Variational Inference

Learning goals

Understand Variational Inference:

- KL divergence, ELBO (evidence lower bound), and their relationship
- Mean-field variation family

Be able to obtain the approximated posterior distribution using the coordinate ascent variational inference (CAVI):

- Derivation
- R implementation

Understand a trade-off between accuracy and efficiency when comparing MCMC vs Variational Inference.

Why Variational Inference (VI)?

We have looked at multiple computational techniques that could be used to approximate the posterior distribution or its properties (e.g., posterior mean and variance).

- Inversion methods
- Rejection sampling
- MCMC: Metropolis-Hastings algorithm, Gibbs sampling

However, there are problems for which we cannot easily use these sampling methods. These arise particularly when datasets are large or models are very complex.

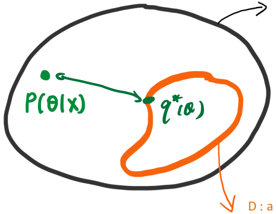
Variational Inference (VI)

- VI is known to be more computationally efficient than MCMC, at the cost of loss in accuracy.
- Compared to MCMC, VI tends to be faster and easier to apply to large data and complex models.

Variational Inference (VI)

VI can be used as an alternative approach to MCMC for Bayesian inference. It approximates the posterior distribution using an optimisation approach.

Given a family of (nice) probability distributions D, VI aims to find the member of D, $q^*(\boldsymbol{\theta})$ which is nearest to the exact posterior, $p(\boldsymbol{\theta}|\mathbf{y})$. Then, $q^*(\boldsymbol{\theta})$ serves as a proxy for the exact posterior.



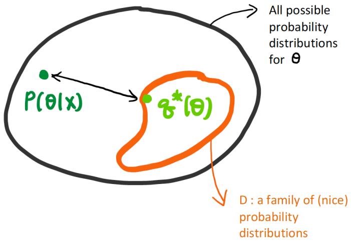
All possible probability distributions for θ

D : a family of (nice) probability distributions

Variational Inference (VI)

VI can be used as an alternative approach to MCMC for Bayesian inference. It approximates the posterior distribution using an optimisation approach.

Given a family of (nice) probability distributions D, VI aims to find the member of D, $q^*(\boldsymbol{\theta})$ which is nearest to the exact posterior, $p(\boldsymbol{\theta}|\mathbf{y})$. Then, $q^*(\boldsymbol{\theta})$ serves as a proxy for the exact posterior.



All possible probability distributions for θ

D : a family of (nice) probability distributions

Variational Inference (VI)

VI can be used as an alternative approach to MCMC for Bayesian inference. It approximates the posterior distribution using an optimisation approach.

Given a family of (nice) probability distributions D, VI aims to find the member of D, $q^*(\boldsymbol{\theta})$ which is nearest to the exact posterior, $p(\boldsymbol{\theta}|\boldsymbol{y})$. Then, $q^*(\boldsymbol{\theta})$ serves as a proxy for the exact posterior.

- Nearest?
- A family of (nice) probability distributions D?

Variational Inference (VI)

Given a family of (nice) probability distributions D, VI aims to find the member of D, $q^*(\boldsymbol{\theta})$ which is nearest to the exact posterior, $p(\boldsymbol{\theta}|\boldsymbol{y})$. Then, $q^*(\boldsymbol{\theta})$ serves as a proxy for the exact posterior.

- **Nearest?**
- A family of (nice) probability distributions D?

Given a family of probability distributions D, VI aims to find the member of D which minimises the Kullback-Leibler (KL) divergence to the exact posterior,

↓not symmetric

$$\begin{aligned} q^*(\boldsymbol{\theta}) &= \underset{q(\boldsymbol{\theta})\in D}{\arg\min} \; KL(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\mathbf{y})) \\ &= \underset{q(\boldsymbol{\theta})\in D}{\arg\min} \int_\theta q(\boldsymbol{\theta})[\log q(\boldsymbol{\theta}) - \log p(\boldsymbol{\theta}|\mathbf{y})]d\theta \\ &= \underset{q(\boldsymbol{\theta})\in D}{\arg\min} \; \mathbb{E}[\log q(\boldsymbol{\theta}) - \log p(\boldsymbol{\theta}|\mathbf{y})], \end{aligned}$$

where the expectation is taken with respect to $q(\boldsymbol{\theta})$.

KL divergence

$$KL(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\mathbf{y})) = \int_\theta q(\boldsymbol{\theta})[\log q(\boldsymbol{\theta}) - \log p(\boldsymbol{\theta}|\mathbf{y})]d\theta$$

We don't know $p(\boldsymbol{\theta}|\mathbf{y})$. How can we solve the optimisation problem?

$\frac{p(\theta,y)}{p(y)} = p(\theta|y).$

$$\begin{aligned} KL(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\mathbf{y})) &= \int_\theta q(\boldsymbol{\theta})[\log q(\boldsymbol{\theta}) - \underline{\log p(\boldsymbol{\theta}|\mathbf{y})}]d\theta \\ &= \int_\theta q(\boldsymbol{\theta})[\log q(\boldsymbol{\theta}) - \boxed{\log p(\boldsymbol{\theta},\mathbf{y}) + \log p(\mathbf{y})}]d\theta \\ &= \int_\theta q(\boldsymbol{\theta})[\log q(\boldsymbol{\theta}) - \log p(\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta})]d\theta + \log p(\mathbf{y}). \end{aligned}$$

$ELBO(q(\theta))$

$= \int_\theta q(\theta)[\log p(\theta,y) - \log(q(\theta)) d\theta$

$- \log q(\theta)]$

$= E_\theta[\log p(\theta,y) - \log q(\theta)]$

↓

with respect to $q(\theta)$

Let's define the evidence lower bound (ELBO) by

$$\text{ELBO}(q(\boldsymbol{\theta})) := -\int_\theta q(\boldsymbol{\theta})[\log q(\boldsymbol{\theta}) - \log p(\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta})]d\theta,$$

which does not require $p(\boldsymbol{\theta}|\mathbf{y})$ and is a tractable quantity to compute (provided that the family of distributions D is simple).

$$\begin{aligned} KL(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\boldsymbol{y})) &= \int_{\theta} q(\boldsymbol{\theta})[\log q(\boldsymbol{\theta}) - \log p(\boldsymbol{\theta}, \boldsymbol{y}) + \log p(\boldsymbol{y})]d\theta \\ &= -\text{ELBO}(q(\boldsymbol{\theta})) + \log p(\boldsymbol{y}). \end{aligned}$$

Because $\log p(\boldsymbol{y})$ does not depend on $q(\boldsymbol{\theta})$, minimising the KL divergence is equivalent to maximising the ELBO:

☆

$$q^*(\boldsymbol{\theta}) = \underbrace{\arg\min_{q(\boldsymbol{\theta})\in D} \quad KL(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\boldsymbol{y}))}_{\downarrow} = \underbrace{\arg\max_{q(\boldsymbol{\theta})\in D} \quad \text{ELBO}(q(\boldsymbol{\theta}))}_{\uparrow}.$$

since this optimisation requires posterior distribution

Remarks

this particular form is selected.

VI finds the member of the family of distributions D which maximise the ELBO. The ELBO is tractable quantity to compute provided that D is simple. This is possible because $KL(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\mathbf{y}))$ has been chosen to measure a distance between $q(\boldsymbol{\theta})$ and the exact posterior $p(\boldsymbol{\theta}|\mathbf{y})$.

Since $KL(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\mathbf{y})) \geq 0$ (for the proof, see the Section 1.6.1 in Christopher Bishop, Pattern Recognition and Machine Learning), $\log p(\mathbf{y}) \geq \text{ELBO}(q(\boldsymbol{\theta}))$. That's why it's called 'the evidence lower bound'.

$$\underbrace{KL(q(\theta) \| p(\theta|y))}_{\geq 0} = \underline{-ELBO(q(\theta)) + \log p(y)}$$

$$\log(p(y)) \geq ELBO(q(\theta))$$

Remarks

The KL divergence is not symmetric. Perhaps we can consider the following KL divergence to measure the distance:

$$
\begin{aligned}
q^*(\boldsymbol{\theta}) &= \underset{q(\boldsymbol{\theta})\in D}{\arg\min} \quad KL(p(\boldsymbol{\theta}|\mathbf{y})||q(\boldsymbol{\theta})) \\
&= \underset{q(\boldsymbol{\theta})\in D}{\arg\min} \quad \int_\theta \underline{p(\boldsymbol{\theta}|\mathbf{y})[\log p(\boldsymbol{\theta}|\mathbf{y}) - \log q(\boldsymbol{\theta})]d\theta.}
\end{aligned}
$$

this format of KL divergence.
w.r.p $p(\theta|y)$.

Expectation propagation (EP) achieves the approximation by minimising $KL(p(\boldsymbol{\theta}|\mathbf{y})||q(\boldsymbol{\theta}))$. We won't cover the EP and see the Section 10.7 in Christopher Bishop, Pattern Recognition and Machine Learning if you are interested.

Variational Inference (VI)

Given a family of (nice) probability distributions D, VI aims to find the member of D, $q^*(\boldsymbol{\theta})$ which is nearest to the exact posterior, $p(\boldsymbol{\theta}|\boldsymbol{y})$. Then, $q^*(\boldsymbol{\theta})$ serves as a proxy for the exact posterior.

- Nearest?
- **A family of (nice) probability distributions D?**

Mean-Field Variational Family

A common choice of D is the mean-field variational family, where the parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_J)$ are mutually independent in q.

VI is applied when parameters are in high dimension

A generic member of the mean-field variational family is

$$q(\boldsymbol{\theta}) = \prod_{j=1}^{J} q_j(\theta_j).$$

Each parameter θ_j is governed by its own variational factor, the probability distribution $q_j(\theta_j)$. In optimisation, these variational factors are chosen to maximise the ELBO.

A common choice of $q_j(\theta_j)$ is in the exponential family.

Mean-Field Variational Family: $q(\boldsymbol{\theta}) = \prod_{j=1}^{J} q_j(\theta_j)$

modelling assumption about $p(y|\theta)$ & $p(\theta)$ & $p(\theta|y)$.

This is not a modelling assumption. This is an assumption on what we approximate the model with.

Sometimes, VI with mean-field variational family provides a decent approximation to the exact posterior distribution. $p(\theta|y)$.

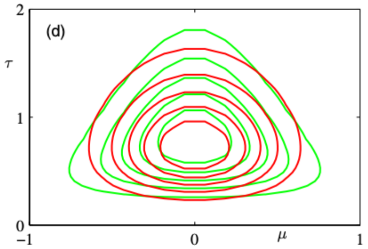
in the posterior space

$p(\mu,\tau|x)$

$\neq p(\mu|x)\cdot p(\tau|x)$

but VI: $q^*(\mu,\tau)$

$= q^*(\mu)\cdot q^*(\tau)$



Note: θ here are not independent in posterior space

Figure: Example taken from the Section 10.1.3 in Christopher Bishop, Pattern Recognition and Machine Learning. Contours of the exact posterior distribution $p(\mu,\tau|\boldsymbol{x})$ are shown in green. Contours of $q^*(\mu,\tau) = q^*_\mu(\mu)q^*_\tau(\tau)$ are shown in red.

Mean-Field Variational Family

$$q(\boldsymbol{\theta}) = \prod_{j=1}^{J} q_j(\theta_j).$$

Researchers have also studied more complex families (beyond the scope of the subject):

- Structured variational inference (add dependencies between parameters) : Saul and Jordan 1996; Barber and Wiegerinck 1999
- Mixtures of variational densities : Bishop et al. 1998

However, there is a trade-off. These families come with a more difficult-to-solve optimisation problem.

Variational Inference (VI)

Given a family of probability distributions D, VI aims to find the member of D which minimises the Kullback-Leibler (KL) divergence to the exact posterior,

$$q^*(\boldsymbol{\theta}) = \underset{q(\boldsymbol{\theta})\in D}{\arg\min} \quad KL(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\mathbf{y})).$$

A common choice of D is the mean-field variational family.

How to solve the optimisation problem?

Coordinate ascent variational inference (CAVI) algorithm

How to solve the optimisation problem?

Here I will describe the coordinate ascent variational inference (CAVI) which solves the optimisation problem for VI with the mean-field variational family.

Researchers have also developed other optimisation methods for VI (beyond the scope of the subject):

- Stochastic variational inference: Hoffman et al, 2013 → stochastic gradient descent
- Black box variational inference: Ranganath et al, 2014
- Automatic differentiation variational inference: Kucukelbir et al, 2017

For the mean-field variational family $D = \{q(\boldsymbol{\theta}) | q(\boldsymbol{\theta}) = \prod_{j=1}^{J} q_j(\theta_j)\}$,

$$q^*(\boldsymbol{\theta}) = \underset{q(\boldsymbol{\theta}) \in D}{\arg\min} \quad KL(q(\boldsymbol{\theta}) || p(\boldsymbol{\theta} | \mathbf{y})).$$

Let's say $\theta=(\mu,\tau)$

$p(\mu,\tau|y)$

$\approx q_\mu(\mu)q_\tau(\tau)$

The CAVI iteratively optimises each factor $q_j(\theta_j)$ as follows while holding the other factors fixed:

$$q_i^*(\theta_i) \propto exp\{\mathbb{E}_{-i}[\log p(\boldsymbol{\theta}, \mathbf{y})]\},$$

$q^*_\mu(\mu) \propto exp\{\int q_\tau(\tau) \log p(y|\mu,\tau)\cdot p(\mu,\tau)d\tau\}$

$= A(y,\mu)$

↓

some function of

where the expectation $\mathbb{E}_{-i}$ is taken with respect to the currently fixed variational factors, $\prod_{j \neq i}^{J} q_j(\theta_j)$. It goes uphill on the ELBO (why? See the Section 10.1.1 in Christopher Bishop, Pattern Recognition and Machine Learning), eventually finding a local optimum.

In practice, we stop the iteration when the ELBO has changed by less than some small number. The CAVI is sensitive to the initial values. We run the CVAI multiple times with different initial values and pick q^* with the highest ELBO.

$q^*_\tau(\tau) \propto exp\{\int q^*_\mu(\mu) \log(y|\mu,\tau)\cdot p(\mu,\tau)d\mu\}$

$\overset{*}{}$ → (alredy optimized μ).

$= B(y.\tau)$

Example

Model: we consider a normal model:

precision τ

$$x_i \sim \mathrm{N}(\mu, \frac{1}{\tau}) \quad \text{for} \quad i = 1, \ldots, n.$$

Prior: we impose the following prior for mean and precision parameters:

$$p(\mu, \tau) = p(\mu|\tau)p(\tau),$$

$$\mu|\tau \sim N(\mu_0, \frac{1}{\lambda_0 \tau}), \quad \tau \sim Gamma(a_0, b_0).$$

Posterior:

$$\underline{p(\mu, \tau|\boldsymbol{x}) = p(\mu|\tau, \boldsymbol{x})p(\tau|\boldsymbol{x}),}$$

μ & τ are not independent

in posterior space

$$\mu|\tau, \boldsymbol{x} \sim N(\frac{\lambda_0\mu_0 + n\bar{x}}{\lambda_0 + n}, \frac{1}{(\lambda_0 + n)\tau})$$

$$\tau|\boldsymbol{x} \sim Gamma(a_0 + \frac{n}{2}, b_0 + \frac{\sum_i (x_i - \bar{x})^2}{2} + \frac{n\lambda_0}{\lambda_0 + n}\frac{(\bar{x} - \mu_0)^2}{2})$$

See `https://en.wikipedia.org/wiki/Normal-gamma_distribution` for the derivation.

Model: we consider a normal model:

$$x_i \sim \mathrm{N}(\mu, \frac{1}{\tau}) \quad \text{for} \quad i = 1, \ldots, n.$$

Prior: we impose the following prior for mean and precision parameters:

$$p(\mu, \tau) = p(\mu|\tau)p(\tau),$$

$$\mu|\tau \sim N(\mu_0, \frac{1}{\lambda_0 \tau}), \quad \tau \sim Gamma(\overset{\text{shape parameter}}{a_0}, \overset{\text{weight parameter}}{b_0}).$$

VI with the mean-field variational family.:

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

CAVI:

- $q_\mu^*(\mu)$: the pdf of $N(\frac{\lambda_0\mu_0 + n\bar{x}}{\lambda_0 + n}, \frac{1}{(\lambda_0 + n)\mathbb{E}_\tau[\tau]})$
- $q_\tau^*(\tau)$: the pdf of $Gamma(\frac{n+1}{2} + a_0, b_0 + \frac{\sum_i \mathbb{E}_\mu[(x_i - \mu)^2]}{2} + \frac{\lambda_0 \mathbb{E}_\mu[(\mu - \mu_0)^2]}{2})$
- 'VI_normal.pdf' provides the derivation for $q_\mu^*(\mu)$, $q_\tau^*(\tau)$ and ELBO.

update $q^*_\mu(\mu)$
$q^*_\tau(\tau)$

$\Updownarrow$

update μ^*, σ^{2*}
a^*, b^*

$q^*_\mu(\mu)$: pdf of $N(\mu^*, \sigma^{2*})$, $q^*_\tau(\tau)$: pdf of $Gamma(a^*, b^*)$

Algorithm 1: CAVI for Normal distribution

```
Input: $x_1, \dots, x_n, \mu_0, \lambda_0, a_0, b_0$
Output: $q^*_\mu(\mu)$ (i.e., $\mu^*, \sigma^{2*}$) and $q^*_\tau(\tau)$ (i.e., $a^*, b^*$)
Initialisation: Initialise $\mu^*, \sigma^{2*}, a^*, b^*$
while ELBO has not converged do
    /* Update $\mu^*, \sigma^{2*}$                                   */
    $\mathbb{E}_\tau[\tau] = \frac{a^*}{b^*}$
    $\mu^* = \frac{\lambda_0\mu_0 + n\bar{x}}{\lambda_0 + n}$
    $\sigma^{2*} = \frac{1}{(\lambda_0+n)\mathbb{E}_\tau[\tau]} \to \frac{a^*}{b^*}$
    /* Update $a^*, b^*$                                             */
    for $i \leftarrow 1$ to $n$ do
        $\mathbb{E}_\mu[(x_i - \mu)^2] = \sigma^{2*} + (\mu^*)^2 - 2x_i\mu^* + x_i^2$
    end
    $\mathbb{E}_\mu[(\mu - \mu_0)^2] = \sigma^{2*} + (\mu^*)^2 - 2\mu_0\mu^* + \mu_0^2$
    $a^* = \frac{n+1}{2} + a_0$
    $b^* = b_0 + \frac{\sum_i \mathbb{E}_\mu[(x_i-\mu)^2]}{2} + \frac{\lambda_0\mathbb{E}_\mu[(\mu-\mu_0)^2]}{2}$
    Compute ELBO($\mu^*, \sigma^{2*}, a^*, b^*$).
end
return $q^*_\mu(\mu)$ (i.e., $\mu^*, \sigma^{2*}$) and $q^*_\tau(\tau)$ (i.e., $a^*, b^*$).
```

See VI.pdf for R implementation of this algorithm.

MCMC vs VI

MCMC is a tool for simulating from densities and VI is a tool for approximating densities.

- MCMC methods tend to be more computationally intensive than VI, but they also provide guarantees of producing (asymptotically) exact samples from the target density. as long as # iteration → ∞ asymptotically
- VI does not provide such guarantees (it can only find a density close to the target; VI generally underestimates the variance of the density), but it tends to be faster than MCMC.

VI is suited to large datasets and scenarios where we want to quickly explore many models.

Empirical comparison between MCMC and VI: Stegle et al (2010). A comparison of inference in sparse factor analysis

Learning goals

Understand Variational Inference:

- KL divergence, ELBO (evidence lower bound), and their relationship
- Mean-field variation family

Be able to obtain the approximated posterior distribution using the coordinate ascent variational inference (CAVI):

- Derivation
- R implementation

Understand a trade-off between accuracy and efficiency when comparing MCMC vs Variational Inference.