# A permutation-based Model for Crowd Labeling

Optimal estimation and robustness

-Summary-

# A permutation-based model for crowd labeling
# - Model and problem formulation

**Notations**

Assume $n$ workers and $d$ questions having two possible answers (+1 or -1)

- $x^* \in \{-1, 1\}^d$ : correct answer vector to $d$ questions
- $Q^* \in [0, 1]^{n \times d}$: probability matrix (unknown)

  ($Q^*_{ij}$ : probability that worker $i$ answers question $j$ correctly)
- $Y_{ij} \in \{-1, 0, 1\}$ : response of worker $i$ to question $j$

  ($Y_{ij} = 0$ if worker $i$ is not asked question $j$)
- $p_{obs} \in [0, 1]$ : probability that worker $i$ asked question $j$

$\Rightarrow$ Goal : Given random matrix $Y$, estimate $x^*$

**Assumption**

1. Worker is never asked the same question twice
2. Given $x^*$ and $Q^*$, $Y'_{ij}s$ are mutually independent such that

$$Y_{ij} = \begin{cases} x_j^* & w.p \;\; p_{obs} Q^*_{ij} \\ -x_j^* & w.p \;\; p_{obs}\left(1 - Q^*_{ij}\right) \\ 0 & w.p \;\; (1 - p_{obs}) \end{cases}$$

# A permutation-based model for crowd labeling
# - Model and problem formulation

**Assumption (cont.)**

There exists two permutation $\pi^*: [n] \rightarrow [n]$ for workers, $\sigma^*: [d] \rightarrow [d]$ for questions such that the probability matrix $Q^*$ satisfies :

$$(\text{Worker monoton}icity)$$

$$if\ \pi^*(i) < \pi^*(i'), \qquad Q^*_{ij} \geq Q^*_{i'j} \text{ for every question j}$$

$$(Question \text{ monoton}icity)$$

$$if\ \sigma^*(j) < \sigma^*(j'), \qquad Q^*_{ij} \geq Q^*_{ij'} \text{ for every worker i}$$

**Conditions**

- (R1) $Q^*_{ij} \geq \frac{1}{2} \qquad \forall i \in [n],\ j \in [d]$
- (R2) $\mathrm{p}_{obs} \geq \frac{1}{n}$ and $d \geq n$

# A permutation-based model for crowd labeling
# - Model and problem formulation

**Definitions**

- $C_{perm} = \{Q \in [0,1]^{n \times d}$ | ∃ permutation $(\pi, \sigma)$ such that worker & question monotonicity hold}
- $C_{Int} = \{Q = \tilde{q}(1-h)^T + \frac{1}{2} 1 h^T \in C_{perm} \big| \, for\ some\ \tilde{q} \in [0,1]^n, h \in [0,1]^d\}$
- $C_{DS} = \{\mathrm{Q} \in C_{perm} \mid Q = q^{DS} 1^T$ for some $q^{DS} \in [0,1]^n\}$
- Estimator for answer vector $x^*$ : $\hat{x} : \; Y \to \{-1, 1\}^d$
- Hamming error : $d_H(\hat{x}, x^*) = \frac{1}{d}\sum_{j=1}^{d} 1\{\widehat{x_j} \neq x_j^*\}$
- $Q^*$-loss function: $L_{Q^*}(\hat{x}, x^*) = \frac{1}{\mathrm{d}}\sum_{j=1}^{d} 1\{\widehat{x_j} \neq x_j^*\}\, \Psi(Q_{1j}^*, \ldots, Q_{nj}^*)$

  (where $\Psi$: $[0,1]^n \to R_+$ is a function capturing difficulty of the task)

**Remark**

- $\Psi\left(Q_{1j}^*, \ldots, Q_{nj}^*\right) = \frac{1}{\mathrm{n}}\sum_{i=1}^{n}\left(2Q_{ij}^* - 1\right)^2$ for each task $j \in [d]$ (from collective intelligence)
- $L_{Q^*}(\hat{x}, x^*) = \frac{1}{d}\sum_{j=1}^{d}\left(1\{\widehat{x_j} \neq x_j^*\}\frac{1}{n}\sum_{i=1}^{n}\left(2Q_{ij}^* - 1\right)^2\right) = \frac{1}{dn}\left\|\left(Q^* - \frac{1}{2} 1 1^T\right) diag(\hat{x} - x^*)\right\|_F^2$

# A permutation-based model for crowd labeling
# - Least square estimator

**Least square estimator under permutation model**

$$(\tilde{x}_{LS}, \tilde{Q}_{LS}) \in \underset{x \in \{-1,1\}^d, Q \in C_{perm}}{\operatorname{argmin}} \left\| \frac{1}{p_{obs}} Y - (2Q - 11^T) diag(x) \right\|_F^2$$

**Theorem 1 (Performance guarantee for least square estimator)**

(a) For any binary vector $x^* \in \{-1, 1\}^d$ and any matrix $Q^* \in C_{perm}$, the least squares estimator $\tilde{x}_{LS}$ has error at most

$$L_{Q^*}(\tilde{x}_{LS}, x^*) \leq \frac{c_v}{np_{obs}} \log^2 d$$

with probability at least $1 - e^{-c_H d log(dn)}$

(b) There exists a matrix $\tilde{Q} \in C_{DS}$ such that any estimator $\hat{x}$ has error at least

$$\sup_{x^* \in \{-1,1\}^d} E[L_{\tilde{Q}}(\hat{x}, x^*)] \geq \frac{c_L}{np_{obs}}$$

# A permutation-based model for crowd labeling
# - Least square estimator

**Remark**

Let $W \in R^{n\times d}$ be a random matrix defined by :

$$W_{ij} = \begin{cases} 1 - p_{obs}\big(2Q^*_{ij} - 1\big)x^*_j & w.p \;\; p_{obs}\left(Q^*_{ij}\left(\dfrac{1+x^*_j}{2}\right) + \big(1 - Q^*_{ij}\big)\left(\dfrac{1-x^*_j}{2}\right)\right) \\ -1 - p_{obs}\big(2Q^*_{ij} - 1\big)x^*_j & w.p \;\; p_{obs}\left(Q^*_{ij}\left(\dfrac{1+x^*_j}{2}\right) + \big(1 - Q^*_{ij}\big)\left(\dfrac{1-x^*_j}{2}\right)\right) \\ -p_{obs}\big(2Q^*_{ij} - 1\big)x^*_j & w.p \;\; 1 - p_{obs} \end{cases}$$

Then, observed matrix $Y$ can be written in the following form :

$$\frac{1}{p_{obs}}Y = \big(2Q^* - 11^T\big)diag(x^*) + \frac{1}{p_{obs}}W$$

Since $E[Y] = p_{obs}\big(2Q^* - 11^T\big)diag(x^*)$, $(\tilde{x}_{LS}, \tilde{Q}_{LS})$ becomes minimizer for $\|Y - E[Y]\|_F^2$

# A permutation-based model for crowd labeling
# - Least square estimator

**Corollary 1**

(a) For any $x^* \in \{-1, 1\}^d$ and any $Q^* \in C_{perm}$, the least squares estimate $\tilde{Q}_{LS}$ has error at most

$$\frac{1}{dn}\left\|\tilde{Q}_{LS} - Q^*\right\|_F^2 \leq \frac{c_v}{np_{obs}} \log^2 d$$

with probability at least $1 - e^{-c_H dlog(dn)}$

(b) Conversely, for any answer vector $x^* \in \{-1, 1\}^d$, any estimator $\hat{Q}$ has error at least

$$\sup_{Q^* \in C_{perm}} E\left[\frac{1}{dn}\left\|\hat{Q} - Q^*\right\|_F^2\right] \geq \frac{c_L}{np_{obs}}$$

# A permutation-based model for crowd labeling - WAN estimator

**WAN estimator : when worker's ordering is approximately known($\pi$)**

Step 1 (Windowing) : Compute integer $k_{WAN}$

$$k_{WAN} \in \underset{k \in \left\{\frac{\log^{1.5}(dn)}{p_{obs}},\dots,n\right\}}{argmax} \sum_{j \in [d]} 1\{|\sum_{i \in [k]} Y_{\pi^{-1}(i)j}| \geq \sqrt{k p_{obs} \log^{1.5}(dn)}\}$$

Step 2 (Aggregating Naively) : Set $\hat{x}_{WAN}(\pi)$ as majority vote of the best $k_{WAN}$ workers

$$[\hat{x}_{WAN}(\pi)]_j \in \underset{b \in \{-1,1\}}{argmax} \sum_{i=1}^{k_{WAN}} 1\left\{Y_{\pi^{-1}(i)j} = b\right\} \; for\ every\ j \in [d]$$

# A permutation-based model for crowd labeling - WAN estimator

**Theorem 2 (Performance guarantee for WAN estimator)**

For any matrix $Q^* \in C_{perm}$ and any binary vector $x^* \in \{-1, 1\}^d$, suppose that the WAN estimator is provided with the permutation $\pi$ of workers.

Consider the subset of the questions given by

$$J = \{j \in [d] | \, \exists k_j \geq \frac{\log^{1.5}(dn)}{p_{obs}} \; s.t \sum_{i=1}^{k_j} \left( Q^*_{\pi^{-1}(i)j} - \frac{1}{2} \right) \geq \frac{3}{4} \sqrt{\frac{k_j}{p_{obs}} \log^{1.5}(dn)}$$

Then the WAN estimator correctly estimates the labels of all questions in set $J$ with high probability:

$$P\left( [\hat{x}_{WAN}(\pi)]_j = x^*_j \; for \; all \; j \in J \right) \geq 1 - e^{-c_H \log^{1.5}(dn)}$$

**Notation**

- $Q^*_j$ = $j$th column of $Q^*$
- $Q^\pi_j$ = vector obtained by permuting the entries of $Q^*_j$ using $\pi$ (approximate ordering)
- $Q^{\pi^*}_j$ = vector obtained by permuting the entries of $Q^*_j$ using $\pi^*$ (correct ordering)

# A permutation-based model for crowd labeling - WAN estimator

**Corollary 2**

For any matrix $Q^* \in C_{perm}$ and any binary vector $x^* \in \{-1, 1\}^d$, suppose that the WAN estimator is provided with the permutation $\pi$ of workers.

Then for every question $j \in [d]$ such that

$$\left\| Q_j^* - \frac{1}{2} \right\|_2^2 \geq \frac{5 \log^{2.5}(dn)}{p_{obs}}, \qquad and\ \left\| Q_j^{\pi} - Q_j^{\pi^*} \right\|_2 \leq \frac{\left\| Q_j^* - \frac{1}{2} \right\|_2}{\sqrt{9 \log(dn)}}$$

We have

$$P\left( [\hat{x}_{WAN}(\pi)]_j = x_j^*\ for\ all\ j \in J \right) \geq 1 - e^{-c_H \log^{1.5}(dn)}$$

Consequently, if $\pi$ is the correct permutation of the workers ($\pi = \pi^*$), then with probability at least $1 - e^{c_H' \log^{1.5}(dn)}$, we have

$$L_{Q^*}(\hat{x}_{WAN}(\pi), x^*) \leq \frac{c_v}{np_{obs}} \log^{2.5} d$$

# A permutation-based model for crowd labeling - OBI-WAN estimator

**OBI-WAN estimator : when worker's ordering$(\pi)$ is unknown**

Note: OBI-WAN = Ordering Based on Inner-products-WAN

***Step 0 (preliminary):***

1. Split $d$ question into two set $(T_0, T_1)$
2. Assign every questions to $T_0$ or $T_1$ uniformly at random.
3. Get $Y_0, Y_1$ : the submatrices of $Y$ containing columns of $Y$ associated to questions in $T_0, T_1$ respectively

***Step 1 (OBI):***

1. Get top left eigenvector of $Y_l Y_l^T$:

$$u_l \in \underset{\|u\|_2=1}{\operatorname{argmax}} \left\| Y_l^T u \right\|_2$$

(Choose sign of $u_l$ so that $\sum_{i\in[n]} [u_l]_i^2 1\{[u_l]_i > 0\} \geq \sum_{i\in[n]} [u_l]_i^2 1\{[u_l]_i < 0\}$)

2. Obtain $\pi_l$ : permutation of $n$ workers in order of the respective entries of $u_l$

***Step 2 (WAN): (To avoid violation of independence assumptions)***

Compute estimators:

$$\hat{x}_{OBI-WAN}(T_0) = \hat{x}_{WAN}(Y_0, \pi_1), \qquad \hat{x}_{OBI-WAN}(T_1) = \hat{x}_{WAN}(Y_1, \pi_0)$$

# A permutation-based model for crowd labeling - OBI-WAN estimator

**Guarantees for OBI-WAN under intermediate model**

- Introduce a parameter $h_j^* \in [0,1]$ that captures the difficulty of each question $j \in [d]$
- Use $\tilde{q}$ associated with the workers as in Dawid-Skene model (instead of $q_{DS}$)
- Under intermediate model

$$P\left(Y_{ij} = x_j^*\right) = \tilde{q}_i\left(1 - h_j^*\right) + \frac{1}{2}h_j^* \qquad \forall\ (i,j)\ such\ that\ Y_{ij} \neq 0$$

( = probability that worker $i$ correctly answers question $j$)

$$C_{Int} = \{Q = \tilde{q}(1-h)^T + \frac{1}{2}1h^T \in C_{perm} \middle|\ for\ some\ \tilde{q} \in [0,1]^n, h \in [0,1]^d\}$$

( = Intermediate model class)

- Obviously, $C_{DS} \subset C_{int} \subset C_{perm}$ holds

# A permutation-based model for crowd labeling - OBI-WAN estimator

**Remark : OBI-WAN under intermediate model**

- Guarantees for the OBI step : Is $u_l \approx \pi_l^*$(correct worker permutation) ?
- Guarantees for the WAN step : Is $u_l$ operates well on $T_{1-l}$?

**(Notations & Assumption)**

- $r^* = \tilde{q} - \frac{1}{2}$,
- $\tilde{r}_l = r^*$ permuted by ordering of $u_l (= \pi_l)$
- Assume $Q_j^* = Q_j^{\pi_l^*}$ => already ordered correctly ($\tilde{q}$ is correctly ordered)

# A permutation-based model for crowd labeling - OBI-WAN estimator

**Lemma 5**

Suppose $(\tilde{q}, h^*)$ satisfies $\left\|\tilde{q} - \frac{1}{2}\right\|_2^2 \|1 - h^*\|_2^2 \geq \frac{\tilde{c}d \log^{2.5}(dn)}{p_{obs}}$ for a large enough constant $\tilde{c}$. Then, with sufficiently high probability, we have

$$P\left(\|\tilde{r}_l - r^*\|_2^2 > \frac{\|r^*\|_2^2}{9\log(dn)}\right) \leq e^{-clog^{1.5}(d)}$$

[From Equation 41] (with high probability)

$$\Leftrightarrow \left\|Q_j^{\pi_l} - Q_j^*\right\|_2 \leq \frac{\left\|Q_j^*\right\|_2}{\sqrt{9\log(dn)}} \text{ for every question j} \in T_{1-l}$$

# A permutation-based model for crowd labeling
# - OBI-WAN estimator

**Theorem 3 (Performance guarantee for OBI-WAN estimator under intermediate model)**

Consider any binary vector $x^* \in \{-1, 1\}^d$ and any matrix $Q^* \in C_{int}$ associated with vectors $(\tilde{q}, h^*)$ satisfying $\left\|\tilde{q} - \frac{1}{2}\right\|_2^2 \|1 - h^*\|_2^2 \geq \frac{\tilde{c} d \log^{2.5}(dn)}{p_{obs}}$ for a large enough constant $\tilde{c}$. Then for every question $j \in [d]$ such that

$$\left(1 - h_j^*\right)^2 \left\|\tilde{q} - \frac{1}{2}\right\|_2^2 \geq \frac{5 \log^{2.5}(dn)}{p_{obs}},$$

We have

$$P\left([\hat{x}_{OBI-WAN}]_j = x_j^*\right) \geq 1 - e^{-c_H \log^{1.5}(dn)}$$

**Corollary 3**

For any $Q^* \in C_{int}$ and any vector $x^* \in \{-1, 1\}^d$, the estimate $\hat{x}_{OBI-WAN}$ has error at most

$$L_{Q^*}(\hat{x}_{OBI-WAN}, x^*) \leq \frac{c_v}{n p_{obs}} \log^{2.5} d$$

With probability at least $1 - e^{-c_H \log^{1.5}(dn)}$

# A permutation-based model for crowd labeling - OBI-WAN estimator

**Guarantees for OBI-WAN under Dawid-Skene model**

- Handle adversarial workers : ignore (R1) condition (i.e : $q_i^{DS} < \frac{1}{2}$ is possible)

  [Recall : (R1) $Q_{ij}^* \geq \frac{1}{2} \qquad \forall i \in [n],\ j \in [d], Q^* = q^{DS}1^T$ in Dawid-Skene model]

- Define two associated vectors $q^{DS+},\ q^{DS-} \in [0,1]^n$

$$q_i^{DS+} = \max\left\{q_i^{DS}, \frac{1}{2}\right\}$$

$$q_i^{DS-} = \min\left\{q_i^{DS}, \frac{1}{2}\right\}$$

# A permutation-based model for crowd labeling
# - OBI-WAN estimator

**Theorem 4 (Performance guarantee for OBI-WAN estimator under Dawid-Skene model)**

Consider any Dawid-Skene matrix of the form $Q^* = q^{DS}1^T$ for some $q^{DS} \in [0,1]^n$. Then:

(a) If $\left\| q^{DS+} - \frac{1}{2} \right\|_2 \geq \left\| q^{DS-} - \frac{1}{2} \right\|_2 + \sqrt{\frac{4 \log^{2.5}(dn)}{p_{obs}}}$ and $\left( q^{DS} - \frac{1}{2} \right)^T 1 \geq 0$, then for any $x^* \in \{-1, 1\}^d$,

the OBI-WAN estimator satisfies :

$$P(\hat{x}_{OBI-WAN} = x^*) \geq 1 - e^{-c_H \log^{1.5}(dn)}$$

(b) Conversely, there exists a positive universal constant $c$ such that for any $q^{DS} \in \left[\frac{1}{10}, \frac{9}{10}\right]^n$ with $\left\| q^{DS} - \frac{1}{2} \right\|_2 \leq \sqrt{\frac{c}{p_{obs}}}$, any

estimator $\hat{x}$ has (normalized) Hamming error at least

$$\sup_{x^* \in \{-1,1\}^d} E\left[\sum_{i=1}^{d} \frac{1}{d} 1\{\hat{x}_i \neq x_i^*\}\right] \geq \frac{1}{10}$$

# A permutation-based model for crowd labeling - OBI-WAN estimator

**Remark : OBI-WAN under Dawid-skene model**

- Guarantees for the OBI step : Is $u_l \approx \pi_l^*$(correct worker permutation) ?
- Guarantees for the WAN step : Is $u_l$ operates well on $T_{1-l}$

**(Notations & Assumption)**

- $r^* = q^{DS} - \frac{1}{2}$,
- $\tilde{r}_l = r^*$ permuted by ordering of $u_l (= \pi_l)$
- $q^{DS}$ is already ordered

**[From equation 50]**

Under Dawid-skene model, the following holds :

$$P\left(\|\tilde{r}_l - r^*\|_2^2 \leq \frac{\log^{1.5} d}{18 p_{obs}}\right) \geq 1 - e^{-c log^{1.5}(d)}$$

Guarantee for the WAN step => theorem 4 (a)

# A permutation-based model for crowd labeling
# - OBI-WAN estimator

### Guarantees for OBI-WAN under permutation-based model

- Previous two sections provided strong guarantees for OBI-WAN for exact recovery and the $Q^*$-loss for Dawid-skene and intermediate model.

### Proposition 1 (Performance guarantee for OBI-WAN estimator under permutation model)

Consider any matrix $Q^* \in \mathrm{C}_{perm}$ and any binary vector $x^* \in \{-1, 1\}^d$. For every question $j \in [d]$ such that $\sum_{i=1}^{n} \left(Q_{ij}^* - \frac{1}{2}\right) \geq \frac{3}{4}\sqrt{\frac{n \log^{1.5}(dn)}{p_{obs}}}$ , the OBI-WAN estimator satisfies

$$P\left([\hat{x}_{OBI-WAN}]_j = x_j^*\right) \geq 1 - e^{-c_H \log^{1.5}(dn)}$$

Consequently for any $Q^* \in C_{perm}$ and any $x^* \in \{-1, 1\}^d$, with probability at least $1 - c^{-c_H \log^{1.5}(dn)}$, the estimator incurs a $Q^*$-loss of at most

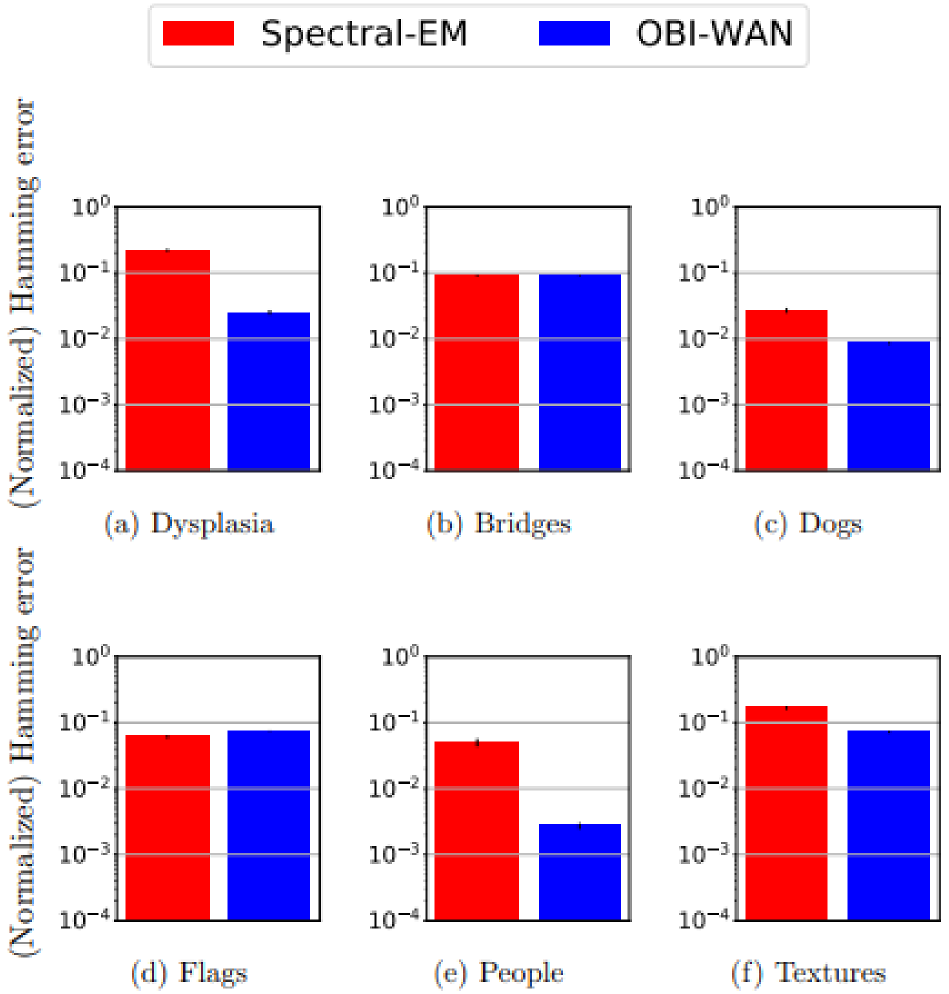
$$L_{Q^*}(\hat{x}_{OBI-WAN}, x^*) \leq \frac{c_v \log d}{\sqrt{n p_{obs}}}$$

# A permutation-based model for crowd labeling - Real world data / Simulation

**Real-world crowdsourcing data / Simulation**

- The experiments reveals that OBI-WAN compares favorably to Spectral-EM (All in all)
- Under synthetic simulation, OBI-WAN shows good performance under $C_{perm} \backslash C_{Int}$
- Under super sparse case, OBI-WAN incurs relatively higher error
- Shows limited performance under small $p_{obs}, n, d$

Real-world data




Simulation