Supplementary Materials

to "Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknokwn Expertise", by Jacob Whitehill, Paul Ruvolo, Tingfan Wu, Jacob Bergsma, and Javier Movellan

1 Full EM Derivation

Recall the probability of correct image label given the labeler's ability α_i and the image's difficulty parameter β_j:

$$p(L_{ij} = Z_j|\alpha_i, \beta_j) = \frac{1}{1 + e^{-\alpha_i \beta_j}} \tag{1}$$

The observed labels are samples from the $\{L_{ij}\}$ random variables. The unobserved variables are the true image labels Z_j, the different labeler accuracies α_i, and the image difficulty parameters $1/\beta_j$. Our goal is to efficiently search for the most probable values of the unobservable variables $\mathbf{Z}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ given the observed data. Here we can use Expectation-Maximization approach (EM) to obtain maximum likelihood estimates of the parameters of interest:

E step: Let the set of all given labels for an image j be denoted as $\mathbf{l}_j = \{l_{ij'} \mid j' = j\}$. Note that not every labeler must label every single image. In this case, the index variable i in $l_{ij'}$ refers only to those labelers who labeled image j. We need to compute the posterior probabilities of all $z_j \in \{0, 1\}$ given the $\boldsymbol{\alpha}, \boldsymbol{\beta}$ values from the last M step and the observed labels:

$$\begin{aligned} p(z_j|\mathbf{l}, \boldsymbol{\alpha}, \boldsymbol{\beta}) &= p(z_j|\mathbf{l}_j, \boldsymbol{\alpha}, \beta_j) \\ &\propto p(z_j|\boldsymbol{\alpha}, \beta_j)p(\mathbf{l}_j|z_j, \boldsymbol{\alpha}, \beta_j) \\ &\propto p(z_j)\prod_i p(l_{ij}|z_j, \alpha_i, \beta_j) \end{aligned}$$

where we noted that $p(z_j|\boldsymbol{\alpha}, \beta_j) = p(z_j)$ using the conditional independence assumptions from the graphical model.

M step: We maximize the auxiliary function Q, which is defined as the expectation of the joint log-likelihood of the observed and hidden variables $(\mathbf{l}, \mathbf{Z})$ given the parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta})$, w.r.t. the posterior probabilities of the $\mathbf{Z}$ values computed during the last E step:

$$\begin{aligned} &Q(\boldsymbol{\alpha}, \boldsymbol{\beta}) \\ &= E\left[\ln p(\mathbf{l}, \mathbf{z}|\boldsymbol{\alpha}, \boldsymbol{\beta})\right] \\ &= E\left[\ln \prod_j \left(p(z_j)\prod_i p(l_{ij}|z_j, \alpha_i, \beta_j)\right)\right] \\ &\quad \text{since } l_{ij} \text{ are cond. indep. given } \mathbf{z}, \boldsymbol{\alpha}, \boldsymbol{\beta} \\ &= \sum_j E\left[\ln p(z_j) + \sum_i \ln p(l_{ij}|z_j, \alpha_i, \beta_j)\right] \\ &= \sum_j E\left[\ln p(z_j)\right] + \sum_{ij} E\left[\ln p(l_{ij}|z_j, \alpha_i, \beta_j)\right] \end{aligned}$$

where the expectation is taken over $\mathbf{z}$ given the old parameter values $\boldsymbol{\alpha}^{old}, \boldsymbol{\beta}^{old}$ as estimated during the last

E-step. Let us define $p^k = p(z_j = k|\mathbf{l}, \boldsymbol{\alpha}^{old}, \boldsymbol{\beta}^{old})$. Then we can expand this expectation as:

$$\begin{aligned} & Q(\boldsymbol{\alpha}, \boldsymbol{\beta}) \\ = & \sum_j \sum_{k=0}^{1} p^k \ln p(z_j = k) + \\ & \sum_{ij} \sum_{k=0}^{1} p^k \ln p(l_{ij}|z_j = k, \alpha_i, \beta_j) \end{aligned}$$

Based on Equation (1), we can compute $p(l_{ij}|z_j = k, \alpha_i, \beta_j)$ as:

$$p(l_{ij}|z_j = 1, \alpha_i, \beta_j) = \sigma(\alpha_i\beta_j)^{l_{ij}}(1 - \sigma(\alpha_i\beta_j))^{1-l_{ij}}$$

and

$$p(l_{ij}|z_j = 0, \alpha_i, \beta_j) = \sigma(\alpha_i\beta_j)^{1-l_{ij}}(1 - \sigma(\alpha_i\beta_j))^{l_{ij}}$$

where $\sigma(x) = 1/(1 + e^{-x})$ is the logistic function. To avoid clutter, we will represent $\sigma(\alpha_i\beta_j)$ simply as σ. Then, after expanding the summation over k into the two cases $z = 0$ and $z = 1$, we get:

$$\begin{aligned} Q(\boldsymbol{\alpha}, \boldsymbol{\beta}) = & \sum_j \left(p^1 \ln p(z_j = 1) + p^0 \ln p(z_j = 0)\right) + \\ & \sum_{ij} p^1 \left[l_{ij} \ln \sigma + (1 - l_{ij}) \ln(1 - \sigma)\right] + \\ & \sum_{ij} p^0 \left[(1 - l_{ij}) \ln \sigma + l_{ij} \ln(1 - \sigma)\right] \end{aligned}$$

Taking the first derivatives causes the first summation to vanish since it is constant w.r.t $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Using the fact that

$$\frac{d}{dx}\sigma(x) = \sigma(x)(1 - \sigma(x))$$

we can differentiate Q to arrive at:

$$\begin{aligned} \frac{\partial Q}{\partial \alpha_i} = & \sum_j p^1(l_{ij}(1 - \sigma)\beta_j - (1 - l_{ij})\sigma\beta_j) + \\ & \sum_j p^0((1 - l_{ij})(1 - \sigma)\beta_j - l_{ij}\sigma\beta_j) \\ = & \sum_j \left(p^1 l_{ij} + p^0(1 - l_{ij}) - (p^1 + p^0)\sigma\right)\beta_j \\ = & \sum_j \left(p^1 l_{ij} + p^0(1 - l_{ij}) - \sigma\right)\beta_j \\ & \text{since } p^0 + p^1 = 1 \end{aligned}$$

Similarly, we can derive:

$$\frac{\partial Q}{\partial \beta_j} = \sum_i \left(p^1 l_{ij} + p^0(1 - l_{ij}) - \sigma\right)\alpha_i$$

The gradient equation for $\frac{\partial Q}{\partial \alpha_i}$ has an intuitive interpretation: The first two terms compute the empirical probability of the given label l_{ij} being correct given posterior probabilities of Z_j from the previous E-Step.

The σ that is subtracted is the model's current estimate of the probability that l_{ij} is correct given the current estimate of the labeler's ability and image's difficulty. Hence, the likelihood function will locally increase by increasing the labeler ability α_i if the empirical estimate of the number of correct images labeled by labeler i (weighted by image difficulty) is greater than its previous belief of correctness (again, weighted by difficulty). Similar intuition applies to $\frac{\partial Q}{\partial \beta_j}$ with regards to image difficulty[1].

To find locally optimal values of the α and β parameter we set the gradient to zero. The resulting equations are non-linear and thus need to be solved using iterative methods.

2 Multi-class Inference Based on the GLAD Model

Here we briefly derive an optimal inference algorithm for the multi-class case. We assume there are K different choices $\{1, \ldots, K\}$ for each image label. We continue under the initial assumption of GLAD as described in the main paper, which is that the probability of correct labeling is

$$p(L_{ij} = k | z_j = k, \alpha_i, \beta_j) = \sigma(\alpha_i \beta_j)$$

where σ is the logistic function. For the multi-class case, we further assume uniform probability over all *incorrect* responses, i.e., for all $k' \neq k$,

$$p(L_{ij} = k' | z_j = k, \alpha_i, \beta_j) = \frac{1}{K-1}(1 - \sigma(\alpha_i \beta_j))$$

The M-step is exactly the same as for the two-class case, except now the posterior probabilities for Z_j must be calculated over K classes, not just 2. For the E-step, we must modify slightly the equations for probability of correctness and the auxiliary function: Then

$$p(l_{ij} | z_j = k, \alpha_i, \beta_j) = \sigma^{\delta(l_{ij}, k)} \left(\frac{1}{K-1}(1 - \sigma) \right)^{1 - \delta(l_{ij}, k)}$$

where $\delta(a, b)$ is the Kronecker delta function. For brevity we write $\delta(l_{ij}, k)$ simply as δ. Then we can define Q as

$$\begin{aligned}
Q &= \sum_j \sum_{k=1}^{K} p^k \ln p(z_j = k) + \sum_j \sum_{k=1}^{K} p^k \ln p(l_{ij} | z_j = k, \alpha_i, \beta_j) \\
\frac{\partial Q}{\partial \alpha_i} &= \sum_j \sum_{k=1}^{K} p^k \left[\delta(1 - \sigma)\beta_j - (1 - \delta)(\sigma \beta_j - \ln(K-1)) \right] \\
&= \sum_j \sum_{k=1}^{K} p^k \left[\delta \beta_j - \delta \sigma \beta_j - \sigma \beta_j + \delta \sigma \beta_j + \ln(K-1) - \delta \ln(K-1) \right] \\
&= \sum_j \sum_{k=1}^{K} p^k \left[(\delta - \sigma)\beta_j + (1 - \delta) \ln(K-1) \right] \\
\frac{\partial Q}{\partial \beta_j} &= \sum_j \sum_{k=1}^{K} p^k \left[(\delta - \sigma)\alpha_i + (1 - \delta) \ln(K-1) \right]
\end{aligned}$$

Similar to the derivation in the paper, $p^k(\delta - \sigma)$ is positive only if $l_{ij} = k$ and represents the difference between the prior belief that the labeler would answer correctly and the empirical correctness of his/her response, weighted by probability that the true label is k. The expression $\ln(K-1)$ is 0 for the two-class problem, and hence the derivation in this supplement reduces to the two-class solution as described in the paper.

[1]Keep in mind that larger β means easier images.