Variational Inference and Laplace Approximation

Advanced Statistical Inference

Simone Rossi

Laplace Approximation

1. Let
$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto \exp(-\mathcal{U}(\boldsymbol{\theta})).$$
Explain how a second-order Taylor expansion of $\mathcal{U}(\boldsymbol{\theta})$ around the mode $\widehat{\boldsymbol{\theta}}$ leads to a Gaussian approximation of the posterior. What are the mean and covariance of the resulting approximation?

2. Why is the gradient term absent in the Laplace approximation when we expand around $\widehat{\boldsymbol{\theta}}$? What condition on the Hessian is needed for the approximation to define a valid Gaussian covariance matrix?

3. Consider the unnormalized posterior
$$p(\theta \mid \boldsymbol{y}) \propto \exp\left(-\frac{1}{2}(\theta - 3)^2\right).$$
   - Write the negative log posterior $\mathcal{U}(\theta)$ up to an additive constant.
   - Compute the mode $\widehat{\theta}$.
   - Compute the Hessian at the mode.
   - Write the Laplace approximation explicitly.

4. Consider the Gamma density
$$p(x \mid \alpha, \beta) \propto x^{\alpha-1} \exp(-\beta x), \qquad x > 0.$$
   - Write the negative log density $\mathcal{U}(x)$ up to an additive constant.
   - Compute the mode for general $\alpha > 1$ and $\beta > 0$.
   - Compute the second derivative at the mode.
   - For $\alpha = 9$ and $\beta = 2$, give the mean and variance of the Laplace approximation numerically.

5. Let the Hessian at a two-dimensional mode be

$$\boldsymbol{H} = \begin{pmatrix} 4 & 1 \\ 1 & 3 \end{pmatrix}.$$

   Compute the covariance matrix of the Laplace approximation.

6. Give two practical limitations of Laplace approximation when applied to complex posteriors such as multimodal or very high-dimensional models.

KL Divergence and ELBO

1. Define the Kullback-Leibler divergence

$$\mathrm{KL}\left(q(\boldsymbol{\theta}) \| p(\boldsymbol{\theta})\right).$$

   State three important properties of this quantity and explain why it is not a true distance.

2. Use Jensen's inequality to show that

$$\mathbb{E}[\log X] \leq \log \mathbb{E}[X]$$

   for a positive random variable X.

3. For one-dimensional Gaussians

$$q(x) = \mathcal{N}(\mu_q, \sigma_q^2), \qquad p(x) = \mathcal{N}(\mu_p, \sigma_p^2),$$

   the KL divergence has a closed form. Compute $\mathrm{KL}\left(q \| p\right)$ numerically for

$$\mu_q = 1, \quad \sigma_q^2 = 4, \quad \mu_p = 0, \quad \sigma_p^2 = 1.$$

4. Assume a mean-field Gaussian variational approximation

$$q(\boldsymbol{\theta}; \boldsymbol{\nu}) = \prod_{j=1}^{J} \mathcal{N}(\theta_j \mid m_j, s_j^2)$$

   and a Gaussian prior

$$p(\boldsymbol{\theta}) = \prod_{j=1}^{J} \mathcal{N}(\theta_j \mid 0, \sigma^2).$$

   Write the closed-form expression of

$$\mathrm{KL}\left(q(\boldsymbol{\theta}; \boldsymbol{\nu}) \| p(\boldsymbol{\theta})\right).$$

   Then simplify it when $J = 2$, $\sigma^2 = 1$, $m_1 = 1$, $m_2 = -1$, $s_1^2 = 2$, and $s_2^2 = 1/2$.

5. Starting from
$$\mathrm{KL}\left(q(\boldsymbol{\theta};\boldsymbol{\nu})\|p(\boldsymbol{\theta}\mid\boldsymbol{y})\right),$$
derive the identity
$$\log p(\boldsymbol{y}) = \mathcal{L}_{\mathrm{ELBO}}(\boldsymbol{\nu}) + \mathrm{KL}\left(q(\boldsymbol{\theta};\boldsymbol{\nu})\|p(\boldsymbol{\theta}\mid\boldsymbol{y})\right).$$
Why does this imply that the ELBO is a lower bound on the log marginal likelihood?

6. Suppose for a given choice of variational parameters you estimate
$$\mathbb{E}_{q(\boldsymbol{\theta};\boldsymbol{\nu})}[\log p(\boldsymbol{y}\mid\boldsymbol{\theta})] \approx -12.4$$
and compute
$$\mathrm{KL}\left(q(\boldsymbol{\theta};\boldsymbol{\nu})\|p(\boldsymbol{\theta})\right) = 1.7.$$
Compute the ELBO. If the true log marginal likelihood is -10.1, what is the implied KL divergence between $q(\boldsymbol{\theta};\boldsymbol{\nu})$ and the true posterior?

7. Explain the role of the two ELBO terms:
   - $\mathbb{E}_{q(\boldsymbol{\theta};\boldsymbol{\nu})}[\log p(\boldsymbol{y}\mid\boldsymbol{\theta})]$;
   - $-\mathrm{KL}\left(q(\boldsymbol{\theta};\boldsymbol{\nu})\|p(\boldsymbol{\theta})\right)$. What behavior would you expect if one dominates the other too strongly during optimization?

8. The model-fitting term in the ELBO is often approximated with Monte Carlo samples
$$\mathbb{E}_{q(\boldsymbol{\theta};\boldsymbol{\nu})}[\log p(\boldsymbol{y}\mid\boldsymbol{\theta})] \approx \frac{1}{S}\sum_{s=1}^{S}\log p(\boldsymbol{y}\mid\boldsymbol{\theta}^{(s)}), \qquad \boldsymbol{\theta}^{(s)} \sim q(\boldsymbol{\theta};\boldsymbol{\nu}).$$
Why is this useful? What happens to the variance of this estimator as S increases?

Variational Inference

1. Describe the three-step recipe of variational inference:
   - choice of variational family;
   - choice of objective;
   - optimization.

2. What does the mean-field assumption buy us computationally, and what important dependency structure can it fail to represent?

3. Compare Laplace approximation and variational inference as approximate inference methods. Give one advantage and one drawback of each.