Literature Review of "Spectral Methods Meet EM: A Provably Optimal Method for Crowdsourcing"

Liam Collins

January 14, 2019

Abstract

Crowdsourcing can be a low cost means of collecting labels, but recovering the true labels from noisy crowdsourced labels is difficult task. Fortunately, recent work has given a provably optimal algorithm for solving this problem. In a 2014 paper titled "Spectral Methods meet EM: A Provably Optimal Method for Crowdsourcing" [5], authors Zhang et al. present a two-stage algorithm to recover the true labels and show that it achieves the optimal convergence rate up to a logarithmic factor. I will summarize their algorithm as well as their theoretical and empirical results here.

1 Introduction

Crowdsourcing is an efficient method solve labelling problems, and has grown in popularity with the rise of online services such as Amazon Mechanical Turk [5]. However, crowdsourcing workers are in general non-experts and potentially adversarial. The labels they generate are therefore noisy, which naturally leads to the problem of recovering the true labels from many noisy ones, known as the crowdsourcing problem.

Zhang, Chen, Zhou and Jordan presented several key contributions to crowdsourcing research in their 2014 paper "Spectral Methods meet EM: A Provably Optimal Method for Crowdsourcing" [5]. They propose an efficient two-stage algorithm to recover the true labels, as well as worker confusion matrices, that uses spectral methods to find initial parameters, then executes an adaptation of the Expectation Maximization (EM) procedure first introduced by Dawid and Skene [2] to estimate the true labels and worker confusion matrices. The (l, c)-th entry in each worker's confusion matrix represents the probability that the worker labels an item in class l as class c. Importantly, Zhang et al. establish bounds on algorithm performance and show that their method achieves minimax convergence rates up to a logarithmic factor under mild conditions, which is an improvement over prior algorithms.

Each section of this paper is a summary of a corresponding 1 or 2 sections in [5]. In Section 2 I briefly overview related work and in Section 3 I formulate the crowdsourcing problem. Section 4 contains an explanation of the two-stage algorithm and Section 5 gives convergence analysis. I conclude by briefly sharing empirical results in Section 6.

2 Related Work

The EM-based algorithm and the generative model involving confusion matrices developed by Dawid and Skene [2] have inspired much of the work in crowdsourcing. Most crowdsourcing algorithms (see for example [4, 3, 1]) are based on their generative model, including the spectral initialized-EM algorithm discussed here. However, including Dawid and Skene's EM algorithm, none of these algorithms have been shown to achieve optimal rates of convergence under reasonable conditions, besides the spectral initialized-EM algorithm.

The two-stage algorithm with spectral initialization draws from recent work that uses spectral methods to estimate latent variable models. As we will see, these spectral methods involve computing second and third-order empirical moments from the data and performing a tensor factorization of these moments using the tensor power method to obtain underlying variable estimates.

3 Problem Formulation

Assume that there are m workers, n items, and k classes. The true label y_j of item j is sampled from a probability distribution $\mathbb{P}[y_j = l] = w_l$. Let $z_{ij} \in \mathbb{R}^k$ be the label that worker i assigns to item j, where $z_{ij} = e_c$ if the label is class c and $z_{ij} = \mathbf{0}$ if the worker does not provide a label for that item. Here, e_c is the c-th canonical basis vector of $\mathbb{R}^k$. Denote π_i as the probability that worker i labels a randomly chosen item. The goal is to estimate the true labels $\{y_j : j \in [n]\}$ from the observed labels $\{z_{ij} : i \in [m], j \in [n]\}$.

Consistent with the work of Dawid and Skene, Zhang et al. assume that the probability that worker i labels an item in class l as class c does not depend on any particular item. Denote this probability μ_{ilc}, and let $\mu_{il} = [\mu_{il1}\, \mu_{il2} \ldots \mu_{ilk}]^T$. Then $C_i = [\mu_{i1}\, \mu_{i2} \ldots \mu_{ik}] \in \mathbb{R}^{k\times k}$ is the confusion matrix for worker i.

4 Two-Stage Algorithm

```
Algorithm 1: Spectral Initialization
  Input: integer $k$, observed labels $z_{ij} \in \mathbb{R}^k$ for $i \in [m]$ and $j \in [n]$
  Output: confusion matrices $\widehat{C}_i \in \mathbb{R}^{k\times k}$ for $i \in [m]$
  (1) Partition the workers into 3 distinct and non-empty groups $G_1, G_2$, and $G_3$ and compute the
      group aggregated labels $Z_{gj}$ by Eq. (1).
  (2) For $(a, b, c) \in \{(1, 2, 3), (2, 3, 1), (3, 1, 2)\}$ compute the second and third order moments
      $\widehat{M}_2 \in \mathbb{R}^{k\times k}$, $\widehat{M}_3 \in \mathbb{R}^{k\times k\times k}$ by Eq. (2), then compute each of the columns of $\widehat{C}_c^\diamond \in \mathbb{R}^{k\times k}$ and the
      diagonal entries of $\widehat{W} \in \mathbb{R}^{k\times k}$ via tensor factorization using the tensor power method.
  (3) Compute each $\widehat{C}_i$ by Eq. (3).
```

The first stage of the two-stage algorithm obtains initial estimates of the confusion matrices using spectral methods. See Algorithm 1 for a summary of this stage. For notation clarification, for any of the three distinct and non-empty groups of workers G_g indexed by $g \in \{1, 2, 3\}$, and $j \in [n]$, let

$$Z_{gj} := \frac{1}{\mid G_g \mid} \sum_{i \in G_g} z_{ij}, \quad \mu_{gl}^\diamond := \frac{1}{\mid G_g \mid} \sum_{i \in G_g} \pi_i \mu_{il} \tag{1}$$

where Z_{gj} is the average labelling within each group, and $\mu_{gl}^\diamond$ represent the group-aggregated confusion matrix columns. Next, the authors cite a result stating that the second and third order moments of the observed labels are equal to

$$M_2 := \textstyle\sum_{l=1}^k w_l \mu_{cl}^\diamond \otimes \mu_{cl}^\diamond, \quad M_3 := \sum_{l=1}^k w_l \mu_{cl}^\diamond \otimes \mu_{cl}^\diamond \otimes \mu_{cl}^\diamond \tag{2}$$

These expressions imply that the class probability estimates $\widehat{w}_k$ and the group aggregate confusion matrix column estimates $\widehat{\mu}_{gl}^\diamond$ can be factored from the empirical moments $\widehat{M}_2$ and $\widehat{M}_3$ using the tensor power method. The empirical moments are computed as functions of the observed labels according to Eqns. (2a-d) in Zhang et al. paper. Then, the confusion matrix estimate $\widehat{C}_i$ for each worker $i \in [m]$ can be computed according to

$$\widehat{C}_i := \text{normalize} \left\{ (\frac{1}{n} \sum_{j=1}^n z_{ij} Z_{aj}^T)(\widehat{W}(\widehat{C}_a^\diamond)^T)^{-1} \right\} \tag{3}$$

where a is one of the two groups that i does not belong to, $\widehat{W} = \text{diag}(\widehat{w}_1, \widehat{w}_2, \ldots. \widehat{w}_k)$ and is averaged across the three groups, and the normalization operator rescales the columns such that they sum to 1.

The next stage optimizes the log likelihood function of the observed and true labels by iteratively updating estimates of the confusion matrices using Dawid and Skene's EM procedure for one iteration or until convergence. It takes as input confusion matrix estimates and on each iteration, executes two steps. First, it computes the intermediate values $\widehat{q}_{jl}$ useful for computing the expected log likelihood of the true and observed labels:

$$\widehat{q}_{jl} = \frac{\exp\left(\sum_{i=1}^{m}\sum_{c=1}^{k}\mathbb{I}(z_{ij}=e_c)\log(\widehat{\mu}_{ilc})\right)}{\sum_{l'=1}^{k}\exp\left(\sum_{i=1}^{m}\sum_{c=1}^{k}\mathbb{I}(z_{ij}=e_c)\log(\widehat{\mu}_{ilc})\right)} \qquad \text{for} \quad j\in[n], l\in[k] \tag{4}$$

Then, it updates the confusion matrices to maximize the expected log likelihood:

$$\widehat{\mu}_{ilc} = \frac{\sum_{j=1}^{n}\widehat{q}_{jl}\mathbb{I}(z_{ij}=e_c)}{\sum_{l'=1}^{k}\sum_{j=1}^{n}\widehat{q}_{jl}\mathbb{I}(z_{ij}=e_c)} \qquad \text{for} \quad j\in[n], l\in[k], c\in[k] \tag{5}$$

After the final iteration, the predicted labels $\widehat{y}_j$ are computed by $y_j = \arg\max_{l\in[k]}\widehat{q}_{jl}$ for the most recently updated $\widehat{q}_{jl}$. Since the likelihood function is not concave, the EM procedure may converge to a local maximum depending on its initialization. However, the authors show in the next section that the EM procedure is guaranteed to converge to an optimal solution when it is initialized using the spectral technique discussed previously.

5 Convergence Analysis

The authors provide novel theoretical guarantees for both their spectral initialization and the EM algorithm under mild assumptions. We can combine their Theorems 1 and 2 to obtain results about the convergence of the the full two-stage algorithm. First define

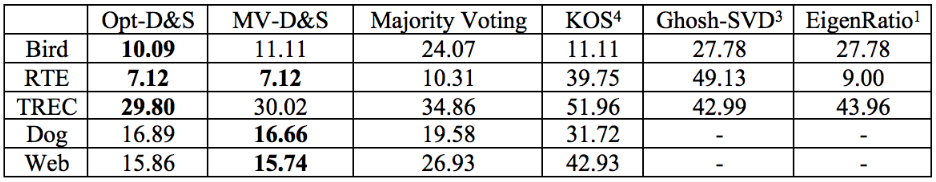
$$\begin{aligned} &w_{\min} := \min\{w_l\}_{l=1}^{k}, \quad \pi_{\min} := \min\{\pi_i\}_{i=1}^{m}, \quad \text{and} \\ &\overline{D} := \min_{l\neq l'}\frac{1}{m}\textstyle\sum_{i=1}^{m}\pi_i\mathbb{D}_{KL}(\mu_{il},\mu_{il'}) \end{aligned} \tag{6}$$

where $\mathbb{D}_{KL}(P,Q)$ is the KL-divergence between probability distributions P and Q. Furthermore, let σ_L be the smallest singular value of a matrix that is a product of the aggregate confusion matrices of any pair of groups. Now, assuming $\mu_{ilc}\geq\rho \quad \forall(i,l,c)\in[m]\times[k]^2$ for some $\rho>0$, if we choose the number of workers m and the number of items n such that

$$m = \tilde{\Omega}\Big(\frac{1}{\overline{D}}\Big), \quad n = \tilde{\Omega}\Big(\frac{k^5}{\pi_{\min}^2 w_{\min}^2\sigma_L^{13}\min\{\rho^2,(\rho\overline{D})^2\}}\Big) \tag{7}$$

then with high probability the spectral-initialized EM algorithm will perfectly recover the true labels and the estimator $\widehat{\mu}$ will be bounded as $\|\widehat{\mu}_{il}-\mu_{il}\|\leq\mathcal{O}(1/(\pi_i w_l n))$. Next, Theorem 3 shows that this convergence rate is optimal; specifically, the number of workers required for perfect recovery meets the lower bound for any number of items, and the accuracy of the confusion matrix estimation meets the lower bound for any number of workers and items. All other previously existing algorithms either have more expensive costs of convergence or make more restrictive assumptions.

6 Experimental Results



	Opt-D&S	MV-D&S	Majority Voting	KOS[4]	Ghosh-SVD[3]	EigenRatio[1]
Bird	**10.09**	11.11	24.07	11.11	27.78	27.78
RTE	**7.12**	**7.12**	10.31	39.75	49.13	9.00
TREC	**29.80**	30.02	34.86	51.96	42.99	43.96
Dog	16.89	**16.66**	19.58	31.72	-	-
Web	15.86	**15.74**	26.93	42.93	-	-

Table 1. Label prediction error rate (%) for real data. The D&S EM algorithms are iterated until convergence.

Zhang et al. provide experimental results that compare their algorithm's performance on five real data sets to five other algorithms. ”Opt-D&S” refers to their algorithm, and ”MV-D&S” refers to another algorithm that also uses D&S EM but is initialized with majority voting. As we observe in Table 1, these two algorithms consistently outperform the others, likely because the others rely on unrealistic assumptions about the data. The authors also obtain empirical results showing that Opt-D&S outperforms MV-D&S when an appropriate lower threshold is placed on the spectral estimates of each element of the confusion matrices.

References

[1] N. Dalvi, A. Dasgupta, R. Kumar, , and V. Rastogi. Aggregating crowdsourced binary ratings. *Proceedings of World Wide Web Conference*, 2013.

[2] A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society*, 1979.

[3] A. Ghosh, S. Kale, and P. McAfee. Who moderates the moderators? crowdsourcing abuse detection in user-generated content. *Proceedings of the ACM Conference on Electronic Commerce*, 2011.

[4] D. R. Karger, S. Oh, , and D. Shah. Budget-optimal task allocation for reliable crowdsourcing systems. *Operations Research*, 2014.

[5] Y. Zhang, X. Chen, D. Zhou, and M. Jordan. Spectral methods meet em: A provably optimal algorithm for crowdsourcing. *Advances in Neural Information Processing Systems*, 2014.