VI by Taking ELBO's Derivatives (An Example)

CS698X: Topics in Probabilistic Modeling and Inference

Piyush Rai

Mean-Field VI by Taking ELBO's Derivatives

- Mean-field assumption $q(\boldsymbol{Z}|\phi) = \prod_{i=1}^{M} q(\boldsymbol{Z}_i|\phi_i)$ results in following optimal distribution

$$q_j^*(\boldsymbol{Z}_j) = \frac{\exp(\mathbb{E}_{i\neq j}[\log p(\boldsymbol{X},\boldsymbol{Z})])}{\int \exp(\mathbb{E}_{i\neq j}[\log p(\boldsymbol{X},\boldsymbol{Z})]\, d\boldsymbol{Z}_j}$$

This approach is applicable even if we don't have mean-field assumption

Note that here we do not have to assume the form of this variational distribution. We simply compute the RHS and find what it is (in the locally-conjugate case, it will be the same distribution as the prior)

- Alternatively, we can take ELBO's partial deriv w.r.t. $\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots, \boldsymbol{\phi}_M$ to find their optimal values
- Consider a Bayesian linear regression model

Likelihood

Prior on $\boldsymbol{w}$

λ assumed fixed

Prior on variance of Gaussian likelihood

$$y_i \sim \text{Normal}(x_i^T w, \alpha^{-1}), \quad w \sim \text{Normal}(0, \lambda^{-1} I), \quad \alpha \sim \text{Gamma}(a, b)$$

Needed in ELBO

Joint distribution on data and unknowns

$$p(y, w, \alpha | x) = p(\alpha)p(w)\prod_{i=1}^{N} p(y_i|x_i, w, \alpha)$$

Assumed variational posterior with mean-field assumption

$$q(w, \alpha) = q(\alpha)q(w) = \text{Gamma}(\alpha|a', b')\text{Normal}(w|\mu', \Sigma')$$

Note that in this approach, we have to assume a form for each variational distribution. It is common to assume them to have the same form as the respective priors

- Now doing VI amounts to maximizing ELBO to find the optimal variational params a', b', μ', Σ'

Mean-Field VI by Taking ELBO’s Derivatives

- The ELBO is

For the Bayesian linear regression model, instead of $p(\boldsymbol{X},\boldsymbol{Z})$, it will be of the form $p(\boldsymbol{y},\boldsymbol{Z}|\boldsymbol{X})$

$$\mathcal{L}(q) = \mathcal{L}(\phi) = \mathbb{E}_q[\log p(\boldsymbol{X},\boldsymbol{Z})] - \mathbb{E}_q[\log q(\boldsymbol{Z})] = \mathbb{E}_q[\log p(\boldsymbol{Z})] + \mathbb{E}_q[\log p(\boldsymbol{X}|\boldsymbol{Z})] - \mathbb{E}_q[\log q(\boldsymbol{Z})]$$
$$= \int q(\boldsymbol{Z}) \log p(\boldsymbol{Z}) d\boldsymbol{Z} + \int q(\boldsymbol{Z}) \log p(\boldsymbol{X}|\boldsymbol{Z}) d\boldsymbol{Z} + \int q(\boldsymbol{Z}) \log q(\boldsymbol{Z}) d\boldsymbol{Z}$$

- Thus the ELBO in the Bayesian linear regression model will be (assuming i.i.d. obs)

Expectations of the log of the likelihood

Expectations of the log of the var. distributions (= their entropies)

$$\mathcal{L}(a',b',\mu',\Sigma') = \int q(\alpha)\ln p(\alpha)d\alpha + \int q(w)\ln p(w)dw$$
$$+ \sum_{i=1}^{N} \int\int q(\alpha)q(w)\ln p(y_i|x_i,w,\alpha)dw d\alpha - \int q(\alpha)\ln q(\alpha)d\alpha - \int q(w)\ln q(w)dw$$

Expectations of the log of the prior

- Substituting the priors, likelihoods, and variational distributions

$$\mathcal{L}(a',b',\mu',\Sigma') = (a-1)(\psi(a') - \ln b') - b\frac{a'}{b'} + \text{constant} - \frac{\lambda}{2}(\mu'^T\mu' + \text{tr}(\Sigma')) + \text{constant} + \frac{N}{2}(\psi(a') - \ln b') - \sum_{i=1}^{N}\frac{1}{2}\frac{a'}{b'}\left((y_i - x_i^T\mu')^2 + x_i^T\Sigma' x_i\right) + \text{constant}$$
$$+a' - \ln b' + \ln\Gamma(a') + (1-a')\psi(a') + \frac{1}{2}\ln|\Sigma'| + \text{constant}$$

Digamma function (log of gamma function)

- Can now maximize the above ELBO w.r.t. $\boldsymbol{a'}, \boldsymbol{b'}, \boldsymbol{\mu'}, \boldsymbol{\Sigma'}$ in an alternating fashion
- For most models, ELBO or its gradients won’t have a simple form (methods like BBVI, reparam trick etc will be needed in those cases)