Variational Parameter Learning

Mark van der Wilk

Department of Computing
Imperial College London

@markvanderwilk
m.vdwilk@imperial.ac.uk

February 27, 2022

Recap: Variational Inference

- KL measures discrepancy between distributions

$$\mathrm{KL}[q(\mathbf{z})||p(\mathbf{z}\,|\,\mathbf{x})] \geqslant 0 \qquad \text{with equality iff } q(\mathbf{z}) = p(\mathbf{z}\,|\,\mathbf{x}) \tag{1}$$

- Find approx $q_{\mathbf{v}}(\mathbf{z}) \approx p(\mathbf{z}\,|\,\mathbf{x})$ by minimising KL divergence:

$$\mathbf{v}^* = \operatorname*{argmin}_{\mathbf{v}} \mathrm{KL}[q_{\mathbf{v}}(\mathbf{z})||p(\mathbf{z}\,|\,\mathbf{x})] \tag{2}$$

- Equivalent to maximising lower bound (ELBO) $\mathcal{L}$ since

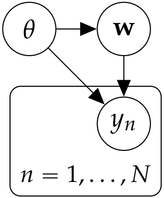
$$\mathrm{KL}[q_{\mathbf{v}}(\mathbf{z})||p(\mathbf{z}\,|\,\mathbf{x})] = \log p(\mathbf{x}) - \mathcal{L}(\mathbf{v}) \tag{3}$$

$$\implies \mathbf{v}^* = \operatorname*{argmax}_{\mathbf{v}} \mathcal{L}(\mathbf{v}) \tag{4}$$

Bayes for hyperparameters

Bayes' rule for everything:

$$
\begin{aligned}
p(\mathbf{w}, \theta \mid \mathbf{y}) &= \frac{p(\mathbf{y}, \mathbf{w}, \theta)}{p(\mathbf{y})} = \frac{p(\mathbf{y} \mid \mathbf{w}, \theta) p(\mathbf{w}|\theta) p(\theta)}{p(\mathbf{y})} && (5) \\
&= \underbrace{\frac{p(\mathbf{y} \mid \mathbf{w}, \theta) p(\mathbf{w}|\theta)}{p(\mathbf{y} \mid \theta)}}_{p(\mathbf{w}|\mathbf{y},\theta)} \underbrace{\frac{p(\mathbf{y} \mid \theta) p(\theta)}{p(\mathbf{y})}}_{p(\theta \mid \mathbf{y})} && (6)
\end{aligned}
$$

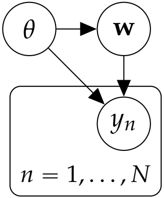


Bayes for hyperparameters

Bayes' rule for everything:

$$p(\mathbf{w}, \theta \,|\, \mathbf{y}) = \frac{p(\mathbf{y}, \mathbf{w}, \theta)}{p(\mathbf{y})} = \frac{p(\mathbf{y} \,|\, \mathbf{w}, \theta) p(\mathbf{w}|\theta) p(\theta)}{p(\mathbf{y})} \tag{5}$$

$$= \underbrace{\frac{p(\mathbf{y} \,|\, \mathbf{w}, \theta) p(\mathbf{w}|\theta)}{p(\mathbf{y} \,|\, \theta)}}_{p(\mathbf{w}|\mathbf{y},\theta)} \underbrace{\frac{p(\mathbf{y} \,|\, \theta) p(\theta)}{p(\mathbf{y})}}_{p(\theta \,|\, \mathbf{y})} \tag{6}$$



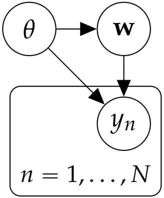
Posterior over f and θ consists of two parts

Bayes for hyperparameters

Bayes' rule for everything:

$$p(\mathbf{w}, \theta \,|\, \mathbf{y}) = \frac{p(\mathbf{y}, \mathbf{w}, \theta)}{p(\mathbf{y})} = \frac{p(\mathbf{y} \,|\, \mathbf{w}, \theta) p(\mathbf{w}|\theta) p(\theta)}{p(\mathbf{y})} \tag{5}$$

$$= \underbrace{\frac{p(\mathbf{y} \,|\, \mathbf{w}, \theta) p(\mathbf{w}|\theta)}{p(\mathbf{y} \,|\, \theta)}}_{p(\mathbf{w}|\mathbf{y},\theta)} \underbrace{\frac{p(\mathbf{y} \,|\, \theta) p(\theta)}{p(\mathbf{y})}}_{p(\theta \,|\, \mathbf{y})} \tag{6}$$



Posterior over f and θ consists of two parts

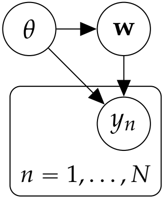
1. The original posterior over f,

Bayes for hyperparameters

Bayes' rule for everything:

$$p(\mathbf{w}, \theta \,|\, \mathbf{y}) = \frac{p(\mathbf{y}, \mathbf{w}, \theta)}{p(\mathbf{y})} = \frac{p(\mathbf{y} \,|\, \mathbf{w}, \theta) p(\mathbf{w}|\theta) p(\theta)}{p(\mathbf{y})} \tag{5}$$

$$= \underbrace{\frac{p(\mathbf{y} \,|\, \mathbf{w}, \theta) p(\mathbf{w}|\theta)}{p(\mathbf{y} \,|\, \theta)}}_{p(\mathbf{w}|\mathbf{y},\theta)} \underbrace{\frac{p(\mathbf{y} \,|\, \theta) p(\theta)}{p(\mathbf{y})}}_{p(\theta \,|\, \mathbf{y})} \tag{6}$$



Posterior over f and θ consists of two parts

1. The original posterior over f,
2. A posterior over θ using the **marginal likelihood**:

$$p(\mathbf{y}|\theta) = \int p(\mathbf{y}|\mathbf{w}, \theta) p(\mathbf{w}|\theta) \mathrm{d}\mathbf{w} \tag{7}$$

Maximum Marginal Likelihood: Logistic Regression

Logistic regression model (e.g. θ_2 controls basis function width):

$$p(\mathbf{w}|\theta) = \mathcal{N}(\mathbf{w}; 0, \theta_1) \tag{8}$$

$$p(y_n|\mathbf{w}, \theta) = \sigma(y_n \boldsymbol{\phi}_{\theta_2}(\mathbf{x}_n)^\top \mathbf{w}) \tag{9}$$

Can we still do Maximum Marginal Likelihood to find θ?

$$p(\mathbf{w}|\mathbf{y}, \theta) = \frac{p(\mathbf{y} \mid \mathbf{w}, \theta) p(\mathbf{w}|\theta)}{p(\mathbf{y} \mid \theta)} \tag{10}$$

- Posterior is intractable.
- Marginal likelihood is intractable.

Maximum Marginal Likelihood: Logistic Regression

Logistic regression model (e.g. θ_2 controls basis function width):

$$p(\mathbf{w}|\theta) = \mathcal{N}(\mathbf{w}; 0, \theta_1) \tag{11}$$

$$p(y_n|\mathbf{w}, \theta) = \sigma(y_n \boldsymbol{\phi}_{\theta_2}(\mathbf{x}_n)^\mathsf{T}\mathbf{w}) \tag{12}$$

Can we still do Maximum Marginal Likelihood to find θ?

$$p(\mathbf{w}|\mathbf{y}, \theta) = \frac{p(\mathbf{y} \mid \mathbf{w}, \theta)p(\mathbf{w}|\theta)}{p(\mathbf{y} \mid \theta)} \tag{13}$$

- Posterior is intractable.
- Marginal likelihood is intractable.

Variational Inference

Variational lower bound:

$$\mathcal{L}(\mathbf{v}, \theta) = \sum_{n=1}^{N} \mathbb{E}_{q_{\mathbf{v}}(\mathbf{w})}[p(y_n|\mathbf{w}, \theta))] - \mathrm{KL}[q(\mathbf{w})||p(\mathbf{w}|\theta)] \tag{14}$$

Standard form has:

- expectations written over smallest dimensional random variable possible, e.g.

$$\mathbb{E}_{p(x_1,x_2)}[\log p(x_1)p(x_2)] = \mathbb{E}_{p(x_1)}[\log p(x_1)] + \mathbb{E}_{p(x_2)}[\log p(x_2)]$$

- KL divergences separated out.
- Highlight when KL can be computed in closed-form.

Finding bounds in standard form is **exam skill** (Example on board).

Variational ML-II

Variational inference actually approximates **two** quantities of interest:

- intractable posterior,
- intractable marginal likelihood.

We can approximate Maximum Marginal Likelihood (or Type-II Maximum Likelihood) using the ELBO!

1. Maximise variational parameters to improve estimate of marglik

$$\mathbf{v}_{t+1} = \underset{\mathbf{v}}{\operatorname{argmax}}\, \mathcal{L}(\mathbf{v}, \theta_t) \tag{15}$$

2. Maximise estimate of marglik

$$\theta_{t+1} = \underset{\theta}{\operatorname{argmax}}\, \mathcal{L}(\mathbf{v}_{t+1}, \theta) \tag{16}$$

Bias of Variational ML-II

- Usually, posterior won't be exact.
- ... so neither will the marginal likelihood (KL gap).

[1] Draw: Plot generalisation, marglik, ELBO.

Bias of Variational ML-II

- Usually, posterior won't be exact.
- ... so neither will the marginal likelihood (KL gap).

$\implies$ optimum of ELBO will be different
to true that of true marginal likelihood (draw).

[1] Draw: Plot generalisation, marglik, ELBO.

Bias of Variational ML-II

- Usually, posterior won't be exact.
- ... so neither will the marginal likelihood (KL gap).

$\implies$ optimum of ELBO will be different
to true that of true marginal likelihood (draw).

- No guarantee whether model selection will work.
- Sometimes can fail catastrophically.
- **Empirical question** whether it works[1]

[1] Draw: Plot generalisation, marglik, ELBO.

References I