Bayesian models for Large-scale Hierarchical Classification (Supplementary Material)

Siddharth Gopal
sgopal1@andrew.cmu.edu

Yiming Yang
yiming@cs.cmu.edu

Bing Bai
bing@nec-labs.com

Alex Niculescu-Mizil
alex@nec-labs.com

1 Variational Inference for HBLR models

For the supplementary material we slightly change the notation as follows,

Define a hierarchy as a set of nodes $\mathcal{N} = \{1, 2...\}$ with the parent relationship $\pi : \mathcal{N} \to \mathcal{N}$ where $\pi(n)$ is the parent of node $n \in \mathcal{N}$. Let $\mathbf{D} = \{(x_i, t_i)\}_{i=1}^N$ denote the training data where $x_i \in \mathbb{R}^d$ is an instance, $t_i \in T$ is a label, where $T \subset \mathcal{N}$ is the set of leaf nodes in the hierarchy labeled from 1 to $|T|$. We assume that each instance is assigned to one of the leaf nodes in the hierarchy. Let C_n be the set of all children of node n.

Models M1,M2 and M3 are defined as follows,

$$\begin{aligned}
\textbf{M1} \quad & w_{root} \sim \mathcal{N}(w_0, \Sigma_0), \qquad \alpha_{root} \sim \quad \Gamma(a_0, b_0) \\
& w_n |\, w_{\pi(n)}, \Sigma_{\pi(n)} \sim \mathcal{N}(w_{\pi(n)}, \Sigma_{\pi(n)}) \quad \forall n, \qquad \alpha_n \sim \quad \Gamma(a_n, b_n) \quad \forall n \notin T \\
& t \mid x, \mathbf{W} \sim \quad \textit{Categorical}(p_1(x), p_2(x), .., p_{|T|}(x)) \quad \forall (x,t) \in \mathbf{D} \\
& p_i(x) = \quad \exp(w_i^\top x) / \Sigma_{t' \in T} \exp(w_{t'}^\top x)
\end{aligned} \tag{1}$$

$$\begin{aligned}
\textbf{M2} \quad & w_n |\, w_{\pi(n)}, \Sigma_{\pi(n)} \sim \mathcal{N}(w_{\pi(n)}, \Sigma_{\pi(n)}) \qquad \forall n \\
& \alpha_n^{(i)} \sim \Gamma(a_n^{(i)}, b_n^{(i)}) \quad i = 1..d, \ \forall n \notin T \\
& \qquad \text{where } \Sigma_{\pi(n)}^{-1} = \text{diag}(\alpha_{\pi(n)}^{(1)}, \alpha_{\pi(n)}^{(2)}, \ldots, \alpha_{\pi(n)}^{(d)})
\end{aligned}$$

$$\begin{aligned}
\textbf{M3} \quad & w_n |\, w_{\pi(n)}, \Sigma_n \sim \mathcal{N}(w_{\pi(n)}, \Sigma_n) \qquad \forall n \\
& \qquad \alpha_n \sim \Gamma(a_n, b_n) \qquad \forall n \notin T
\end{aligned}$$

1.1 Variational Inference for model M2

The posterior of the model parameters is given by,

$$\begin{aligned}
p(\mathbf{W}, \boldsymbol{\alpha} | \mathbf{D}) \propto & p(\mathbf{D} | \mathbf{W}, \boldsymbol{\alpha}) p(\mathbf{W}, \boldsymbol{\alpha}) \\
p(\mathbf{D} | \mathbf{W}, \boldsymbol{\alpha}) = & \prod_{(x,t) \in D} \frac{\exp(w_t^\top x)}{\sum_{t' \in T} \exp(w_{t'}^\top x)} \qquad (2) \\
p(\mathbf{W}, \boldsymbol{\alpha}) = & \prod_{n \in \mathcal{N} \setminus T} \prod_{i=1}^{d} p(\alpha_n^{(i)} | a_n^{(i)}, b_n^{(i)}) \prod_{n \in \mathcal{N}} p(w_n | w_{\pi(n)}, \Sigma_{\pi(n)}) \\
= & \prod_{n \in \mathcal{N} \setminus T} \prod_{i=1}^{d} \Gamma(\alpha_n^{(i)} | a_n^{(i)}, b_n^{(i)}) \prod_{n \in \mathcal{N}} \mathcal{N}(w_n | w_{\pi(n)}, \Sigma_{\pi(n)}) \qquad (3)
\end{aligned}$$

The posterior has a logistic likelihood term with the $\boldsymbol{W}$ in (2) and with a Gamma and Normal prior over $\boldsymbol{\alpha}$, $\boldsymbol{W}$ in (3). The convolution between a normal-gamma prior and logistic likelihood cannot be computed in closed form; therefore one has to resort to approximate methods to calculate the posterior.

Variational methods try to compute an approximate posterior having a simplified factored form which is closest in KL divergence to the true posterior. They rely on the following bound for the log-marginal probability of D. For any distribution $q(\boldsymbol{W}, \boldsymbol{\alpha})$,

$$\begin{aligned}
\log P(\boldsymbol{D}) &= \int q(\boldsymbol{W}, \boldsymbol{\alpha}) \log \frac{p(\boldsymbol{W}, \boldsymbol{\alpha}, \boldsymbol{D})}{q(\boldsymbol{W}, \boldsymbol{\alpha})} d\boldsymbol{W} d\boldsymbol{\alpha} - \int q(\boldsymbol{W}, \boldsymbol{\alpha}) \log \frac{p(\boldsymbol{W}, \boldsymbol{\alpha}|\boldsymbol{D})}{q(\boldsymbol{W}, \boldsymbol{\alpha})} d\boldsymbol{W} d\boldsymbol{\alpha} \quad (4)\\
&\geq \int q(\boldsymbol{W}, \boldsymbol{\alpha}) \log \frac{p(\boldsymbol{W}, \boldsymbol{\alpha}, \boldsymbol{D})}{q(\boldsymbol{W}, \boldsymbol{\alpha})} d\boldsymbol{W} d\boldsymbol{\alpha}\\
&= \int q(\boldsymbol{W}, \boldsymbol{\alpha}) \log p(\boldsymbol{W}, \boldsymbol{\alpha}, \boldsymbol{D}) d\boldsymbol{W} d\boldsymbol{\alpha} - \int q(\boldsymbol{W}, \boldsymbol{\alpha}) \log q(\boldsymbol{W}, \boldsymbol{\alpha}) d\boldsymbol{W} d\boldsymbol{\alpha} \quad (5)\\
VLB &= E_q \left[\log p(\boldsymbol{W}, \boldsymbol{\alpha}, \boldsymbol{D})\right] + H(q) \quad (6)
\end{aligned}$$

Firstly, not that equation(4) can be easily verified by combining the RHS terms. Equation (6) represents the variational lower-bound for the likelihood of the data, this is quantity that we maximize w.r.t q.

In order to compute such a q, we start by assuming a simplified factored form using independent distributions for each parameter. Note that this does not neglect the dependence between the various parameters in the model; it just finds the suitable factored form which best approximates the dependencies.

$$\begin{aligned}
q(\mathbf{W}, \boldsymbol{\alpha}) &= \prod_{n \in \mathcal{N} \setminus T} q(\alpha_n) \prod_{n \in \mathcal{N}} q(w_n)\\
q(w_n) &\propto \mathcal{N}(.|\mu_n, \Psi_n)\\
q(\alpha_n) &= \prod_{i=1}^{d} q(\alpha_n^{(i)}) \propto \prod_{i=1}^{d} \Gamma(.|\tau_n^{(i)}, \upsilon_n^{(i)})
\end{aligned}$$

Here $q(\mathbf{W}, \boldsymbol{\alpha}) \equiv q(\mathbf{W}|\boldsymbol{\mu}, \boldsymbol{\Psi}) q(\boldsymbol{\alpha}|\boldsymbol{\tau}, \boldsymbol{\upsilon})$ where $\boldsymbol{\tau}, \boldsymbol{\upsilon}, \boldsymbol{\mu}, \boldsymbol{\Sigma}$ are variational parameters which we can optimize one at a time to maximize (6).

For example, to optimize w_n, we differentiate VLB w.r.t $q(w_n)$,

$$VLB = \int q(w_n|\mu_n, \Psi_n) \left[q(\boldsymbol{W}^{-w_n}, \boldsymbol{\alpha}) \log p(\boldsymbol{W}, \boldsymbol{\alpha}, \boldsymbol{D}) d\boldsymbol{W} d\boldsymbol{\alpha}\right] - \int q(w_n|\mu_n, \Psi_n) \log q(w_n|\mu_n, \Psi_n) dw_n$$

Setting the gradient w.r.t $q(w_n)$ to zero yields (7)

$$\log q^*(w_n|\mu_n, \Psi_n) = E_{q^{-w_n}} \left[\log p(\boldsymbol{W}, \boldsymbol{\alpha}, \boldsymbol{D})\right] \quad + constant \quad (8)$$

where $-w_n$ denotes all parameters other than w_n. Similary,

$$\log q^*(\alpha_n|\tau_n, \upsilon_n) = E_{q^{-\alpha_n}} \left[\log p(\boldsymbol{W}, \boldsymbol{\alpha}, \boldsymbol{D})\right] \quad + constant \quad (9)$$

In general, to update the parameters we need to calculate the expected log-likelihood with variational posterior i.e. $E_q[\log p(\boldsymbol{W}, \boldsymbol{\alpha}, \boldsymbol{D})]$.

$$E_q[\log p(\boldsymbol{W}, \boldsymbol{\alpha}, \boldsymbol{D})] = E_q \left[\log P(\boldsymbol{D}|\boldsymbol{W}, \boldsymbol{\alpha})\right] + E_q \left[\log P(\boldsymbol{W}, \boldsymbol{\alpha})\right]$$

Where,

$$\begin{aligned}
E_q \left[\log P(\boldsymbol{W}, \boldsymbol{\alpha})\right] = &\sum_{n \in \mathcal{N}} -\frac{1}{2} E_q \left[(w_n - w_{\pi(n)})^\top \Sigma_{\pi(n)}^{-1} (w_n - w_{\pi(n)})\right]\\
&- \sum_{n \in \mathcal{N}} \frac{1}{2} E_q \left[\log |\Sigma_n|\right] - \sum_{n \in \mathcal{N}/T} \sum_{i=1}^{d} E_q \left[\alpha_n^{(i)}\right] b_n + E_q \left[(a_n^{(i)} - 1) \log(a_n^{(i)})\right]
\end{aligned} \quad (10)$$

$$E_q\left[\log P(\boldsymbol{D}|\boldsymbol{W},\boldsymbol{\alpha})\right] = \sum_{(x,t)\in D} \mu_n^\top x - \sum_{(x,t)\in D} E_q\left[\log\left(\sum_{n\in T}\exp(w_n^\top x)\right)\right] \tag{11}$$

The expectation of the log-sum-exp function is not computable in closed form. Therefore we replace the log-sum-exp with suitable lower-bound proposed in [1] whose expectation can be computed in closed form.

$$\begin{aligned}
\log\left(\sum_{n\in T}\exp(w_n^\top x)\right) \leq & \beta_x + \sum_{n\in T}\frac{(w_n^\top x - \beta_x - \xi_{xn})}{2} + \sum_{n\in T}\lambda(\xi_{xn})\left(\left(w_n^\top x - \beta_x\right)^2 - \xi_{xn}^2\right) \\
& + \sum_{n\in T}\log(1+e^{\xi_{xn}}) \\
\leq & -(\frac{|T|}{2}-1)\beta_x + \sum_{n\in T}\frac{w_n^\top x}{2} - \sum_{n\in T}\frac{\xi_{xn}}{2} \\
& + \sum_{n\in T}\lambda(\xi_{xn})\left(w_n^\top(xx^\top)w_n - 2\beta_x(w_n^\top x) + \beta_x^2 - \xi_{xn}^2\right) + \sum_{n\in T}\log(1+e^{\xi_{xn}})
\end{aligned}$$

Here we have introduced variational parameter β_x and ξ_{xn} for every $x \in D, n \in \mathcal{N}$. In order to get the tightest possible bound, we can optimize over these variational parameters. The expected log-sum-exp function can be computed as followed (the function $\lambda(x)$ is defined in the paper)

$$\begin{aligned}
E_q\left[\log\left(\sum_{n\in T}\exp(w_n^\top x)\right)\right] \leq & \sum_{n\in\mathcal{N}}\left(\frac{1}{2} - 2\lambda(\xi_{xn})\beta_x\right)\mu_n^\top x + \sum_{n\in\mathcal{N}}\mu_n^\top(2\lambda(\xi_{xn})xx^\top)\mu_n \\
& - (\frac{|T|}{2}-1)\beta_x - \sum_{n\in\mathcal{N}}\left(\lambda(\xi_{xn})\beta_x^2 - \frac{\xi_{xn}}{2} - \lambda(\xi_{xn})\xi_{xn}^2 + \log(1+e^{\xi_{xn}})\right)
\end{aligned} \tag{12}$$

1.1.1 Optimizing $q^*(w_n)$

Combining the bound for the logistic soft-max function with (11) and (10), the update for parameter $w_n, n \in T$ can be written as

$$\begin{aligned}
\log q^*(w_n|\mu_n,\Psi_n) = & \left(\sum_{(x,t)\in D}(I(t=n) - \frac{1}{2} + 2\lambda(\xi_{xn})\beta_x)x + E_{q^{-w_n}}\left[\Sigma_{\pi(n)}^{-1}\right]\mu_\pi(n)) - w_n\right)^\top \\
& \left(\sum_{(x,t)\in D}2\lambda(\xi_{xn})xx^\top + E_{q^{-w_n}}\left[\Sigma_{\pi(n)}^{-1}\right]\right) \\
& \left(\sum_{(x,t)\in D}(I(t=n) - \frac{1}{2} + 2\lambda(\xi_{xn})\beta_x)x + E_{q^{-w_n}}\left[\Sigma_{\pi(n)}^{-1}\right]\mu_\pi(n)) - w_n\right) + constant
\end{aligned}$$

From the above, we can directly match the sufficient statistics i.e. mean and the covariance matrix of $q^*(w_n)$ and set μ_n and Ψ_n as in the main paper. Note that we have used the fact that

$$\begin{aligned}
E_{q^{-w_n}}\left[\Sigma_n^{-1}\right] = & \operatorname{diag}\left(E_{q^{-w_n}}\left[\alpha_n^{(1)}\right], E_{q^{-w_n}}\left[\alpha_n^{(2)}\right], \ldots, E_{q^{-w_n}}\left[\alpha_n^{(d)}\right])\right) \\
= & \operatorname{diag}\left(\frac{\tau_n^{(1)}}{\upsilon_n^{(1)}}, \frac{\tau_n^{(2)}}{\upsilon_n^{(2)}}, \ldots, \frac{\tau_n^{(d)}}{\upsilon_n^{(d)}})\right)
\end{aligned}$$

For $n \notin T$, fortunately there is not convolution with a logistic function, therefore the posterior can be derived similar to convolutions of Normal distributions.

1.1.2 Optimizing $q^*(\alpha_n)$

For optimizing $q^*(\alpha_n)$, we essentially follow the same strategy as above.

$$\begin{aligned}\log q^*(\alpha_n^{(i)}|\tau_n^{(i)},\upsilon_n^{(i)}) &= -\sum_{c\in C_n}\frac{\alpha_n^{(i)}}{2}E_q\left[(w_n^{(i)}-w_c^{(i)})^2\right]+\left(\sum_{c\in C_n}\frac{\log(\alpha_n^{(i)})}{2}\right)-\alpha_n^{(i)}b_n^{(i)}+(a_n^{(i)}-1)\log(\alpha_n^{(i)})\\ &= -\alpha_n^{(i)}\left(b_n^{(i)}+\sum_{c\in C_n}\frac{1}{2}\left((\mu_c^{(i)}-\mu_n^{(i)})^2+\Psi_n^{(i,i)}+\Psi_c^{(i,i)}\right)\right)\\ &+(a_n^{(i)}+\frac{|C_n|}{2}-1)\log(\alpha_n^{(i)})+constant\end{aligned}$$

Since $q^*(\alpha_n)$ is a gamma distribution, we simply match the sufficient statistcs and update. Note that we have used the fact that

$$-\log|\Sigma_n| = \log|\Sigma_n^{-1}| = \log\prod_{i=1}^{d}\alpha_n^{(i)} = \sum_{i=1}^{d}\log\alpha_n^{(i)}$$

1.2 Variational Inference for model M1

Most of the derivation simply goes through changed except that $\alpha_n^{(1)}=\alpha_n^{(2)}=\ldots=\alpha_n$. Although, this makes a difference only in how the α's are updated; we present the full update equations.

If $n\in T$

$$\begin{aligned}\Psi_n^{-1} &= \sum_{(x,t)\in D}2\lambda(\xi_{xn})xx^\top+E_{q^{-w_n}}\left[\Sigma_{\pi(n)}^{-1}\right]\\ \mu_n &= \Psi_n\left(\sum_{(x,t)\in D}(I(t=n)-\frac{1}{2}+2\lambda(\xi_{xn})\beta_x)x+E_{q^{-w_n}}\left[\Sigma_{\pi(n)}^{-1}\right]\mu_{\pi(n)}\right)\end{aligned}$$

If $n\notin T$

$$\begin{aligned}\Psi_n^{-1} &= E_{q^{-w_n}}\left[\Sigma_{\pi(n)}^{-1}\right]+|C_n|E_{q^{-w_n}}\left[\Sigma_n^{-1}\right]\\ \mu_n &= \Psi_n\left(E_{q^{-w_n}}\left[\Sigma_{\pi(n)}^{-1}\right]\mu_{\pi(n)}+E_{q^{-w_n}}\left[\Sigma_n^{-1}\right]\sum_{c\in C_n}\mu_c\right)\\ \upsilon_n &= b_n+\frac{1}{2}\sum_{c\in C_n}trace(\Psi_n)+trace(\Psi_c)+(\mu_n-\mu_c)^\top(\mu_n-\mu_c)\\ \tau_n &= a_n+\frac{|C_n|d}{2}\end{aligned}$$

where

$$E_{q^{-w_n}}\left[\Sigma_n^{-1}\right]=\text{diag}\left(\frac{\tau_n}{\upsilon_n},\frac{\tau_n}{\upsilon_n},\ldots,\frac{\tau_n}{\upsilon_n})\right)$$

1.3 Variational Inference for model M3

The extension to HBLR-M3 follows on similar lines.

If $n \in T$

$$\Psi_n^{-1} = \sum_{(x,t)\in D} 2\lambda(\xi_{xn}) x x^\top + E_{q^{-w_n}}\left[\Sigma_n^{-1}\right]$$

$$\mu_n = \Psi_n \left(\sum_{(x,t)\in D} (I(t=n) - \frac{1}{2} + 2\lambda(\xi_{xn})\beta_x) x + E_{q^{-w_n}}\left[\Sigma_n^{-1}\right] \mu_{\pi(n)} \right)$$

If $n \notin T$

$$\Psi_n^{-1} = E_{q^{-w_n}}\left[\Sigma_n^{-1}\right] + \sum_{c\in C_n} E_{q^{-w_n}}\left[\Sigma_c^{-1}\right]$$

$$\mu_n = \Psi_n \left(E_{q^{-w_n}}\left[\Sigma_n^{-1}\right] \mu_{\pi(n)} + \sum_{c\in C_n} E_{q^{-w_n}}\left[\Sigma_c^{-1}\right] \mu_c \right)$$

$$\upsilon_n = b_n + \frac{1}{2}\left(trace(\Psi_{\pi(n)}) + trace(\Psi_n) + (\mu_n - \mu_{\pi(n)})^\top (\mu_n - \mu_{\pi(n)}) \right)$$

$$\tau_n = a_n + \frac{d}{2}$$

2 Empirical Bayes

In this section, we show why Empirical Bayes route for learning the hyperparameters in our model does not work. Let us model M1 for instance. The general procedure for Empirical Bayes is to maximize the marginal likelihood w.r.t to the hyperparameters and get point estimates for them (another approach would be use MCMC, which we do not pursue in the interested of scalability).

In M1, the marginal likelihood of the data can be computed as

$$P(D|a,b) \propto \int_{\boldsymbol{\alpha}} P(\boldsymbol{\alpha}|a,b) \int_{\boldsymbol{W}} P(D|\boldsymbol{W}) P(\boldsymbol{W}|\boldsymbol{\alpha})$$

As before the integral cannot be computed in closed form and hence cannot be maximized analytically. Therefore we use the variational lower-bound as a proxy which is more amenable to maximization.

$$\begin{aligned}
\log P(D|a,b) &\geq E_q[logP(D|a,b)] \\
&\geq E_q[logP(\alpha|a,b)] + \text{ terms that do not depend on } a,b \\
&\geq \sum_{n\in\mathcal{N}\setminus T} E_q[log(P(\alpha_n|a_n,b_n)] \\
&\geq \sum_{n\in\mathcal{N}\setminus T} E_q\left[b_n^{a_n} \frac{1}{\Gamma(a_n)} \alpha_n^{a_n-1} e^{-b_n\alpha} \right]
\end{aligned}$$

Note that the maximization can be carried out independently for all the a_n's. This leads to a Gamma distribution type MLE of the following form,

$$(a_n^*, b_n^*) = \arg\max a_n \log(b_n) - \log\Gamma(a_n) + (a_n - 1) E_q[log(\alpha_n)] - b_n E_q[\alpha_n]$$

Note that $E_q[log(\alpha_n)] = \Psi(\tau_n) - \log(\upsilon_n)$ and $E_q[\alpha_n] = \frac{\tau_n}{\upsilon_n}$; both of which can be considered as constants in the maximization. But since there is exactly only one sample, one cannot learn the a_n, b_n effectively [2].

One way to overcome this, is to assume that all the α_n's are commonly drawn from a single a, b. This enables a larger number of samples to succesfully estimate the the single a, b. The downside is

that, this commonly shrinks all the α_n's to the expected value of the distribution $\frac{a}{b}$, which might be not a good thing.

We conducted several preliminary experiments were we tried sharing all α_y's under single a, b as well sharing the α_y's of sinling nodes etc. None of the models seemed to achieve competitive performance. For example, using a common a, b on the best performing model $M3$ on the CLEF dataset achieved a Micro-F_1 of of 80.32 and Macro-F_1 of 54.82 both of which are lower than M3-var. Further investigation is required to establish how empirical can be succesfully applied.

References

[1] G. Bouchard. Efficient bounds for the softmax function. 2007.

[2] George Casella. Empirical bayes method - a tutorial. Technical report.