Achieving budget-optimality with adaptive schemes in crowdsourcing

Sewoong Oh

Department of Industrial and Enterprise Systems Engineering
University of Illinois Urbana-Champaign

joint work with Ashish Khetan
(David Karger, Devavrat Shah, Jungseul Ok, Jinwoo Shin, Yung Yi)
Crowdsourcing systems

Marketplace to get labels for training data

airplanes?  
<table>
<thead>
<tr>
<th>Yes</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
</table>

Quality of the labels can be very low

70/100  30/100
Add redundancy to cope with noise
Add redundancy to cope with noise.
Add redundancy to cope with noise
Add redundancy to cope with noise
Add redundancy to cope with noise
Add redundancy to cope with noise
Add redundancy to cope with noise
Add redundancy to cope with noise
Add redundancy to cope with noise

Majority Voting
Add redundancy to cope with noise

Tradeoff: redundancy vs. accuracy

- special to crowdsourcing system:
  - we pay for each response
  - we design the graph
  - workers arrive online fashion
Notations

- \( m \) questions
- unknown true answers \( t \in \{-1, +1\}^m \)
- task-assignment graph \( G([m], [n], E) \)
  \((i, j) \in E\) indicates that question \( i \) is asked to the \( j \)-th arriving worker
- \( n \) workers arrive online fashion and submit responses
- response matrix \( A \in \{-1, 0, 1\}^{m \times n} \)
  \[ A_{ij} = \begin{cases} 
    0 & \text{if not assigned} \\
    +1 & \text{if worker answers “+1”} \\
    -1 & \text{if worker answers “−1”} 
  \end{cases} \]

- Dawid-Skene model from 1979:
  - Each worker is parametrized by a scalar value \( p_j \in [0, 1] \)
  - For each assigned question, answers correctly with probability \( p_j \)
    \[ A_{ij} = \begin{cases} 
    t_i & \text{with probability} \ p_j \\
    -t_i & \text{with probability} \ 1 - p_j 
  \end{cases} \]
  - Criticism: all tasks are assumed to be equally difficult
Design:

1. Task assignment graph $E$
2. Inference algorithm $\hat{t}(A) \in \{-1, +1\}^m$

How does error rate trade-off with cost/redundancy?

$$\mu \triangleq \frac{1}{n} \sum_{j=1}^{n} (2p_j - 1)$$

$$\sigma^2 \triangleq \frac{1}{n} \sum_{j=1}^{n} (2p_j - 1)^2$$

<table>
<thead>
<tr>
<th>$t_1$</th>
<th>$t_2$</th>
<th>$t_3$</th>
<th>$t_3$</th>
<th>$t_4$</th>
<th>$t_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_1$</td>
<td>$-1$</td>
<td>$+1$</td>
<td>$-1$</td>
<td>$-1$</td>
<td>$-1$</td>
</tr>
<tr>
<td>$p_2$</td>
<td>$+1$</td>
<td>$+1$</td>
<td>$-1$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p_3$</td>
<td></td>
<td></td>
<td>$+1$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p_4$</td>
<td></td>
<td></td>
<td></td>
<td>$+1$</td>
<td></td>
</tr>
<tr>
<td>$p_5$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$-1$</td>
</tr>
<tr>
<td>$p_6$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$-1$</td>
</tr>
</tbody>
</table>
Design:

1. Task assignment graph $E$
2. Inference algorithm $\hat{t}(A) \in \{-1, +1\}^m$

How does error rate trade-off with cost/redundancy?

$$\mu \triangleq \frac{1}{n} \sum_{j=1}^{n} (2p_j - 1)$$

$$\sigma^2 \triangleq \frac{1}{n} \sum_{j=1}^{n} (2p_j - 1)^2$$

- Majority Voting
  - $\sim e^{-c\mu^2 \ell}$
- Oracle Estimator
  - $\sim e^{-c'\sigma^2 \ell}$

Cost = Number of assignments per task $\ell$

Probability of error

$\ell$ answers per task
Design:

1. Task assignment graph $E$
2. Inference algorithm $\hat{t}(A) \in \{-1, +1\}^m$

How does error rate trade-off with cost/redundancy?

\[
\mu \triangleq \frac{1}{n} \sum_{j=1}^{n} (2p_j - 1)
\]

\[
\sigma^2 \triangleq \frac{1}{n} \sum_{j=1}^{n} (2p_j - 1)^2
\]
What is known for DS model [KargerOhShah2011]

quality of the crowd \( \sigma^2 \triangleq \frac{1}{n} \sum_{j=1}^{n} (2p_j - 1)^2 \)

**Achievability:** random graph \( E \) and **Iterative Algorithm** achieves

\[
P_{\text{err}} \leq e^{-c\sigma^2 \ell}
\]

**Fundamental limit:** the best task-assignment and inference is limited by

\[
\min_{E, \hat{t}} \max_{t \in \{\pm 1\}^m, p \in \mathcal{F}_{\sigma^2}} P_{\text{err}} \geq e^{-c'\sigma^2 \ell}
\]
What is known for DS model [KargerOhShah2011]

quality of the crowd \( \sigma^2 \triangleq \frac{1}{n} \sum_{j=1}^{n} (2p_j - 1)^2 \)

Achievability: random graph \( E \) and Iterative Algorithm achieves

\[
P_{\text{err}} \leq e^{-c \sigma^2 \ell}
\]

Fundamental limit: the best task-assignment and inference is limited by

\[
\min_{E, \hat{t}} \max_{t \in \{\pm 1\}^m, p \in \mathcal{F}_{\sigma^2}} P_{\text{err}} \geq e^{-c' \sigma^2 \ell}
\]

Disturbing fact: the best adaptive scheme is limited by

\[
\min_{\text{all adaptive schemes}, \hat{t}} \max_{t \in \{\pm 1\}^m, p \in \mathcal{F}_{\sigma^2}} P_{\text{err}} \geq e^{-c'' \sigma^2 \ell}
\]
Experiments using Amazon Mechanical Turk

- Learning similarities using Crowdsourcing
  - recommendations
  - searching

You wanted to get the tie on the top, but it was not available. Which one would be a better substitute?

- the tie the left
- the tie on the right
Experiments using Amazon Mechanical Turk

- Learning similarities using Crowdsourcing
  - recommendations
  - searching

Which tie matches my shirt better?

- [ ] the tie on the left
- [x] the tie on the right
Experiments using Amazon Mechanical Turk

- Number of responses per task
- Probability of error

- Which colors...
  - left
  - right

- 0.14 EM Algorithm
- 0.12 Majority voting
- 0.06 Iterative Algo.
Experiments using Amazon Mechanical Turk

Which colors...

- left
- right

Number of responses per task

Average probability of error

- EM Algorithm
- Majority voting
- ITERATIVE ALGO.
Adaptive schemes are common in practice

- in practice, adaptive schemes improve significantly
- in theory, the gain is minimal
Adaptive schemes are common in practice

- in practice, adaptive schemes improve significantly
- in theory, the gain is minimal
Adaptive schemes are common in practice

- in practice, adaptive schemes improve significantly
- in theory, the gain is minimal
Adaptive schemes are common in practice

- in practice, adaptive schemes improve significantly
- in theory, the gain is minimal
Adaptive schemes are common in practice

- in practice, adaptive schemes improve significantly
- in theory, the gain is minimal
Adaptive schemes are common in practice

- in practice, adaptive schemes improve significantly
- in theory, the gain is minimal
Adaptive schemes are common in practice

- in practice, adaptive schemes improve significantly
- in theory, the gain is minimal
Adaptive schemes are common in practice

- in practice, adaptive schemes improve significantly
- in theory, the gain is minimal
When does adaptivity help?
Generalized DS model [Zhou et al. ’15]

- **Definition**
  - Each task is parametrized by a scalar value $q_i \in [0, 1]$
  - When a task is presented to a worker, it is perceived as a positive task with probability $q_i$

$$A_{ij} = \begin{cases} 
+1 & \text{with probability } q_ip_j + (1 - q_i)(1 - p_j) \\
-1 & \text{with probability } q_i(1 - p_j) + (1 - q_i)p_j
\end{cases}$$

- ground truth is $t_i = \mathbb{I}(q_i > 0.5) - \mathbb{I}(q_i < 0.5)$
- difficulty level of a task $i$ measured by $(2q_i - 1)^2$
- recovers the DS model when all $q_i \in \{0, 1\}$
Generalized DS model [Zhou et al. ’15]

Definition

- Each task is parametrized by a scalar value $q_i \in [0, 1]$
- When a task is presented to a worker, it is perceived as a positive task with probability $q_i$

$$A_{ij} = \begin{cases} +1 & \text{with probability } q_ip_j + (1 - q_i)(1 - p_j) \\ -1 & \text{with probability } q_i(1 - p_j) + (1 - q_i)p_j \end{cases}$$

- ground truth is $t_i = \mathbb{I}(q_i > 0.5) - \mathbb{I}(q_i < 0.5)$
- difficulty level of a task $i$ measured by $(2q_i - 1)^2$
- recovers the DS model when all $q_i \in \{0, 1\}$
Generalized DS model with adaptive scheme

Collective task difficulty:

$$\rho^2 = \frac{1}{m \sum_{i=1}^{m} \frac{1}{(2q_i - 1)^2}}$$

$$\approx e^{-c(2q_{\min} - 1)^2 \mu^2 \ell}$$
Non-adaptive $E$
Majority voting

$$\approx e^{-c(2q_{\min} - 1)^2 \sigma^2 \ell}$$
Non-adaptive $E$
Iterative Algorithm

$$\approx e^{-c' \rho^2 \sigma^2 \ell}$$
Adaptive $E$
Iterative Algorithm
Minimax rate for adaptive scenario [KhetanOh16]

- task difficulty $\rho^2 = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{(2q_i-1)^2}$
Minimax rate for adaptive scenario [KhetanOh16]

- task difficulty \( \rho^2 = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{(2q_i - 1)^2} \)

**Fundamental limit:** the best adaptive scheme and inference is limited by

\[
\min_{\text{adaptive}} \mathbb{E}, \hat{t} \quad \max_{q \in \mathcal{G}_{\rho^2}, p \in \mathcal{F}_{\sigma^2}} P_{\text{err}} \geq e^{-C' \rho^2 \sigma^2 \ell}
\]

**Achievability:** efficient adaptive scheme and inference achieves

\[
P_{\text{err}} \leq e^{-C \rho^2 \sigma^2 \ell}
\]

**Fundamental limit:** the best non-adaptive scheme is limited by

\[
\min_{\mathbb{E}, \hat{t}} \max_{q \in \mathcal{G}_{\rho^2}, p \in \mathcal{F}_{\sigma^2}} P_{\text{err}} \geq e^{-C'' (2q_{\text{min}} - 1)^2 \sigma^2 \ell}
\]
**Iterative Algorithm**

- **Message-passing algorithm**
  - Two sets of messages:
    - Task messages \( \{ T_{i \rightarrow j} \} \), and worker messages \( \{ W_{j \rightarrow i} \} \)
    - Initialize worker messages as random Gaussian: \( W_{j \rightarrow i} \sim \mathcal{N}(1, 1) \)
    - Iteratively update messages

  Task-likelihood update
  
  \[
  T_{i \rightarrow j} = \sum_{j' \neq j} W_{j' \rightarrow i} A_{ij'}
  \]

  Worker-reliability update
  
  \[
  W_{j \rightarrow i} = \sum_{i' \neq i} T_{i' \rightarrow j} A_{i'j}
  \]

- A task is likely to be ‘+’ if reliable workers agree that it is ‘+’
- A worker is reliable if the worker agreed with our belief on other tasks
1. DS model + non-adaptive scheme
1. DS model + non-adaptive scheme

\[ P(T_{i \rightarrow j} | q_i = 0) \]

\[ P(T_{i \rightarrow j} | q_i = 1) \]

\[ T_{i \rightarrow j} = \sum A_{ik} W_{k \rightarrow i} \]
1. DS model + non-adaptive scheme

\[ \mathbb{P}(T_{i \rightarrow j} | q_i = 0) \]

\[ \mathbb{P}(T_{i \rightarrow j} | q_i = 1) \]
1. DS model + non-adaptive scheme

\[ \mathbb{P}(T_{i \rightarrow j} | q_i = 0) \]

\[ \mathbb{P}(T_{i \rightarrow j} | q_i = 1) \]

\[ 2q_i - 1 \]
1. DS model + non-adaptive scheme
1. DS model + non-adaptive scheme

\[ P(T_{i \rightarrow j} | q_i = 0) \]

\[ \hat{t}_i = -1 \quad \hat{t}_i = +1 \]

\[ P(T_{i \rightarrow j} | q_i = 1) \]

\[ P_{\text{err}} \leq e^{-C \sigma^2 \ell} \]
2. Generalized DS model + non-adaptive scheme
2. Generalized DS model + non-adaptive scheme
2. Generalized DS model + non-adaptive scheme

\[
\hat{t}_i = -1 \quad \hat{t}_i = +1
\]

\[
P(T_{i \rightarrow j} | q_i)
\]
2. Generalized DS model + non-adaptive scheme

\[ \hat{t}_i = -1 \quad \hat{t}_i = +1 \]

\[ \mathbb{P}(T_{i \rightarrow j} | q_i) \]

\[ P_{err} \leq e^{-C(2q_{\min} - 1)^2 \sigma^2 \ell} \]
3. Generalized DS model + adaptive scheme

- Repeat rounds \( t \in \{1, 2, \ldots\} \)
  - Assign tasks with random \((\ell_t, r)\)-regular random graph
  - Run Iterative Algorithm
  - Classify high-confidence tasks with threshold \( \mathcal{X}_t \)
3. Generalized DS model + adaptive scheme

- Repeat rounds \( t \in \{1, 2, \ldots \} \)
  - Assign tasks with random \((\ell_t, r)\)-regular random graph
  - Run Iterative Algorithm
  - Classify high-confidence tasks with threshold \( X_t \)
3. Generalized DS model + adaptive scheme

- Repeat rounds $t \in \{1, 2, \ldots\}$
  - Assign tasks with random $(\ell_t, r)$-regular random graph
  - Run **Iterative Algorithm**
  - Classify high-confidence tasks with threshold $\mathcal{X}_t$
3. Generalized DS model + adaptive scheme

Repeat rounds $t \in \{1, 2, \ldots\}$

- Assign tasks with random $(\ell_t, r)$-regular random graph
- Run Iterative Algorithm
- Classify high-confidence tasks with threshold $\mathcal{X}_t$
3. Generalized DS model + adaptive scheme

- Repeat rounds $t \in \{1, 2, \ldots\}$
  - Assign tasks with random $(\ell_t, r)$-regular random graph
  - Run Iterative Algorithm
  - Classify high-confidence tasks with threshold $\mathcal{X}_t$
3. Generalized DS model + adaptive scheme

Repeat rounds $t \in \{1, 2, \ldots\}$
- Assign tasks with random $(\ell_t, r)$-regular random graph
- Run Iterative Algorithm
- Classify high-confidence tasks with threshold $\mathcal{X}_t$
3. Generalized DS model + adaptive scheme

- Repeat rounds $t \in \{1, 2, \ldots\}$
  - Assign tasks with random $(\ell_t, r)$-regular random graph
  - Run Iterative Algorithm
  - classify high-confidence tasks with threshold $\mathcal{X}_t$
Where did this algorithm come from?

• Spectral method

\[
A = \begin{bmatrix}
+ & - & + & - \\
- & + & - & - \\
+ & - & + & - \\
- & - & - & - \\
\end{bmatrix}
\]

\[
\mathbb{E}[A|t, p] = \frac{\ell}{n} \cdot t \cdot (2p - 1)
\]

\[
\mathbb{E}[A|t] = \frac{\ell}{n} \cdot t \cdot (2p - 1)^T
\]

▶ Singular vector of a non-backtracking matrix (widely used in community detection [Mossel et al. ’13, Krzkala et al. ’13, Bordenave et al. ’14, Saade et al. ’15, etc.])

• Belief Propagation for approximating [Peng et al. 12]

\[
P(t_i|A)
\]
Recap

- Each worker with quality $p_j$

\[ \sigma^2 = \frac{1}{n} \sum_{j=1}^{n} (2p_j - 1)^2 \]

\[ P_{\text{err}} \propto e^{-C\sigma^2 \ell} \]

- Criticism: homogeneous tasks
- No gain in adaptivity
Recap

**Dawid-Skene model**

- Each worker with quality $p_j$
  \[
  \sigma^2 = \frac{1}{n} \sum_{j=1}^{n} (2p_j - 1)^2
  \]
  \[
  P_{\text{err}} \sim e^{-C\sigma^2\ell}
  \]

**generalized Dawid-Skene**

- Each task with difficulty $q_i$
  \[
  \rho^2 = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{(2q_i - 1)^2}
  \]
  \[
  P_{\text{err}} \sim e^{-C\rho^2\sigma^2\ell}
  \]

- Criticism: homogeneous tasks
- No gain in adaptivity

**non-adaptive limit:**

\[
P_{\text{err}} \sim e^{-C(2q_{\text{min}} - 1)^2\sigma^2\ell}
\]
What is the gain of knowing prior on $p_j$'s: analysis of Belief Propagation.

![Graph showing the comparison between Majority Voting, Iterative Algo., and Lower Bound on probability of error as a function of the number of workers per task.](image)
Ongoing/future work

What is the gain of knowing prior on $p_j$'s: analysis of Belief Propagation.

\[ P_{err}^{BP} - P_{err}^{LB} \leq m^{-\gamma} \quad \text{for some } \gamma > 0 \]

[OkOhShinYi16]
Related work

- Crowdsourcing in machine learning
  - [ShengProvostIpeirotis2008] - first modern application of DS model
  - [WhitehillWuBergsmaMovellanRuvolo2009] - practical impact
  - [WelinderBransonPeronaBelongie2010] - practical impact
  - [GhoshKaleMcAfee2011] - weakness of analysis: $P_{\text{err}} \leq \frac{C}{\sigma^2 \ell}$

- Lots of recent advances following our initial work...
  - Spectral meets EM [ZhangChenZhouJordan2016]
    + generalized to $k$ classes
    - assumes dense graph
    - open question: how to deal with sampling bias?
  - Permutation based model [ShahBalakrishnanWainwright2016]
    + introduce general model based on permutation
    - permutation model is only defined for binary classification
    - can only show a loose bound of $P_{\text{err}} \leq \frac{C}{\sigma^2 \ell}$
    - open question: what is the right model of human computation?
  - open question: how to model complex crowdsourcing tasks?
In this talk...
