Spectral Methods Meet EM: A Provably Optimal Algorithm for Crowdsourcing

Yuchen Zhang[⋆] **Xi Chen**[†] **Dengyong Zhou**[‡] **Michael I. Jordan**[⋆]

[⋆]University of California, Berkeley [†]New York University [‡]Microsoft Research

Problem Formulation

Given:

- m workers and n items, each worker labels a subset of items.
- Workers have distinct labeling accuracies.

Goal:

- Infer the true item labels.
- Infer the accuracy of workers.

An Example

Given labels, one possible inference result:

	Wkr 1	Wkr 2	Wkr 3	Wkr 4	Wkr 5	True
item 1	A	B	B	A	B	B
item 2	A	B	A			A
item 3	B	A	A	B		A
item 4		B	B	B	B	B
item 5	A				B	B
Accuracy	25%	75%	100%	33%	100%	–

Dawid & Skene Model

Dawid & Skene propose a generative model:

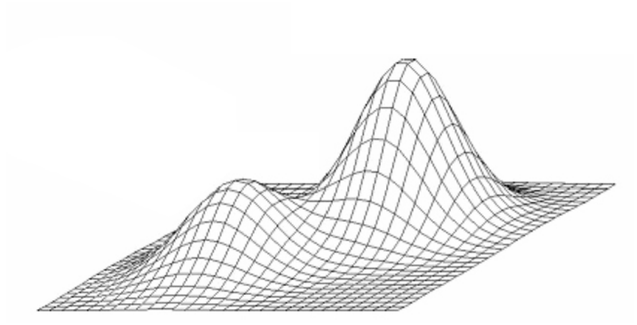
$$\mathbb{P}(\text{worker } i \text{ assigns label } j \mid \text{true label} = y) = \mu_{ijy}$$

EM Algorithm: Estimate the true label y and the confusion matrix μ to maximize the likelihood of observation.

1. Initialize label estimate $\widehat{y}$ by majority voting.
2. based on (1): update confusion matrix $\widehat{\mu}$ to maximize the likelihood.
3. based on (2): re-udpate label $\widehat{y}$ to maximize the likelihood.
4. repeat (2) and (3) until converge.

Global Maximum vs Local Maximum

Fact: $(\widehat{y}, \widehat{\mu})$ is the optimal estimate if they are the global likelihood maximum (Gao and Zhou, 2013).

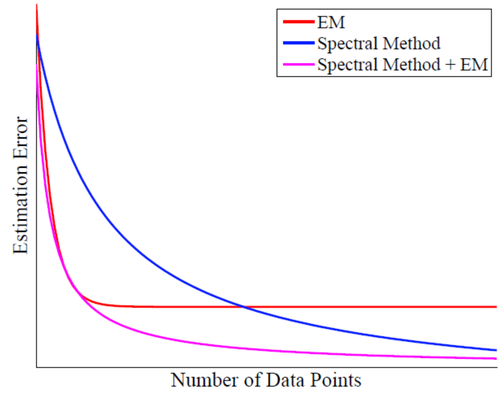


EM Algorithm: converge to a local maximum of the likelihood function.

Our Algorithm

Stage 1: Initialize confusion matrix estimate $\widehat{\mu}$ by a new algorithm based on spectral method.

Stage 2: Taking the initialization of Stage 1, re-update $\widehat{y}$ and $\widehat{\mu}$ using EM algorithm for one or more iterations.



Spectral + EM ⇒ optimal convergence rate.

Theory

Quantities measuring the ability of workers:

$$\pi_i = P(\text{worker } i \text{ labels an item})$$

$$\overline{D} = \min_{y \neq y'} \frac{1}{m} \sum_{i=1}^{m} \pi_i \mathbb{D}_{\mathrm{KL}}(\mu_{iy}, \mu_{iy'})$$

Theorem 1 *If there are $\widetilde{\Omega}(1/\overline{D})$ workers, then with high probability our algorithm provides $\widehat{y} = y$. Otherwise, any algorithm suffers $P[\widehat{y} \neq y] \geq 1/4$.*

Theorem 2 *Our algorithm estimates worker accuracies as $\mathbb{E}[\|\widehat{\mu}_{iy} - \mu_{iy}\|_2^2] \lesssim \frac{1}{w_y \pi_i n}$. On the other hand, any algorithm has $\mathbb{E}[\|\widehat{\mu}_{iy} - \mu_{iy}\|_2^2] \gtrsim \frac{1}{w_y \pi_i n}$.*

Experiments

Five real datasets of different scales:

Dataset name	# classes	# items	# workers	# worker labels
Bird	2	108	39	4,212
RTE	2	800	164	8,000
TREC	2	19,033	762	88,385
Dog	4	807	52	7,354
Web	5	2,665	177	15,567

Compare with Dawid & Skene estimator and recently proposed algorithms on label prediction error (%):

	Spectral+EM	MV+EM	MV	KOS	Ghosh-SVD	EigenRatio
Bird	**10.09**	11.11	24.07	11.11	27.78	27.78
RTE	**6.88**	7.12	10.31	39.75	49.13	9.00
TREC	**29.80**	30.02	34.86	51.96	42.99	43.96
Dog	**15.24**	15.74	26.93	42.93	–	–
Web	**12.29**	16.66	19.58	31.72	–	–

(MV = majority voting; MV+EM has no theoretical guarantee)

Conclusion: Our algorithm is comparable with the best empirical approach, outperforming other recent algorithms.