ELBO Derivations (Version 4)

Joe Benton

October 29, 2021

1 Introduction

The likelihood is given by

$$\log p(y|X, \beta) = -\sum_{i=1}^{n} \log(1 + e^{-y_i \beta^T x_i}) \tag{1}$$

We put a normal prior $\beta \sim N(0, \tau^2 I_m)$ on β.

We use the Gaussian mean-field variational family

$$q(\beta) = \prod_{j=1}^{m} q_j(\beta_j; \mu_j, \rho_j) \tag{2}$$

where $q_j(\beta_j; \mu_j, \rho_j) = \mathcal{N}(\beta_j, \mu_j, \sigma_j^2)$ is a Gaussian distribution with

$$\sigma_j = \log(1 + \exp(\rho_j)) \tag{3}$$

and $\{\mu_j, \rho_j\}_{j=1}^m$ is our set of variational parameters.

The ELBO is

$$\text{ELBO}(q) = \mathbb{E}_{q(\beta|\mu,\rho)}[\log q(\beta|\mu, \rho)] - \mathbb{E}_{q(\beta|\mu,\rho)}[\log p(y, \beta|X)] \tag{4}$$

We can decouple the randomness in β from μ, ρ by defining

$$\beta_j = \mu_j + \log(1 + \exp(\rho_j))\epsilon_j, \qquad \epsilon \sim \mathcal{N}(0, I_m) \tag{5}$$

Then we can calculate

$$\log q(\beta|\mu,\rho) = \sum_{j=1}^{m} \log q_j(\beta_j|\mu_j,\rho_j) \tag{6}$$

$$= -\frac{1}{2}\sum_{j=1}^{m}\log(2\pi\sigma_j^2) - \frac{1}{2}\sum_{j=1}^{m}\frac{(\beta_j-\mu_j)^2}{\sigma_j^2} \tag{7}$$

$$\frac{\partial}{\partial\beta_j}\log q(\beta|\mu,\rho) = -\frac{\beta_j-\mu_j}{\sigma_j^2} \tag{8}$$

$$\frac{\partial}{\partial\mu_j}\log q(\beta|\mu,\rho) = \frac{\beta_j-\mu_j}{\sigma_j^2} \tag{9}$$

$$\frac{\partial}{\partial\rho_j}\log q(\beta|\mu,\rho) = -\frac{1}{1+\exp(-\rho_j)}\Big(\frac{1}{\sigma_j} - \frac{(\beta_j-\mu_j)^2}{\sigma_j^3}\Big) \tag{10}$$

and

$$\log p(y,\beta|X) = \log p(y|X,\beta) + \log p(\beta) \tag{11}$$

$$= -\sum_{i=1}^{n}\log(1+e^{-y_i\beta^T x_i}) - \frac{m}{2}\log(2\pi\tau^2) - \frac{1}{2\tau^2}||\beta||^2 \tag{12}$$

$$\frac{\partial}{\partial\beta_j}\log p(y,\beta|X) = \sum_{i=1}^{n} x_{ij}y_i\cdot\frac{e^{-y_i\beta^T x_i}}{1+e^{-y_i\beta^T x_i}} - \frac{\beta_j}{\tau^2} \tag{13}$$

$$\frac{\partial}{\partial\mu_j}\log p(y,\beta|X) = \frac{\partial}{\partial\rho_j}\log p(y,\beta|X) = 0 \tag{14}$$

In addition, we have

$$\frac{\partial\beta_k}{\partial\mu_j} = \delta_{jk}, \qquad \frac{\partial\beta_k}{\partial\rho_j} = \frac{\epsilon_j\delta_{jk}}{1+\exp(-\rho_j)} \tag{15}$$

Putting this all together, we get

$$\frac{\partial}{\partial\mu_j}\mathrm{ELBO}(q) = \frac{\partial}{\partial\mu_j}\mathbb{E}_{q(\beta|\mu,\rho)}[\log q(\beta|\mu,\rho)] - \frac{\partial}{\partial\mu_j}\mathbb{E}_{q(\beta|\mu,\rho)}[\log p(y,\beta|X)] \tag{16}$$

$$= \mathbb{E}_{q(\epsilon)}\Big[\sum_{k=1}^{m}\frac{\partial\beta_k}{\partial\mu_j}\frac{\partial}{\partial\beta_k}\log q(\beta|\mu,\rho) + \frac{\partial}{\partial\mu_j}\log q(\beta|\mu,\rho)\Big] \tag{17}$$

$$- \mathbb{E}_{q(\epsilon)}\Big[\sum_{k=1}^{m}\frac{\partial\beta_k}{\partial\mu_j}\frac{\partial}{\partial\beta_k}\log p(y,\beta|X) + \frac{\partial}{\partial\mu_j}\log p(y,\beta|X)\Big] \tag{18}$$

$$= \mathbb{E}_{q(\epsilon)}\Big[-\frac{\beta_j-\mu_j}{\sigma_j^2} + \frac{\beta_j-\mu_j}{\sigma_j^2}\Big] \tag{19}$$

$$- \mathbb{E}_{q(\epsilon)}\Big[\sum_{i=1}^{n} x_{ij}y_i\cdot\frac{e^{-y_i\beta^T x_i}}{1+e^{-y_i\beta^T x_i}} - \frac{\beta_j}{\tau^2}\Big] \tag{20}$$

$$= -\mathbb{E}_{q(\epsilon)}\Big[\sum_{i=1}^{n} x_{ij}y_i\cdot\frac{e^{-y_i\beta^T x_i}}{1+e^{-y_i\beta^T x_i}} - \frac{\beta_j}{\tau^2}\Big] \tag{21}$$

Similarly, we get

$$\frac{\partial}{\partial \rho_j}\text{ELBO}(q) = \frac{\partial}{\partial \rho_j}\mathbb{E}_{q(\beta|\mu,\rho)}[\log q(\beta|\mu,\rho)] - \frac{\partial}{\partial \rho_j}\mathbb{E}_{q(\beta|\mu,\rho)}[\log p(y,\beta|X)] \tag{22}$$

$$= \mathbb{E}_{q(\epsilon)}\Big[\sum_{k=1}^{m}\frac{\partial \beta_k}{\partial \rho_j}\frac{\partial}{\partial \beta_k}\log q(\beta|\mu,\rho) + \frac{\partial}{\partial \rho_j}\log q(\beta|\mu,\rho)\Big] \tag{23}$$

$$- \mathbb{E}_{q(\epsilon)}\Big[\sum_{k=1}^{m}\frac{\partial \beta_k}{\partial \rho_j}\frac{\partial}{\partial \beta_k}\log p(y,\beta|X) + \frac{\partial}{\partial \rho_j}\log p(y,\beta|X)\Big] \tag{24}$$

$$= \mathbb{E}_{q(\epsilon)}\Big[- \frac{\epsilon_j}{1+\exp(-\rho_j)} \cdot \frac{\beta_j - \mu_j}{\sigma_j^2} \tag{25}$$

$$- \frac{1}{1+\exp(-\rho_j)}\Big(\frac{1}{\sigma_j} - \frac{(\beta_j-\mu_j)^2}{\sigma_j^3}\Big)\Big] \tag{26}$$

$$- \mathbb{E}_{q(\epsilon)}\Big[\frac{\epsilon_j}{1+\exp(-\rho_j)} \cdot \Big(\sum_{i=1}^{n} x_{ij} y_i \cdot \frac{e^{-y_i\beta^T x_i}}{1+e^{-y_i\beta^T x_i}} - \frac{\beta_j}{\tau^2}\Big)\Big] \tag{27}$$

[Comment: Line (24) probably has an analytic solution. Is that worth evaluating?]

2 Multinomial Logisitic Regression

We denote our data by $\{x_i, y_i\}_{i=1}^n$ where each x_i is an $m \times 1$ feature vector and y_i is a $k \times 1$ one-hot representation of the class of the ith data point.

Our likelihood is then given by

$$\log p(Y|X,\beta) = \sum_{i=1}^{n}\log\Big(\frac{\exp(y_i^T\beta x_i)}{\sum_{j=1}^{k}\exp(z_j^T\beta x_i)}\Big) \tag{28}$$

where β is a $k \times m$ matrix of parameters and z_j is the one-hot representation of the jth class. As before, we put a normal prior on each component of β, so that $\beta_{ij} \sim \mathcal{N}(0,\tau^2)$.

We then perform variational inference on the posterior $p(\beta|X,Y)$ using the variational family

$$q(\beta|\mu,\rho) = \prod_{i=1}^{k}\prod_{j=1}^{m} q_{ij}(\beta_{ij};\mu_{ij},\rho_{ij}) \tag{29}$$

where $q_{ij}(\beta_{ij};\mu_{ij},\rho_{ij}) = \mathcal{N}(\beta_{ij};\mu_{ij},\sigma_{ij}^2)$ with $\sigma_{ij} = \log(1+\exp(\rho_{ij}))$ and our variational parameters are $\{\mu_{ij},\rho_{ij}\}_{i=1,\dots,k;\ j=1,\dots,m}$.

The ELBO is given by

$$\text{ELBO}(q) = \mathbb{E}_{q(\beta|\mu,\rho)}[\log p(Y,\beta|X)] - \mathbb{E}_{q(\beta|\mu,\rho)}[\log q(\beta|\mu,\rho)] \tag{30}$$

In order to calculate the gradient of the ELBO, we can decouple the randomness in β from μ, ρ by writing

$$\beta_{ij} = \mu_{ij} + \sigma_{ij}\epsilon_{ij}, \qquad \epsilon_{ij} \sim \mathcal{N}(0,1) \tag{31}$$

We can then compute

$$\log q(\beta|\mu,\rho) = -\frac{1}{2}\sum_{i,j}\log(2\pi\sigma_{ij}^2) - \frac{1}{2}\sum_{i,j}\frac{(\beta_{ij}-\mu_{ij})^2}{\sigma_{ij}^2} \tag{32}$$

and so

$$\frac{\partial}{\partial \beta_{ij}}\log q(\beta|\mu,\rho) = -\frac{\beta_{ij}-\mu_{ij}}{\sigma_{ij}^2} = -\frac{\epsilon_{ij}}{\sigma_{ij}} \tag{33}$$

$$\frac{\partial}{\partial \mu_{ij}}\log q(\beta|\mu,\rho) = \frac{\beta_{ij}-\mu_{ij}}{\sigma_{ij}^2} = \frac{\epsilon_{ij}}{\sigma_{ij}} \tag{34}$$

$$\frac{\partial}{\partial \sigma_{ij}}\log q(\beta|\mu,\rho) = -\frac{1}{\sigma_{ij}} + \frac{(\beta_{ij}-\mu_{ij})^2}{\sigma_{ij}^3} \tag{35}$$

$$= -\frac{1}{\sigma_{ij}} + \frac{\epsilon_{ij}^2}{\sigma_{ij}} \tag{36}$$

Similarly,

$$\log p(Y,\beta|X) = \sum_{r=1}^{n}\log\Big(\frac{\exp(y_r^T\beta x_r)}{\sum_{s=1}^{k}\exp(z_s^T\beta x_r)}\Big) - \frac{km}{2}\log(2\pi\tau^2) - \frac{1}{2\tau^2}||\beta||^2 \tag{37}$$

and so

$$\frac{\partial}{\partial \beta_{ij}}\log p(Y,\beta|X) = \sum_{r=1}^{n}\Big\{x_{rj}y_{ri} - \frac{x_{rj}\exp(z_i^T\beta x_r)}{\sum_{s=1}^{k}\exp(z_s^T\beta x_r)}\Big\} - \frac{\beta_{ij}}{\tau^2} \tag{38}$$

$$\frac{\partial}{\partial \mu_{ij}}\log p(Y,\beta|X) = \frac{\partial}{\partial \sigma_{ij}}\log p(Y,\beta|X) = 0 \tag{39}$$

So, computing the gradient of the ELBO, we get

$$\frac{\partial}{\partial \mu_{ij}}\mathrm{ELBO}(q) = \mathbb{E}_{q(\epsilon)}\Big[\sum_{r,s}\frac{\partial \beta_{rs}}{\partial \mu_{ij}}\frac{\partial}{\partial \beta_{rs}}\log p(Y,\beta|X) + \frac{\partial}{\partial \mu_{ij}}\log p(Y,\beta|X)\Big] \tag{40}$$

$$- \mathbb{E}_{q(\epsilon)}\Big[\sum_{r,s}\frac{\partial \beta_{rs}}{\partial \mu_{ij}}\frac{\partial}{\partial \beta_{rs}}\log q(\beta|\mu,\rho) + \frac{\partial}{\partial \mu_{ij}}\log q(\beta|\mu,\rho)\Big] \tag{41}$$

$$= \mathbb{E}_{q(\epsilon)}\Big[\sum_{r=1}^{n}\Big\{x_{rj}y_{ri} - \frac{x_{rj}\exp(z_i^T\beta x_r)}{\sum_{s=1}^{k}\exp(z_s^T\beta x_r)}\Big\} - \frac{\beta_{ij}}{\tau^2}\Big] \tag{42}$$

and

$$\frac{\partial}{\partial \rho_{ij}} \text{ELBO}(q) = \frac{\partial \sigma_{ij}}{\partial \rho_{ij}} \frac{\partial}{\partial \sigma_{ij}} \text{ELBO}(q) \tag{43}$$

$$= \frac{1}{1+\exp(-\rho_{ij})} \Big\{ \mathbb{E}_{q(\epsilon)} \Big[\sum_{r,s} \frac{\partial \beta_{rs}}{\partial \sigma_{ij}} \frac{\partial}{\partial \beta_{rs}} \log p(Y, \beta | X) + \frac{\partial}{\partial \sigma_{ij}} \log p(Y, \beta | X) \Big] \tag{44}$$

$$- \mathbb{E}_{q(\epsilon)} \Big[\sum_{r,s} \frac{\partial \beta_{rs}}{\partial \sigma_{ij}} \frac{\partial}{\partial \beta_{rs}} \log q(\beta | \mu, \rho) + \frac{\partial}{\partial \sigma_{ij}} \log q(\beta | \mu, \rho) \Big] \Big\} \tag{45}$$

$$= \frac{1}{1+\exp(-\rho_{ij})} \Big\{ \mathbb{E}_{q(\epsilon)} \Big[\epsilon_{ij} \Big\{ \sum_{r=1}^{n} \Big(x_{rj} y_{ri} - \frac{x_{rj} \exp(z_i^T \beta x_r)}{\sum_{s=1}^{k} \exp(z_s^T \beta x_r)} \Big) - \frac{\beta_{ij}}{\tau^2} \Big\} + \frac{1}{\sigma_{ij}} \Big] \Big\} \tag{46}$$