Supplementary Material of Analysis of Minimax Error Rate for Crowdsourcing and Its Application to Worker Cluster Model

Hideaki Imamura [1 2] **Issei Sato** [1 2] **Masashi Sugiyama** [2 1]

A. Appendix: Proof of Lemma 1

Lemma 1 (Fano's inequality Fano (1949)).
For any Markov chain $V \to X \to \hat{V}$, we have

$$h(P(\hat{V} \neq V)) + P(\hat{V} \neq V)(\log(|\mathcal{V}| - 1) \geq H(V|\hat{V}),$$

where

$$h(p) = -p \log p - (1-p) \log(1-p),$$

$\mathcal{V}$ is the set of possible value of V, and $H(V|\hat{V})$ is the entropy of V conditioned on $\hat{V}$.

Proof of Lemma 1.
This inequality follows by expanding the entropy in two different ways. Let E be the indicator random variable for the event that $\hat{V} \neq V$, that is, $E = 1$ if $\hat{V} \neq V$ and is 0 otherwise. Then we have

$$\begin{aligned} H(V, E|\hat{V}) &= H(V|E, \hat{V}) + H(E|\hat{V}) \\ &= P(E=1)H(V|E=1, \hat{V}) \\ &\quad + P(E=0)H(V|E=0, \hat{V}) + H(E|\hat{V}) \\ &= P(E=1)H(V|E=1, \hat{V}) + H(E|\hat{V}), \end{aligned}$$

where the last equation follows because there is no error, and V has no variability given $\hat{V}$ if $E = 0$. Expanding the entropy by the chain rule in a different order, we have

$$\begin{aligned} H(V, E|\hat{V}) &= H(V|\hat{V}) + H(E|\hat{V}, V) \\ &= H(V|\hat{V}), \end{aligned}$$

because E is perfectly determined by V and $\hat{V}$. Combining these two equations, we have

$$\begin{aligned} H(V|\hat{V}) &= P(E=1)H(V|E=1, \hat{V}) + H(E|\hat{V}) \\ &= P(\hat{V} \neq V)H(V|E=1, \hat{V}) + H(E|\hat{V}) \\ &\leq P(\hat{V} \neq V)\log(|\mathcal{V}| - 1) + H(E|\hat{V}) \\ &\leq P(\hat{V} \neq V)\log(|\mathcal{V}| - 1) + h(P(\hat{V} \neq V)), \end{aligned}$$

where the first inequality follows because V can take on at most $|\mathcal{V}| - 1$ values when there is an error, and the last inequality follows by $H(E|\hat{V}) \leq H(E) = h(P(\hat{V} \neq V))$. The proof is complete. □

B. Appendix: Detailed Derivation of Inference Algorithm

In this section, we show the detailed derivation of an inference algorithm for the proposed WC model.

B.1. Empirical Variational Inference

To estimate latent variables and parameters, we adopt the strategy of *empirical variational inference*.

We want to maximize a following marginal log likelihood or *model evidence*.

$$\log p(X|\alpha, \beta, \Lambda) = \log \int p(X, G, \pi|\rho, \tau, \Lambda) dG d\pi.$$

However it is impossible to optimize the RHS's integral analytically. Therefore, we will instead maximize the following evidence lower bound (ELBO).

$$\begin{aligned} &\log \int p(X, G, \pi|\rho, \tau, \Lambda) dG d\pi \\ &\geq \int q(G, \pi) \log \left(\frac{p(X, G, \pi|\rho, \tau, \Lambda)}{q(G, \pi)} \right) dG d\pi \\ &:= ELBO. \end{aligned}$$

We make an assumption of the mean field approximation for $q(G, \pi)$, that is,

$$q(G, \pi) = \left[\prod_{i=1}^{n} q(G_i) \right] \left[\prod_{j=1}^{n} q(\pi^j) \right].$$

[1]The University of Tokyo, Tokyo, Japan [2]RIKEN, Tokyo, Japan. Correspondence to: Hideaki Imamura <imamura@ms.k.u-tokyo.ac.jp>.

Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

On the basis on the variational method, maximizing ELBO yields the following expression.

$$
\begin{aligned}
q(G_i) &= Mutinomial(G_i|\hat{\theta}_i), \\
q(\pi^j) &= Mutinomial(f_\Lambda^{-1}(\pi^j)|\hat{\phi}_j),
\end{aligned}
$$

where $\hat{\theta} = \{\hat{\theta}_i\}_{i=1}^n, \hat{\phi} = \{\hat{\phi}_j\}_{j=1}^m$ are variational parameters. They satisfy the following eqautions.

$$
\hat{\theta}_{i,k} = \frac{\exp\left(\sum_{j=1}^m \sum_{k'=1}^K \sum_l^L \delta(X_{i,j}=k')\hat{\phi}_{j,l} \log \Lambda_{l,k,k'} + \log \rho_k\right)}{\sum_{k=1}^K \exp\left(\sum_{j=1}^m \sum_{k'=1}^K \sum_l^L \delta(X_{i,j}=k')\hat{\phi}_{j,l} \log \Lambda_{l,k,k'} + \log \rho_k\right)}, \tag{1}
$$

$$
\hat{\phi}_{j,l} = \frac{\exp\left(\sum_{i=1}^n \sum_{k'=1}^K \sum_k^K \delta(X_{i,j}=k')\hat{\theta}_{i,k} \log \Lambda_{l,k,k'} + \log \tau_l\right)}{\sum_{l=1}^L \exp\left(\sum_{i=1}^n \sum_{k'=1}^K \sum_k^K \delta(X_{i,j}=k')\hat{\theta}_{i,k} \log \Lambda_{l,k,k'} + \log \tau_l\right)}. \tag{2}
$$

We can calculate ELBO analytically as follows.

$$
\begin{aligned}
ELBO = \sum_{i=1}^n \sum_{j=1}^m \sum_{k=1}^K \sum_{k'=1}^K \sum_{l=1}^L \hat{\theta}_{i,k}\hat{\phi}_{j,l}\delta(X_{i,j} = k') \log \Lambda_{l,k,k'} \\
+ \sum_{i=1}^n \sum_{k=1}^K \hat{\theta}_{i,k}(\log \rho_k - \log \hat{\theta}_{i,k}) + \sum_{j=1}^m \sum_{l=1}^L \hat{\phi}_{j,l}(\log \tau_l - \log \hat{\phi}_{j,l}).
\end{aligned} \tag{3}
$$

We want to maximize ELBO with respect to $\hat{\theta}, \hat{\phi}, \Lambda, \rho$, and τ. Deriving the stationary condition for $\{\hat{\theta}, \hat{\phi}\}$, we get equations (1) and (2). Deriving the stationary condition for $\{\Lambda, \rho, \tau\}$, we get the following equations.

$$
\rho_k = \frac{1}{n} \sum_{i=1}^n \hat{\theta}_{i,k}, \tag{4}
$$

$$
\tau_l = \frac{1}{m} \sum_{j=1}^m \hat{\phi}_{j,l}, \tag{5}
$$

$$
\Lambda_{l,k,k'} = \frac{\sum_{i=1}^n \sum_{j=1}^m \hat{\theta}_{i,k}\hat{\phi}_{j,l}\delta(X_{i,j} = k')}{\sum_{i=1}^n \sum_{j=1}^m \sum_{k'=1}^K \hat{\theta}_{i,k}\hat{\phi}_{j,l}\delta(X_{i,j} = k')}. \tag{6}
$$

Using (1), (2) and (4)~(6) iteratively, we can get a local optimal solution of ELBO. We refer to the estimated values of G, π as $\hat{G}, \hat{\pi}$. From the above, it is possible to calculate the $\hat{G}, \hat{\pi}$ as follows.

$$
\hat{G}_i = \arg\max_k \hat{\theta}_{i,k}, \tag{7}
$$

$$
\hat{\pi}^j = \Lambda_{\ell_j}. \tag{8}
$$

where the group to which the j-th worker belong is written as ℓ_j and calculated as $\ell_j = \arg\max_l \hat{\phi}_{j,l}$.

Finally, we have derived Algorithm 1.

B.2. Initialization Process

In this section, we discuss the initialization process of step [1] in Algorithm 1. We need to initialize $\hat{\theta}, \hat{\phi}, \rho, \tau$, and Λ.

The initialization process of ρ and τ is simple. We use the equations (4) and (5) to initialize ρ and τ. However, to initialize ρ and τ, we need $\hat{\theta}$ and $\hat{\phi}$. Thus, we need to initialize $\hat{\theta}$ and $\hat{\phi}$ first.

Algorithm 1 Empirical Variational Inference for Worker Clustering Model

Input: X
Output: $\hat{G}, \hat{\pi}$
1 Initialize $\hat{\theta}, \hat{\phi}, \rho, \tau, \Lambda$ appropriately (discussed later).
2 Update variational parameters by using (1) and (2)
3 Update hyper parameters by using (4)~(6)
4 Calculate ELBO (3), then compare it with previous ELBO value. If the inclement is smaller than the threshold, go to the step [5], otherwise repeat from step [2]
5 Calculate estimated values $\hat{G}, \hat{\pi}$ by using (7) and (8)

The initialization process of $\hat{\theta}$ is the same as that ofDawid & Skene (1979) given by

$$
\hat{\theta}_{i,k} = \frac{\sum_{j=1}^m \delta(X_{i,j} = k)}{\sum_{k=1}^K \sum_{j=1}^m \delta(X_{i,j} = k)}.
$$

The initialization process of $\hat{\phi}$ and Λ is a bit more complicated. First, we approximately calculate confusion matrices $\pi = \{\pi^j\}_{j=1}^m$ of each worker j as done by Dawid & Skene (1979) as follows.

$$
\pi^j_{k,k'} = \frac{\sum_{i=1}^n \hat{\theta}_{i,k}\delta(X_{i,j} = k')}{\sum_{k'=1}^K \sum_{i=1}^n \hat{\theta}_{i,k}\delta(X_{i,j} = k')}.
$$

We want to separate all m workers into L groups. The ability of each group $l \in \{1, 2, ..., L\}$ is measured by Λ_l. The bigger the diagonal components of each Λ_l are, the higher the accuracy of group l is. Thus, we first calculate the trace norm of each π^j, namely $\|\pi^j\|_*$ and then rearrange all workers in descending order according to $\{\|\pi^j\|_*\}_{j=1}^m$ to obtain the permutation of workers $\sigma = \begin{pmatrix} 1 & 2 & \dots & m \\ \sigma(1) & \sigma(2) & \dots & \sigma(m) \end{pmatrix}$, where $\sigma(j') = j$ means that the trace norm of the worker j is the j'-th largest. Second, we separate $\{\sigma(j')\}_{j'=1}^m$ into L groups. In other words, let $J_l = \left\{\sigma\left(\frac{m(l-1)}{L} + 1\right), ..., \sigma\left(\frac{ml}{L}\right)\right\}$. Using this set, the initial values of Λ and ϕ are below.

$$
\begin{aligned}
\Lambda_{l,k,k'} &= \frac{L}{m} \sum_{j \in J_l} \pi^j{}_{k,k'}, \\
\phi_{j,l} &= \delta(j \in J_l).
\end{aligned}
$$

References

Dawid, A. and Skene, A. Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied Statistics*, 28(1):20–28, 1979.

Fano, R. M. *Transmission of Information*. M.I.T. Press, 1949.