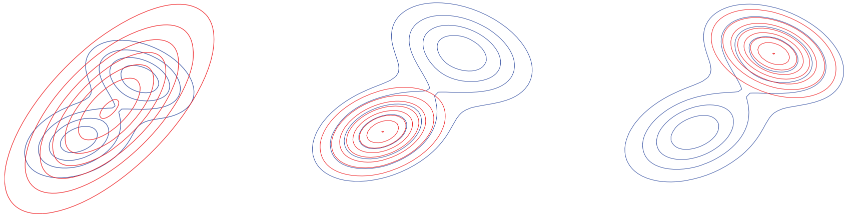
Expectation propagation: Minimize forward KL divergence

$$q^*(\boldsymbol{z}) = \arg \min_{q(\boldsymbol{z}) \in \Omega} \mathsf{KL}[p(\boldsymbol{z}|\boldsymbol{x}) \parallel q(\boldsymbol{z})]$$

An important application of VI: Variational autoencoders

D. P. Kingma, and M. Welling, "An Introduction to Variational Autoencoders," Foundations and Trends in Machine Learning, vol. 12, no. 4, pp. 307–392, 2019.

Deterministic approximate inference



- Blue: Bimodal distribution
- Red: Single Gaussian ($\Omega = \{\mathcal{N}(\mu, \sigma^2)\}$)
  - Left: $q^*(\boldsymbol{z}) = \arg\min_{q(\boldsymbol{z})\in\Omega} \mathsf{KL}[p(\boldsymbol{z}|\boldsymbol{x}) \parallel q(\boldsymbol{z})]$
  - Middle and right: $q^*(\boldsymbol{z}) = \arg\min_{q(\boldsymbol{z})\in\Omega} \mathsf{KL}[q(\boldsymbol{z}) \parallel p(\boldsymbol{z}|\boldsymbol{x})]$

Variational inference

Idea: Approximate $p(\boldsymbol{z}|\boldsymbol{x})$ with a tractable $q(\boldsymbol{z}) \in \Omega$ that minimizes

$$q^*(\boldsymbol{z}) = \arg \min_{q(\boldsymbol{z}) \in \Omega} \mathsf{KL}[q(\boldsymbol{z}) \parallel p(\boldsymbol{z}|\boldsymbol{x})]$$

- Not tractable! (requires knowledge of posterior $p(\boldsymbol{z}|\boldsymbol{x})$)

But can rewrite $\mathsf{KL}[q(\boldsymbol{z}) \parallel p(\boldsymbol{z}|\boldsymbol{x})]$ as

$$\begin{aligned}
\mathsf{KL}[q(\boldsymbol{z}) \parallel p(\boldsymbol{z}|\boldsymbol{x})] &= -\int q(\boldsymbol{z}) \ln \frac{p(\boldsymbol{z}|\boldsymbol{x})}{q(\boldsymbol{z})} \mathrm{d}\boldsymbol{z} \\
&= -\int q(\boldsymbol{z}) \ln \frac{p(\boldsymbol{x},\boldsymbol{z})}{q(\boldsymbol{z})p(\boldsymbol{x})} \mathrm{d}\boldsymbol{z} \\
&= \ln p(\boldsymbol{x}) - \underbrace{\int q(\boldsymbol{z}) \ln \frac{p(\boldsymbol{x},\boldsymbol{z})}{q(\boldsymbol{z})} \mathrm{d}\boldsymbol{z}}_{\mathcal{L}(q)} \\
&= \ln p(\boldsymbol{x}) + \mathsf{KL}[q(\boldsymbol{z}) \parallel p(\boldsymbol{x},\boldsymbol{z})]
\end{aligned}$$

Variational inference

It follows:

$$
\begin{aligned}
\ln p(\boldsymbol{x}) &= \ln \int p(\boldsymbol{x}, \boldsymbol{z}) \mathsf{d}\boldsymbol{z} \\
&= \ln \int q(\boldsymbol{z}) \frac{p(\boldsymbol{x}, \boldsymbol{z})}{q(\boldsymbol{z})} \mathsf{d}\boldsymbol{z} \\
&= \ln \left(\mathbb{E}_{q(\boldsymbol{z})} \left[\frac{p(\boldsymbol{x}, \boldsymbol{z})}{q(\boldsymbol{z})} \right] \right) \\
&\geq \mathbb{E}_{q(\boldsymbol{z})} \left[\ln \left(\frac{p(\boldsymbol{x}, \boldsymbol{z})}{q(\boldsymbol{z})} \right) \right] \\
&= \int q(\boldsymbol{z}) \ln \frac{p(\boldsymbol{x}, \boldsymbol{z})}{q(\boldsymbol{z})} \mathsf{d}\boldsymbol{z} \\
&\triangleq \mathcal{L}(q)
\end{aligned}
$$

$\mathcal{L}(q)$: A lower bound on $\ln p(\boldsymbol{x})$ (evidence lower bound (ELBO)).

Variational inference

$$\mathsf{KL}[q(\boldsymbol{z}) \parallel p(\boldsymbol{z}|\boldsymbol{x})] = \ln p(\boldsymbol{x}) - \mathcal{L}(q)$$

Thus, solving

$$q^*(\boldsymbol{z}) = \arg \min_{q(\boldsymbol{z}) \in \Omega} \mathsf{KL}[q(\boldsymbol{z}) \parallel p(\boldsymbol{z}|\boldsymbol{x})]$$

equivalent to solving

$$q^*(\boldsymbol{z}) = \arg \max_{q(\boldsymbol{z}) \in \Omega} \mathcal{L}(q) \triangleq \int q(\boldsymbol{z}) \ln \frac{p(\boldsymbol{x}, \boldsymbol{z})}{q(\boldsymbol{z})} \mathsf{d}\boldsymbol{z} = -\mathsf{KL}[q(\boldsymbol{z}) \parallel p(\boldsymbol{x}, \boldsymbol{z})]$$

With no restrictions on $q(\boldsymbol{z})$, $\mathcal{L}(q)$ maximized for $q(\boldsymbol{z}) = p(\boldsymbol{z}|\boldsymbol{x})$.

Variational inference

$$q^*(\boldsymbol{z}) = \arg\max_{q(\boldsymbol{z})\in\Omega} \mathcal{L}(q)$$
$$= \arg\min_{q(\boldsymbol{z})\in\Omega} \mathsf{KL}[q(\boldsymbol{z}) \parallel p(\boldsymbol{x}, \boldsymbol{z})]$$

In general intractable!

Idea: Choose a parametric distribution $q(\boldsymbol{z}|\boldsymbol{\omega})$ that is tractable, but rich enough to provide a good approximation of the true posterior.

- $\mathcal{L}(q)$ a function of $\boldsymbol{w}$ $\longrightarrow$ Can exploit standard nonlinear optimization techniques to determine optimal $\boldsymbol{w}$

Idea (2) (mean field variational inference): Restrict $q(\boldsymbol{z})$ so that factorizes as

$$q(\boldsymbol{z}) = \prod_{i=1}^{M} q_i(\boldsymbol{z}_i)$$

where $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_M$ are disjoint partitions of $\boldsymbol{z}$.

Mean field variational inference

$$q(\boldsymbol{z}) = \prod_{i=1}^{M} q_i(\boldsymbol{z}_i)$$

Goal: Solve optimization problem

$$\max_{q_1,\ldots,q_M} \mathcal{L}(q)$$

Amongst all $q(\boldsymbol{z}) = \prod_{i=1}^{M} q_i(\boldsymbol{z}_i)$, we want to find distribution with largest $\mathcal{L}(q)$.

- We will do optimization one term at a time

Mean field variational inference

Goal: Solve

$$q^*(\boldsymbol{z}) = \arg\max_{q(\boldsymbol{z})\in\Omega} \mathcal{L}(q) \triangleq \int q(\boldsymbol{z}) \ln \frac{p(\boldsymbol{x},\boldsymbol{z})}{q(\boldsymbol{z})} \mathrm{d}\boldsymbol{z}$$

with

$$q(\boldsymbol{z}) = \prod_{i=1}^{M} q_i(\boldsymbol{z}_i)$$

Singling out terms that involve q_j:

$$\begin{aligned}
\mathcal{L}(q) &= \int \prod_i q_i \left(\ln p(\boldsymbol{x},\boldsymbol{z}) - \sum_k \ln q_k \right) \mathrm{d}\boldsymbol{z} \\
&= \left(\int \prod_i q_i \ln p(\boldsymbol{x},\boldsymbol{z}) \mathrm{d}\boldsymbol{z} \right) - \left(\int \prod_i q_i \left(\sum_k \ln q_k \right) \mathrm{d}\boldsymbol{z} \right) \\
&= \left(\int \prod_i q_i \ln p(\boldsymbol{x},\boldsymbol{z}) \mathrm{d}\boldsymbol{z} \right) - \left(\int \prod_i q_i \ln q_j \mathrm{d}\boldsymbol{z} \right) - \left(\int \prod_i q_i \left(\sum_{k\neq j} \ln q_k \right) \mathrm{d}\boldsymbol{z} \right)
\end{aligned}$$

First term:

$$
\begin{aligned}
\int \prod_i q_i \ln p(\boldsymbol{x}, \boldsymbol{z}) \mathsf{d}\boldsymbol{z} &= \int q_j \left(\int \ln p(\boldsymbol{x}, \boldsymbol{z}) \prod_{i \neq j} q_i \mathsf{d}\boldsymbol{z}_i \right) \mathsf{d}\boldsymbol{z}_j \\
&= \int q_j \mathbb{E}_{\{\mathbf{z}_i\}_{i \neq j} \sim \prod_{i \neq j} q_i(\boldsymbol{z}_i)} \Big[\ln p(\boldsymbol{x}, \boldsymbol{z}) \Big] \mathsf{d}\boldsymbol{z}_j \\
&= \int q_j \mathbb{E}_{i \neq j} \Big[\ln p(\boldsymbol{x}, \boldsymbol{z}) \Big] \mathsf{d}\boldsymbol{z}_j
\end{aligned}
$$

Mean field variational inference

Second term:

$$\int \prod_i q_i \ln q_j \mathsf{d}\boldsymbol{z} = \int q_j \ln q_j \prod_{i \neq j} q_i \mathsf{d}\boldsymbol{z}_j \mathsf{d}\boldsymbol{z}_{i \neq j}$$

$$= \left(\int q_j \ln q_j \mathsf{d}\boldsymbol{z}_j \right) \left(\int \prod_{i \neq j} q_i \mathsf{d}\boldsymbol{z}_{i \neq j} \right)$$

$$= \int q_j \ln q_j \mathsf{d}\boldsymbol{z}_j$$

Mean field variational inference

Third term:

$$\begin{aligned}
\int \prod_i q_i \left(\sum_{k\neq j} \ln q_k \right) \mathsf{d}\boldsymbol{z} &= \int q_j \prod_{i\neq j} q_i \left(\sum_{k\neq j} \ln q_k \right) \mathsf{d}\boldsymbol{z}_j \mathsf{d}\boldsymbol{z}_{i\neq j} \\
&= \left(\int q_j \mathsf{d}\boldsymbol{z}_j \right) \left(\int \prod_{i\neq j} q_i \left(\sum_{k\neq j} \ln q_k \right) \mathsf{d}\boldsymbol{z}_{i\neq j} \right) \\
&= \int \prod_{i\neq j} q_i \left(\sum_{k\neq j} \ln q_k \right) \mathsf{d}\boldsymbol{z}_{i\neq j}
\end{aligned}$$

A constant that does not depend on $q(\boldsymbol{z}_j)$!

Mean field variational inference

Goal: Solve

$$q^*(\boldsymbol{z}) = \arg\max_{q(\boldsymbol{z})\in\Omega} \mathcal{L}(q) \triangleq \int q(\boldsymbol{z}) \ln \frac{p(\boldsymbol{x},\boldsymbol{z})}{q(\boldsymbol{z})} \mathsf{d}\boldsymbol{z}$$

with

$$q(\boldsymbol{z}) = \prod_{i=1}^{M} q_i(\boldsymbol{z}_i)$$

Singling out terms that involve q_j:

$$\begin{aligned}\mathcal{L}(q) &= \int q_j \mathbb{E}_{i\neq j}\Big[\ln p(\boldsymbol{x},\boldsymbol{z})\Big] \mathsf{d}\boldsymbol{z}_j - \int q_j \ln q_j \mathsf{d}\boldsymbol{z}_j + \text{const.}\\ &= \int q_j \ln \tilde{p}(\boldsymbol{x},\boldsymbol{z}_j) \mathsf{d}\boldsymbol{z}_j - \int q_j \ln q_j \mathsf{d}\boldsymbol{z}_j + \text{const.}\end{aligned}$$

with $\ln \tilde{p}(\boldsymbol{x},\boldsymbol{z}_j) = \mathbb{E}_{i\neq j}\Big[\ln p(\boldsymbol{x},\boldsymbol{z})\Big] + \text{const.}$

Mean field variational inference

$$\begin{aligned}
\mathcal{L}(q) &= \int q_j \ln \tilde{p}(\boldsymbol{x}, \boldsymbol{z}_j) \mathsf{d}\boldsymbol{z}_j - \int q_j \ln q_j \mathsf{d}\boldsymbol{z}_j + \text{const.} \\
&= \int q_j \ln \frac{\tilde{p}(\boldsymbol{x}, \boldsymbol{z}_j)}{q_j} \mathsf{d}\boldsymbol{z}_j + \text{const.} \\
&= -\mathsf{KL}[q_j(\boldsymbol{z}_j) \parallel \tilde{p}(\boldsymbol{x}, \boldsymbol{z}_j)] + \text{const.}
\end{aligned}$$

Keeping $q_{i \neq j}$ fixed, maximizing $\mathcal{L}(q)$ with respect to $q_j(\boldsymbol{z}_j)$, we obtain:

$$\begin{aligned}
q_j^*(\boldsymbol{z}_j) &= \arg\max_{q_j} \ \mathcal{L}(q) \\
&= \arg\max_{q_j} \ -\mathsf{KL}[q_j(\boldsymbol{z}_j) \parallel \tilde{p}(\boldsymbol{x}, \boldsymbol{z}_j)] + \text{const.} \\
&= \arg\min_{q_j} \ \mathsf{KL}[q_j(\boldsymbol{z}_j) \parallel \tilde{p}(\boldsymbol{x}, \boldsymbol{z}_j)] \\
&= \tilde{p}(\boldsymbol{x}, \boldsymbol{z}_j) \\
&= \exp\left(\mathbb{E}_{i \neq j}\left[\ln p(\boldsymbol{x}, \boldsymbol{z})\right] + \text{const.}\right)
\end{aligned}$$

Mean field variational inference

$$q_j^*(\boldsymbol{z}_j) = \exp\left(\mathbb{E}_{i\neq j}\left[\ln p(\boldsymbol{x},\boldsymbol{z})\right] + \text{const.}\right)$$

Equivalently,

$$\ln q_j^*(\boldsymbol{z}_j) = \mathbb{E}_{i\neq j}[\ln p(\boldsymbol{z},\boldsymbol{x})] + \text{const.}$$

$\ln q_j^*(\boldsymbol{z}_j)$ obtained by considering logarithm of $\ln p(\boldsymbol{z},\boldsymbol{x})$ and taking expectation with respect to $\{q_i\}_{i\neq j}$.

- Constant chosen so that q_j^* is a normalized distribution

Mean field variational inference

Goal: Solve optimization problem

$$\max_{q_1,\ldots,q_M} \mathcal{L}(q)$$

Algorithm:

1. Initialization: Set $\{q_i(\boldsymbol{z}_i)\}$
2. For $\ell = 1, \ldots, \ell_{\text{max}}$:
   - Fix $\{q_i(\boldsymbol{z}_i)\}_{i\neq j}$ to their last estimated values $q_i^*(\boldsymbol{z}_i)$
   - Update $q_j^*(\boldsymbol{z}_j)$ as

     $$q_j^*(\boldsymbol{z}_j) = \exp\left(\mathbb{E}_{\{q_i\}_{i\neq j}}[\ln p(\boldsymbol{z}, \boldsymbol{x})] + \text{const.}\right)$$

   - Normalize $q_j^*(\boldsymbol{z}_j)$
3. Repeat Step 2 until ELBO ($\mathcal{L}(q)$) converges

Mean field variational inference solves $\max_{q_1,\ldots,q_M} \mathcal{L}(q)$ iteratively for one hidden variable $\mathbf{z}_j$ at a time, while fixing $q_i(\boldsymbol{z}_i)$ for other latent variables $\{\mathbf{z}_i\}_{i\neq j}$.

Variational linear regression

Probabilistic model:

$$p(t_{\mathcal{D}}|\boldsymbol{w}) = \prod_{i=1}^{N} \mathcal{N}\left(t_i|\boldsymbol{w}^{\mathsf{T}}\boldsymbol{\phi}(\boldsymbol{x}_i), \beta^{-1}\right)$$

$$p(\boldsymbol{w}|\alpha) = \mathcal{N}(\boldsymbol{w}|0, \alpha^{-1}\boldsymbol{I})$$

with

$$\mathcal{N}\left(t_i|\boldsymbol{w}^{\mathsf{T}}\boldsymbol{\phi}(\boldsymbol{x}_i), \beta^{-1}\right) = \left(\frac{\beta}{2\pi}\right)^{1/2} \exp\left(-\frac{\beta}{2}\left(t_i - \boldsymbol{w}^{\mathsf{T}}\boldsymbol{\phi}(\boldsymbol{x}_i)\right)^2\right)$$

$$\mathcal{N}(\boldsymbol{w}|0, \alpha^{-1}\boldsymbol{I}) = \left(\frac{\alpha}{2\pi}\right)^{M/2} \exp\left(-\frac{\alpha}{2}\boldsymbol{w}^{\mathsf{T}}\boldsymbol{w}\right)$$

- We will assume β known

How do we pick α? $\longrightarrow$ Introduce prior, $p(\alpha) = \text{Gamma}(\alpha|a_0, b_0)$.

Variational linear regression

Probabilistic model:

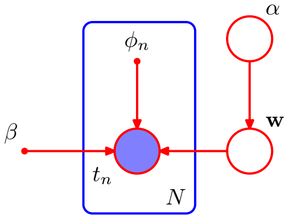
$$p(t_{\mathcal{D}}|\boldsymbol{w}) = \prod_{i=1}^{N} \mathcal{N}\left(t_i|\boldsymbol{w}^{\mathsf{T}}\boldsymbol{\phi}(\boldsymbol{x}_i), \beta^{-1}\right)$$

$$p(\boldsymbol{w}|\alpha) = \mathcal{N}(\boldsymbol{w}|0, \alpha^{-1}\boldsymbol{I}) = \left(\frac{\alpha}{2\pi}\right)^{M/2} \exp\left(-\frac{\alpha}{2}\boldsymbol{w}^{\mathsf{T}}\boldsymbol{w}\right)$$

$$p(\alpha) = \mathsf{Gamma}(\alpha|a_0, b_0)$$

Joint distribution:

$$p(t_{\mathcal{D}}, \boldsymbol{w}, \alpha) = p(t_{\mathcal{D}}|\boldsymbol{w})p(\boldsymbol{w}|\alpha)p(\alpha)$$



Variational linear regression

$$\ln p(t_{\mathcal{D}}|\boldsymbol{w}) = \sum_{i=1}^{N} \ln \left(\left(\frac{\beta}{2\pi} \right)^{1/2} \exp \left(-\frac{\beta}{2} \left(t_i - \boldsymbol{w}^{\mathsf{T}} \boldsymbol{\phi}(\boldsymbol{x}_i) \right)^2 \right) \right)$$

$$= -\frac{\beta}{2} \sum_{i=1}^{N} \left(t_i - \boldsymbol{w}^{\mathsf{T}} \boldsymbol{\phi}(\boldsymbol{x}_i) \right)^2 + \text{const.}$$

$$\ln p(\boldsymbol{w}|\alpha) = \frac{M}{2} \ln \alpha - \frac{\alpha}{2} \boldsymbol{w}^{\mathsf{T}} \boldsymbol{w} + \text{const.}$$

$$\ln p(\alpha) = (a_0 - 1) \ln \alpha - b_0 \alpha + \text{const.}$$

Predictive distribution (recall):

$$p(t|\mathcal{D}, \boldsymbol{x}, \beta) = \int p(\boldsymbol{w}|\mathcal{D}, \beta) p(t|\boldsymbol{x}, \boldsymbol{w}, \beta) d\boldsymbol{w}$$

Goal: Find an approximation of $p(\boldsymbol{w}, \alpha|\mathcal{D}, \beta) = p(\boldsymbol{w}, \alpha|\mathcal{D}) \longrightarrow$ Variational framework!

Variational linear regression

Goal: Find an approximation of $p(\boldsymbol{w}, \alpha|\mathcal{D}, \beta) = p(\boldsymbol{w}, \alpha|\mathcal{D}) \longrightarrow$ Variational framework!

We will consider a posterior $p(\boldsymbol{w}, \alpha|\mathcal{D}, \beta) \approx q(\boldsymbol{w}, \alpha)$ that factorizes as

$$q(\boldsymbol{w}, \alpha) = q(\boldsymbol{w})q(\alpha)$$

with $q(\boldsymbol{w}, \alpha) \equiv p(\boldsymbol{w}, \alpha|\mathcal{D})$, $q(\boldsymbol{w}) \equiv q(\boldsymbol{w}|\mathcal{D})$ and $q(\alpha) \equiv p(\alpha|\mathcal{D})$

Goal: Want to minimize ELBO.

- Recall: for each factor, we take the log of joint distribution, then average with respect to other variables

Variational linear regression

We need to iterate equations:

$$\ln q^*(\alpha) = \mathbb{E}_{q(\boldsymbol{w})}[\ln(p(t_{\mathcal{D}}, \boldsymbol{w}, \alpha))] + \text{const.}$$
$$\ln q^*(\boldsymbol{w}) = \mathbb{E}_{q(\alpha)}[\ln(p(t_{\mathcal{D}}, \boldsymbol{w}, \alpha))] + \text{const.}$$

with

$$p(t_{\mathcal{D}}, \boldsymbol{w}, \alpha) = p(t_{\mathcal{D}}|\boldsymbol{w})p(\boldsymbol{w}|\alpha)p(\alpha)$$

Optimum $q^*(\alpha)$:

$$\begin{aligned}
\ln q^*(\alpha) &= \mathbb{E}_{q(\boldsymbol{w})}[\ln(p(t_{\mathcal{D}}, \boldsymbol{w}, \alpha))] + \text{const.} \\
&= \mathbb{E}_{q(\boldsymbol{w})}[\ln(p(\boldsymbol{w}|\alpha)) + \ln(p(\alpha))] + \text{const.} \\
&= \ln p(\alpha) + \mathbb{E}_{q(\boldsymbol{w})}[\ln p(\boldsymbol{w}|\alpha)] + \text{const.} \\
&= (a_0 - 1)\ln\alpha - b_0\alpha + \frac{M}{2}\ln\alpha - \frac{\alpha}{2}\mathbb{E}_{q(\boldsymbol{w})}\left[\boldsymbol{w}^\mathsf{T}\boldsymbol{w}\right] + \text{const.}
\end{aligned}$$

$q^*(\alpha) = \text{Gamma}(\alpha|a_N, b_N)$, $a_N = a_0 + \frac{M}{2}$, $b_N = b_0 + \frac{1}{2}\mathbb{E}_{q(\boldsymbol{w})}[\boldsymbol{w}^\mathsf{T}\boldsymbol{w}]$

Variational linear regression

Optimum $q^*(\boldsymbol{w})$:

$$
\begin{aligned}
\ln q^*(\boldsymbol{w}) &= \mathbb{E}_{q(\alpha)}[\ln(p(t_{\mathcal{D}}, \boldsymbol{w}, \alpha))] + \text{const.} \\
&= \mathbb{E}_{q(\alpha)}[\ln(p(t_{\mathcal{D}}|\boldsymbol{w})) + \ln(p(\boldsymbol{w}|\alpha))] + \text{const.} \\
&= \ln(p(t_{\mathcal{D}}|\boldsymbol{w})) + \mathbb{E}_{q(\alpha)}[\ln p(\boldsymbol{w}|\alpha)] + \text{const.} \\
&= -\frac{\beta}{2}\sum_{i=1}^{N}\left(t_i - \boldsymbol{w}^\mathsf{T}\boldsymbol{\phi}(\boldsymbol{x}_i)\right)^2 - \frac{1}{2}\mathbb{E}_{q(\alpha)}[\alpha]\boldsymbol{w}^\mathsf{T}\boldsymbol{w} + \text{const.} \\
&= -\frac{1}{2}\boldsymbol{w}^\mathsf{T}\left(\mathbb{E}_{q(\alpha)}[\alpha]\boldsymbol{I} + \beta\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}\right)\boldsymbol{w} + \beta\boldsymbol{w}^\mathsf{T}\boldsymbol{\Phi}^\mathsf{T}t_{\mathcal{D}} + \text{const.}
\end{aligned}
$$

$q^*(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}_N, \boldsymbol{S}_N)$, $\boldsymbol{m}_N = \beta\boldsymbol{S}_N\boldsymbol{\Phi}^\mathsf{T}t_{\mathcal{D}}$, $\boldsymbol{S}_N = \left(\mathbb{E}_{q(\alpha)}[\alpha]\boldsymbol{I} + \beta\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}\right)^{-1}$

We get (see Bishop, Appendix B):

$$
\mathbb{E}_{q(\alpha)}[\alpha] = \frac{a_N}{b_N} \qquad \mathbb{E}_{q(\boldsymbol{w})}[\boldsymbol{w}\boldsymbol{w}^\mathsf{T}] = \boldsymbol{m}_N\boldsymbol{m}_N^\mathsf{T} + \boldsymbol{S}_N
$$

Variational linear regression

$$\mathbb{E}_{q(\alpha)}[\alpha] = \frac{a_N}{b_N} \qquad\qquad \mathbb{E}_{q(\boldsymbol{w})}[\boldsymbol{w}\boldsymbol{w}^\mathsf{T}] = \boldsymbol{m}_N\boldsymbol{m}_N^\mathsf{T} + \boldsymbol{S}_N$$

Algorithm:

1. Initialization: Set $q(\boldsymbol{w})$
2. For $\ell = 1, \ldots, \ell_{\mathsf{max}}$:
   - Compute
   $$\begin{aligned} a_N &= a_0 + \frac{M}{2} \\ b_N &= b_0 + \frac{1}{2}\mathbb{E}_{q(\boldsymbol{w})}[\boldsymbol{w}^\mathsf{T}\boldsymbol{w}] = b_0 + \frac{1}{2}\left(\boldsymbol{m}_N\boldsymbol{m}_N^\mathsf{T} + \boldsymbol{S}_N\right) \end{aligned}$$
   - Compute
   $$\begin{aligned} \boldsymbol{m}_N &= \beta\boldsymbol{S}_N\boldsymbol{\Phi}^\mathsf{T}t_{\mathcal{D}} \\ \boldsymbol{S}_N &= \left(\mathbb{E}_{q(\alpha)}[\alpha]\boldsymbol{I} + \beta\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}\right)^{-1} = \left(\frac{a_N}{b_N}\boldsymbol{I} + \beta\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}\right)^{-1} \end{aligned}$$
3. Repeat Step 2 until ELBO ($\mathcal{L}(q)$) converges

Variational linear regression

Predictive distribution:

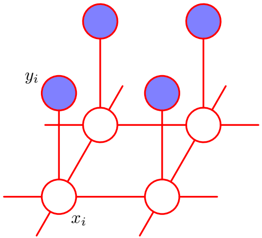
$$
\begin{aligned}
p(t|\mathcal{D},\boldsymbol{x},\beta) &= \int p(\boldsymbol{w}|\mathcal{D},\beta)p(t|\boldsymbol{x},\boldsymbol{w},\beta)d\boldsymbol{w} \\
&= \int p(\boldsymbol{w}|\mathcal{D})p(t|\boldsymbol{x},\boldsymbol{w})d\boldsymbol{w} \\
&\approx \int p(t|\boldsymbol{x},\boldsymbol{w})q(\boldsymbol{w})\mathrm{d}\boldsymbol{w} \\
&= \int \mathcal{N}\left(t|\boldsymbol{w}^{\mathsf{T}}\boldsymbol{\phi}(\boldsymbol{x}),\beta^{-1}\right)\mathcal{N}(\boldsymbol{w}|\boldsymbol{m}_N,\boldsymbol{S}_N)\mathrm{d}\boldsymbol{w} \\
&= \mathcal{N}\left(t|\boldsymbol{m}_N^{\mathsf{T}}\boldsymbol{\phi}(\boldsymbol{x}),\sigma^2(\boldsymbol{x})\right)\mathrm{d}\boldsymbol{w}
\end{aligned}
$$

with

$$
\sigma^2(\boldsymbol{x}) = \frac{1}{\beta} + \boldsymbol{\phi}(\boldsymbol{x})^{\mathsf{T}}\boldsymbol{S}_N\boldsymbol{\phi}(\boldsymbol{x})
$$

Solution takes same form as seen previously in class! (with fixed α)

Mean field variational inference for the Ising model



$$p(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{Z} \prod_{i,j} \psi_{i,j}(x_i, x_j) \prod_i \psi_i(x_i, y_i)$$

- $\mathsf{x}_i, \mathsf{y}_i \in \{+1, -1\}$ (Ising model)
- $\psi_i(x_i, y_i) = e^{\eta x_i y_i}$ and $\psi_{i,j}(x_i, x_j) = e^{\beta x_i x_j}$

For a Gaussian model $y_i = x_i + n_i$, with $\mathsf{n}_i \sim \mathcal{N}(0, \sigma^2)$,

$$\psi_{i,j}(x_i, y_i) = e^{-E(x_i, y_i)} = e^{L_i(x_i, y_i)}$$

Mean field variational inference for the Ising model

Goal:

$$\hat{\boldsymbol{x}} = \arg\max_{\boldsymbol{x}} p(\boldsymbol{x}|\boldsymbol{y})$$

with

$$\begin{aligned}
p(\boldsymbol{x}|\boldsymbol{y}) &= \frac{p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})}{p(\boldsymbol{y})} \\
&= \frac{1}{Z_1}\, p(\boldsymbol{x},\boldsymbol{y}) \\
&= \frac{1}{Z_2} \prod_i \psi_i(x_i, y_i) \prod_{i,j:\text{clique}} \psi_{i,j}(x_i, x_j) \\
&= \frac{1}{Z_2} \prod_i e^{L_i(x_i,y_i)} \prod_{i,j:\text{clique}} e^{\beta x_i x_j} \\
&= \frac{1}{Z_2}\, e^{\sum_{i,j:\text{clique}} \beta x_i x_j + \sum_i L_i(x_i,y_i)}
\end{aligned}$$

Mean field variational inference for the Ising model

Goal:

$$\hat{\boldsymbol{x}} = \arg\max_{\boldsymbol{x}} p(\boldsymbol{x}|\boldsymbol{y})$$

with

$$p(\boldsymbol{x}|\boldsymbol{y}) \propto \exp\left(\sum_{i,j:\text{clique}} \beta x_i x_j + \sum_i L_i(x_i, y_i)\right)$$

Idea: Approximate $p(\boldsymbol{x}|\boldsymbol{y})$ by a fully factorized approximation

$$q(\boldsymbol{x}) = \prod_i q(x_i, \mu_i), \quad \mu_i : \text{mean value of } \mathrm{x}_i$$

then apply mean field variational inference.

Mean field variational inference for the Ising model

Optimal factor $q_j^*(x_j)$:

$$\begin{aligned}
q_j^*(x_j) &= \exp\left(\mathbb{E}_{\{q_i\}_{i\neq j}}\left[\ln p(\boldsymbol{x},\boldsymbol{y})\right] + \text{const.}\right)\\
&= \frac{1}{Z}\exp\left(\mathbb{E}_{\{q_i\}_{i\neq j}}\left[\ln p(\boldsymbol{x},\boldsymbol{y})\right]\right)\\
&= \frac{1}{Z}\exp\left(\mathbb{E}_{\{q_i\}_{i\neq j}}\left[\sum_{i,j:\text{clique}}\beta x_i x_j + \sum_i L_i(x_i,y_i)\right]\right)
\end{aligned}$$

Need to consider only terms that involve x_j,

$$\begin{aligned}
q_j^*(x_j) &\propto \exp\left(\mathbb{E}_{\{q_i\}_{i\neq j}}\left[x_j\sum_{i\in\mathcal{N}(j)}\beta x_i + L_j(x_j,y_j)\right]\right)\\
&= \exp\left(x_j\sum_{i\in\mathcal{N}(j)}\beta\mathbb{E}_{q_i}\left[x_i\right] + L_j(x_j,y_j)\right)\\
&= \exp\left(x_j\sum_{i\in\mathcal{N}(j)}\beta\mu_i + L_j(x_j,y_j)\right)
\end{aligned}$$

Mean field variational inference for the Ising model

$$q_j^*(x_j) \propto \exp\left(x_j m_j + L_j(x_j, y_j)\right)$$

with $m_j \triangleq \sum_{i \in \mathcal{N}(j)} \beta \mu_i$

Define $L_j^+ \triangleq L_j(+1, y_j)$ and $L_j^- \triangleq L_j(-1, y_j)$

Approximate marginal posterior given by

$$\begin{aligned} q_j^*(\mathsf{x}_j = +1) &= \frac{\exp\left(m_j + L_j^+\right)}{\exp\left(m_j + L_j^+\right) + \exp\left(-m_j + L_j^-\right)} \\ &= \sigma(2a_j) \end{aligned}$$

with

$$a_j \triangleq m_j + 0.5(L_j^+ - L_j^-)$$

and

$$q_j^*(\mathsf{x}_j = -1) = \sigma(-2a_j)$$

Mean field variational inference for the Ising model

We can now compute new mean for x_j as

$$\begin{aligned}\mu_j = \mathbb{E}_{q_j}[x_j] &= q_j(\mathsf{x}_j = +1)\cdot(+1) + q_j(\mathsf{x}_j = -1)\cdot(-1)\\ &= \frac{1}{1+e^{-2a_j}} - \frac{1}{1+e^{2a_j}}\\ &= \tanh(a_j)\\ &= \tanh\left(\sum_{i\in\mathcal{N}(j)} \beta\mu_i + 0.5(L_j^+ - L_j^-)\right)\end{aligned}$$

We are done!

We can now update the parameters $\{\mu_j\}$ iteratively as

$$\mu_j^{(\ell)} = \tanh\left(\sum_{i\in\mathcal{N}(j)} \beta\mu_i^{(\ell-1)} + 0.5(L_j^+ - L_j^-)\right)$$

Mean field variational inference for the Ising model

Algorithm:

1. Initialization: Set $\{\mu_i^{(1)}\}$, e.g., to the noisy pixel values
2. For $\ell = 2, \ldots, \ell_{\text{max}}$:
   - Update $q_j^*(\mathsf{x}_j = +1)$ and $q_j^*(\mathsf{x}_j = -1)$ according to

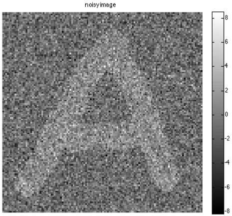
     $$q_j^*(\mathsf{x}_j = +1) = \sigma(2a_j) \quad \text{and} \quad q_j^*(\mathsf{x}_j = -1) = \sigma(-2a_j)$$

     using $\{\mu_i^{(\ell-1)}\}$ from previous iteration
   - Compute new mean values $\{\mu_i^{(\ell)}\}$ as

     $$\mu_j^{(\ell)} = \tanh\left(\sum_{i \in \mathcal{N}(j)} \beta\mu_i^{(\ell-1)} + 0.5(L_j^+ - L_j^-)\right)$$

3. Repeat Step 2 until convergence

Mean field variational inference for the Ising model

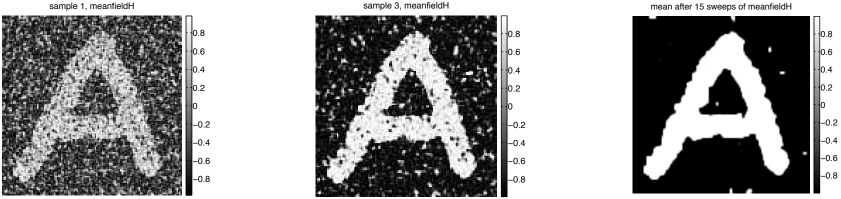


Model:

- $\beta_{i,j} = 1$, Gaussian model with $\sigma = 2$
- Parallel updates with $\lambda = 0.5$,

$$\mu_j^{(\ell)} = (1-\lambda)\mu_j^{(\ell-1)} + \lambda \tanh\left(\sum_{i\in\mathcal{N}(j)} \beta\mu_i^{(\ell-1)} + 0.5(L_j^+ - L_j^-)\right)$$

Mean field variational inference for the Ising model



Figures:

- Left: One iteration
- Center: Three iterations
- Right: Mean over 15 iterations

Deterministic approximate inference

Two possibilities:

Variational inference: Minimize

$$q^*(\boldsymbol{z}) = \arg\min_{q(\boldsymbol{z})\in\Omega} \mathsf{KL}[q(\boldsymbol{z}) \parallel p(\boldsymbol{z}|\boldsymbol{x})]$$

Expectation propagation: Minimize

$$q^*(\boldsymbol{z}) = \arg\min_{q(\boldsymbol{z})\in\Omega} \mathsf{KL}[p(\boldsymbol{z}|\boldsymbol{x}) \parallel q(\boldsymbol{z})]$$

Expectation propagation

Consider

$$p(\mathcal{D}, \boldsymbol{\theta}) = \prod_{i}^{I} f_i(\boldsymbol{\theta})$$

Goal: Evaluate $p(\boldsymbol{\theta}|\mathcal{D})$.

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{1}{p(\mathcal{D})} \prod_{i}^{I} f_i(\boldsymbol{\theta})$$

Idea: Approximate $p(\boldsymbol{\theta}|\mathcal{D})$ with a tractable distribution $q(\boldsymbol{z}) \in \Omega$,

$$q(\boldsymbol{\theta}) = \frac{1}{Z} \prod_{i}^{I} q_i(\boldsymbol{\theta})$$

Expectation propagation

Often assumed that factors come from exponential family, e.g.,

$$q(\boldsymbol{\theta}) = \frac{1}{Z} \prod_i^I \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$$

Find $q(\boldsymbol{\theta})$ which minimizes

$$q^*(\boldsymbol{\theta}) = \arg \min_{q(\boldsymbol{\theta}) \in \Omega} \mathsf{KL}[p(\boldsymbol{\theta}|\mathcal{D}) \parallel q(\boldsymbol{\theta})]$$

- Not tractable! (requires knowledge of posterior $p(\boldsymbol{\theta}|\mathcal{D})$)

Idea: Optimizing each factor in turn (keeping others constant)

1. Initializate factors $q_i(\boldsymbol{\theta})$
2. Until convergence, cycle through factors $q_j(\boldsymbol{\theta})$ and optimize as

$$q_j^\star(\boldsymbol{\theta}) = \arg \min_{q_j(\boldsymbol{\theta}) \in \Omega} \mathsf{KL}\left[\frac{1}{p(\mathcal{D})} f_j(\boldsymbol{\theta}) \prod_{i \neq j} q_i^\star(\boldsymbol{\theta}) \,\middle\|\, \frac{1}{Z} q_j(\boldsymbol{\theta}) \prod_{i \neq j} q_i^\star(\boldsymbol{\theta})\right]$$

Expectation propagation in practice

TrueSkill:

Microsoft's method to rank players of Xbox 360 Live online gaming system (one of the largest application of Bayesian statistics to date—processes over 105 games per day)

Reading

"Pattern recognition and machine learning,"

Chapter 10 (Intro, 10.1, 10.3 (not 10.3.3), 10.6 (optional), 10.7 (not 10.7.1))

Appendix A: Variational Bayes for a univariate Gaussian

Goal: Infer the posterior distribution for μ and τ for $\mathcal{N}(\mu, \tau^{-1})$ given $\mathcal{D} = \{x_1, \ldots, x_N\}$

Likelihood function:

$$p(\mathcal{D}|\mu,\tau) = \left(\frac{\tau}{2\pi}\right)^{N/2} \exp\left(-\frac{\tau}{2}\sum_{i=1}^{N}(x_i-\mu)^2\right)$$

We introduce conjugate priors:

$$p(\mu|\tau) = \mathcal{N}(\mu|\mu_0, (\lambda_0\tau)^{-1}) = \left(\frac{\lambda_0\tau}{2\pi}\right)^{1/2} \exp\left(-\frac{\lambda_0\tau}{2}(\mu-\mu_0)^2\right)$$

$$p(\tau) = \mathrm{Gamma}(\tau|a_0, b_0)$$

with $\ln\left(\mathrm{Gamma}(\tau|a_0,b_0)\right) = (a_0-1)\ln\tau - b_0\tau + \text{const.}$

Assume factorized variational approximation of posterior:

$$q(\mu,\tau) = q_\mu(\mu)q_\tau(\tau)$$

Appendix A: Variational Bayes for a univariate Gaussian

Goal: Finding optimum factors $q_\mu(\mu)$ and $q_\tau(\tau)$

We will use

$$\ln q_j^*(\boldsymbol{z}_j) = \mathbb{E}_{i\neq j}[\ln p(\boldsymbol{z}, \boldsymbol{x})] + \text{const.}$$

We proceed:

$$\begin{aligned} p(\mathcal{D}, \boldsymbol{\theta}) &= p(\mathcal{D}, \mu, \tau) \\ &= p(\mathcal{D}|\mu, \tau)p(\mu, \tau) \\ &= p(\mathcal{D}|\mu, \tau)p(\mu|\tau)p(\tau) \end{aligned}$$

We obtain:

$$\ln p(\mathcal{D}, \mu, \tau) = \ln p(\mathcal{D}|\mu, \tau) + \ln p(\mu|\tau) + \ln p(\tau)$$

$$= \frac{N}{2}\ln\tau - \frac{\tau}{2}\sum_{i=1}^{N}(x_i-\mu)^2 + \frac{1}{2}\ln\tau - \frac{\tau\lambda_0}{2}(\mu-\mu_0)^2 + (a_0-1)\ln\tau - b_0\tau + \text{const.}$$

Appendix A: Variational Bayes for a univariate Gaussian

We can now easily derive $q_\mu(\mu)$ and $q_\tau(\tau)$:

$q_\mu(\mu)$: Can focus on terms involving only μ

$$\begin{aligned}\ln q_\mu^\star(\mu) &= \mathbb{E}_{q(\tau)}[\ln p(\mathcal{D}|\mu,\tau) + \ln p(\mu|\tau)] + \text{const.}\\ &= -\frac{\mathbb{E}_{q(\tau)}(\tau)}{2}\left[\sum_{i=1}^{N}(x_i-\mu)^2 - \lambda_0(\mu-\mu_0)^2\right] + \text{const.}\end{aligned}$$

From this, $q_\mu^\star(\mu) = \mathcal{N}(\mu|\mu_N, \tau_N^{-1})$ with

$$\mu_N = \frac{\lambda_0\mu_0 + \sum_{i=0}^{N} x_i}{\lambda_0 + N} \qquad \tau_N = (\lambda_0 + N)\mathbb{E}_{q(\tau)}(\tau)$$

Appendix A: Variational Bayes for a univariate Gaussian

$q_\tau(\tau)$: Can focus on terms involving only τ

$$\ln q_\tau^\star(\tau) = \mathbb{E}_{q(\mu)}[\ln p(\mathcal{D}|\mu,\tau) + \ln p(\mu|\tau) + \ln p(\tau)] + \mathsf{const}$$

From this, $q_\tau^\star(\tau) = \mathsf{Gam}(a_N, b_N)$ with

$$a_N = a_0 + \frac{N+1}{2}$$

$$b_N = b_0 + \frac{1}{2}\mathbb{E}_{q(\mu)}\left[\sum_{i=1}^{N}(x_i-\mu)^2 + \lambda_0(\mu-\mu_0)^2\right]$$

Appendix A: Variational Bayes for a univariate Gaussian

$q^\star_\mu(\mu) = \mathcal{N}(\mu|\mu_N, \tau_N^{-1})$ with

$$\mu_N = \frac{\lambda_0\mu_0 + \sum_{i=0}^N x_i}{\lambda_0 + N} \qquad \tau_N = (\lambda_0 + N)\mathbb{E}_{q(\tau)}(\tau)$$

$q^\star_\tau(\tau) = \text{Gamma}(a_N, b_N)$ with

$$a_N = a_0 + \frac{N+1}{2} \qquad b_N = b_0 + \frac{1}{2}\mathbb{E}_{q(\mu)}\left[\sum_{i=1}^N (x_i - \mu)^2 + \lambda_0(\mu - \mu_0)^2\right]$$

Algorithm:

1. Initialization: Set $q_\tau(\tau)$
2. For $\ell = 1, \ldots, \ell_{\text{max}}$:
   - Fix $q_\tau(\tau)$ to its last estimated values $q^*_\tau(\tau)$
   - Update $q^*_\mu(\mu)$, i.e., update μ_N and τ_N
   - Fix $q_\mu(\mu)$ to its last estimated values $q^*_\mu(\mu)$
   - Update $q^*_\tau(\tau)$, i.e., update a_N and b_N
3. Repeat Step 2 until ELBO ($\mathcal{L}(q)$) converges

Appendix A: Variational Bayes for a univariate Gaussian