Performance as a Constraint: An Improved Wisdom of Crowds Using Performance Regularization

Jiyi Li[1*], **Yasushi Kawase**[2], **Yukino Baba**[3] and **Hisashi Kashima**[4]
[1]University of Yamanashi
[2]Tokyo Institute of Technology
[3]University of Tsukuba
[4]Kyoto University
jyli@yamanashi.ac.jp, kawase.y.ab@m.titech.ac.jp, baba@cs.tsukuba.ac.jp, kashima@i.kyoto-u.ac.jp

Abstract

Quality assurance is one of the most important problems in crowdsourcing and human computation, and it has been extensively studied from various aspects. Typical approaches for quality assurance include unsupervised approaches such as introducing task redundancy (i.e., asking the same question to multiple workers and aggregating their answers) and supervised approaches such as using worker performance on past tasks or injecting qualification questions into tasks in order to estimate the worker performance. In this paper, we propose to utilize the worker performance as a global constraint for inferring the true answers. The existing semi-supervised approaches do not consider such use of qualification questions. We also propose to utilize the constraint as a regularizer combined with existing statistical aggregation methods. The experiments using heterogeneous multiple-choice questions demonstrate that the performance constraint not only has the power to estimate the ground truths when used by itself, but also boosts the existing aggregation methods when used as a regularizer.

1 Introduction

In spite of the recent significant advances in artificial intelligence technologies, it is still difficult for machines to solve open-world and knowledge-intensive problems. Human computation [Law and Ahn, 2011] is a promising idea of combining human intelligence and machine intelligence to solve such "AI-hard" problems. On the other hand, the recent rise of crowdsourcing allows one to recruit human labor that can be scaled in an on-demand manner. This not only has such a quantitative advantage but also has a qualitative advantage, that is, the crowds often include experts who know correct solutions to the problems at hand. Consequently, crowdsourcing is often regarded as a favorable execution platform for human computation. However, despite such huge merits, there are still several challenges in the effective use of crowdsourcing. One of the typical problems is the quality control of crowd work. Because there is huge variability in the diligence and ability of crowd workers, the quality of crowdsourcing results also has a huge variance. There are at least two typical approaches to the quality control problem: one is the unsupervised approach and the other is the supervised approach. The unsupervised approach is based on the statistical inference of latent true answers. One of the typical methods is majority voting. More sophisticated statistical models that consider worker ability and task difficulty have also been proposed (e.g., Dawid and Skene (1979), Karger et al. (2011), Zhou et al. (2012), Venanzi et al. (2014), Whitehill et al. (2009), Welinder et al. (2010), Wauthier and Jordan (2011), Bachrach et al. (2012), Ma et al. (2015), Zhou and He (2016), and Yin et al. (2017)).

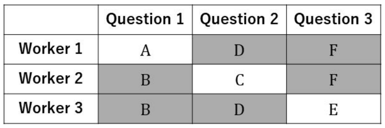
> Q: Which of the following drugs is most likely to cause
> Cushing's syndrome with long-term use?
> (a) Heparin, (b) Insulin, (c) Theophylline, (d) Prednisolone

Figure 1: An example of *heterogeneous multiple-choice questions*.

However, because most of the unsupervised aggregation methods basically enhance the majority opinions (and the workers who provided them), their performance is limited in difficult cases where majority voting fails [Li *et al.*, 2017]. Figure 1 shows a difficult multiple-choice question requiring professional medical knowledge. Among the four candidate answers, the correct one is '(d) Prednisolone'. However, '(b) insulin' is probably the most familiar medical term for non-experts, and therefore, incapable workers are likely to choose it, which sometimes results in the wrong answer winning in the majority voting. This is the so-called availability heuristic by which people tend to choose answers that are easier to recall [Baker *et al.*, 2004]. A simple remedy for the difficult aggregation problem is the supervised approach that estimates worker performance more directly. It typically uses qualification questions whose true answers are known to the requester, but unknown to the workers. The estimated performance allows us to give large importance weights to the expert workers or screen out incapable workers. Instead of qualification questions, the performance on past tasks the workers engaged in is sometimes available.

In this paper, we propose a hybrid approach that we call *performance constraint* that combines the supervised and un-

[*]Contact Author



	Question 1	Question 2	Question 3
Worker 1	A	D	F
Worker 2	B	C	F
Worker 3	B	D	E

Figure 2: An example of three heterogeneous multiple-choice questions answered by three equal-performance workers. A, C, and E are the correct answers of the three questions, respectively, and the others are wrong answers. The simple majority voting and performance-weighted voting fail to find the correct answers, while the proposed method successfully finds the solutions by using the worker performance information (i.e., $1/3$ of the answers are correct) as a constraint of inference.

supervised approaches in a novel manner. Our approach does not involve the typical usages of worker performance such as the use as the seeds or local constraints for the answers in semi-supervised aggregation methods. Instead, we solve a constraint satisfaction problem to find answers that are consistent with given worker performance. In other words, we use worker performance as a *global constraint* for inferring the true answer. Our theoretical analysis guarantees that we can find true answers with high probability if we have a sufficient number of workers even when used by itself.

We demonstrate the idea of our approach using an example. Assume that we have three heterogeneous multiple-choice questions and obtain the answers shown in Figure 2. The ground truth answers to the three questions are A, C, and E, respectively. Like the previous example (Figure 1), the majority voting fails and results in wrong answers B, D, and F that are more popular than the correct answers. What if we know the exact performance of each worker? Even in this case, since the performance of all the workers are equally $1/3$, the answers of all workers should be considered equally important, and therefore the performance-weighted majority voting still results in the wrong answers. However, if we change the way in which we exploit the worker performance, it is possible to estimate the correct answers. Instead of using the performance information as the reliability of the workers, we use it as a constraint on the ratio of correct answers of each worker. In the present example, assuming that the correct answer to question 1 is B, workers 2 and 3 are correct on this question, and worker 1 is not. Considering the constraint that the performance of all workers is $1/3$, workers 2 and 3 cannot make any more correct answers, and therefore none of C, D, E, F can be correct answers for the other questions; this is a contradiction for worker 1. Similarly, we can exclude D and F from the answers of questions 2 and 3, respectively, and obtain the correct answers, A, C, and E. Note that if we set the ground truth answers as B, D, and F, although majority voting succeeds; our method can also give the correct answers; our method has higher capacity than majority voting.

In practice, the constraints cannot always uniquely identify the correct answers, or the worker performance is known only approximately. We relax the above constraint satisfaction problem as an optimization problem to minimize the discrepancy between the performance on the estimated answers and the performance constraint. As well as standard divergence measures such as the Euclidean distance and KL divergence, we also propose a ranking-based relaxation.

In addition to using the performance information as a constraint for aggregation, we also propose to use it as a regularizer for existing the statistical aggregation methods, which we call *performance regularization*. Because the performance constraint itself is independent of the underlying probabilistic generative models of worker answers, we use it in combination with such models to boost their performance.

The experimental results show the proposed method is quite powerful when we know the exact worker performance. The aggregation accuracy degrades when the performance estimates are not accurate, but the ranking-based constraint relaxation alleviates the weakness. In addition, when combined with existing statistical aggregation methods, the performance regularizer consistently boosts the performance.

2 Problem Setting

We address a problem of aggregating answers provided by crowd workers for rather difficult questions. In contrast to the standard unsupervised aggregation setting, we additionally suppose that we know the performance of each worker.

Assume that we have $\mathcal{M}$ questions, each of which is a heterogeneous $\mathcal{K}$-choice question that has its own $\mathcal{K}$ potential answers (e.g., the medical knowledge question in Figure 2). For simplicity, we assume that $\mathcal{K}$ is the same for all questions, but this is not a requirement. We ask the questions to $\mathcal{N}$ workers, and obtain a set of responses $\{r_{ij}\}_{i,j}$, where r_{ij} is the response of worker i to question j. For notational simplicity, we assume that all workers answer all questions, but this assumption is not necessarily required. Here, r_{ij} is represented as a $\mathcal{K}$-dimensional one-hot vector only one of whose elements is 1, that is, $r_{ij} = (0, \dots, 0, 1, 0, \dots, 0)^\top \in \{0, 1\}^{\mathcal{K}}$. The goal is to estimate the ground truths $\{g_j\}_j$, where g_j is the estimated ground truth for question j and denoted by a $\mathcal{K}$-dimensional one-hot encoded binary vector, that is, $g_j = (0, \dots, 0, 1, 0, \dots, 0)^\top \in \{0, 1\}^{\mathcal{K}}$. An important assumption we make in this paper is that we somehow know the performance p_i of worker i; in other words, p_i is the ratio of correct answers by worker i.

In summary, the aggregation problem with worker performance takes the input as $\{r_{ij}\}_{i,j}$, the set of worker responses, and $\{p_i\}_i$, the performance levels of the workers, and the outputs are the estimated ground truths $\{g_j\}_j$.

One might be afraid that it is a too strong assumption that we know the worker performance; however, it is often the case that we can know (or at least guess) the worker performance based on the past tasks the worker has accomplished or the answers to qualification questions.

3 Performance Constraint

3.1 Worker Performance as a Constraint

In crowdsourcing, we often use qualification questions to test if the workers have sufficient capacity to execute the given tasks. Unlike that in the case of the main tasks, the requester knows the true answers for the qualification questions. They are given to workers before they start the main tasks, or are just mixed in with the main tasks in an indistinguishable way.

The requester can evaluate the worker ability using the qualification questions. The qualification questions can also be used in statistical aggregation methods in a semi-supervised manner. Most of the existing aggregation models are unsupervised and usually estimate the true answers as latent variables; however, they can easily incorporate the qualification questions by fixing their latent answers to their true answers.

In this paper, we consider a totally different way to exploit the results of the qualifications questions, that is, to use the performance on the qualification questions as a global constraint for aggregation. More precisely, suppose that the performance (i.e., accuracy) of worker i on the qualification questions is p_i, we force the performance (i.e., the ratio of correct answers) on the other (main) questions also to be p_i. Our idea is represented as a constraint satisfaction problem to find $\{g_j\}_j$ that satisfies the constraint

$$\frac{1}{\mathcal{M}} \sum_j I(r_{ij} = g_j) = p_i, \tag{1}$$

where I is the indicator function. Note that $I(r_{ij} = g_j)$ can also be written as $I(r_{ij} = g_j) = r_{ij}^\top g_j$.

3.2 Theoretical Justification

We gave an intuitive explanation of how the performance constraint helps aggregation, but it is still not theoretically clear if the performance constraint is generally capable of finding the ground truth answers. The following theorem guarantees that we can obtain the ground truth answers with high probability if we have a sufficient number of workers. For simplicity, we consider a simple case where $\mathcal{K} = 2$ and each worker i gives the ground truth answers for $p_i \mathcal{M}$ questions independently and uniformly at random (i.e., $\binom{\mathcal{M}}{p_i \mathcal{M}}$ possibilities occur with the same probability). We assume $p_i \neq 1/2$ for some i, because otherwise the all-wrong answers $(g_j^\dagger)_j$ (i.e., $g_j^\dagger \neq g_j^*$ for all question j) always satisfy the constraint (1).

Theorem 1. *If $p_1 \neq 1/2$ and $\mathcal{N} > \mathcal{M} \ln \mathcal{M}$, then the ground truth answers are uniquely determined by the performance constraint* (1) *with high probability.*

Proof. Without loss of generality, we assume that $p_i \notin \{0, 1\}$ for all i, because if $p_i \in \{0, 1\}$, the ground truth answers are determined by the answers of worker i.

For $s = 1, \ldots, \mathcal{N}$, consider an integer linear system:

$$\text{IP}(s)\text{:} \begin{cases} \sum_j a_{ij} x = (2p_i - 1)\mathcal{M} & (\forall i \in \{1, \ldots, s\}), \\ x_j \in \{-1, 1\} & (\forall j \in \{1, \ldots, \mathcal{M}\}), \end{cases}$$

$$\text{where } a_{ij} = \begin{cases} +1 & (g_j^* = r_{ij}), \\ -1 & (g_j^* \neq r_{ij}). \end{cases}$$

Then, (1) has a unique solution if and only if IP($\mathcal{N}$) has a unique solution. Indeed, if $(\hat{g}_j)_j$ is a solution of (1), then

$$\hat{x}_i = \begin{cases} +1 & (\hat{g}_j = g_j^*), \\ -1 & (\hat{g}_j \neq g_j^*) \end{cases}$$

is a solution of IP($\mathcal{N}$), and vice versa. In particular, the all-one vector of length $\mathcal{M}$ (which corresponds to the ground truth answers) is always a solution of IP(s) for any s.

If $\operatorname{rank}(a_1^\top, \ldots, a_s^\top) = \mathcal{M}$ (i.e., full rank), then the solution of IP(s) is uniquely determined. Suppose that the solution of IP(s) is not uniquely determined and let $\rho = \operatorname{rank}(a_1^\top, \ldots, a_s^\top)$ ($< \mathcal{M}$). Then, for any $s' > s$, we have $a_{s'}^\top \notin \operatorname{span}\{a_1^\top, \ldots, a_s^\top\}$ with probability at least (i) $(\mathcal{M} - \rho - 1)/\mathcal{M}$ if $\rho < \mathcal{M} - 1$, and (ii) $1/\mathcal{M}$ if $\rho = \mathcal{M} - 1$ (see supplementary material for proof). Note that $a_{s'}^\top \notin \operatorname{span}\{a_1^\top, \ldots, a_s^\top\}$ implies $\operatorname{rank}(a_1^\top, \ldots, a_{s'}^\top) > \rho$. Hence, by the standard analysis of the *coupon collector's problem*, IP($\mathcal{N}$) has a uniquely solution with high probability (see, e.g., Motwani and Raghavan [1995]). □

A detailed proof can be found in the appendix.

4 Relaxation and Performance Regularization

4.1 Distance-based Constraint Relaxation

In practice, there might be neither an exact solution nor a unique solution satisfying Eq. (1), so we relax the ground truth vector; let $g_j \in [0, 1]^{\mathcal{K}}$ be a $\mathcal{K}$-dimensional vector in a simplex such that $g_j \geq 0$ and $|g_j|_1 = 1$. We obtain a relaxed optimization problem as

$$\{g_j^*\}_j = \operatorname*{arg\,min}_{\{g_j\}_j} \sum_i d\left(\frac{1}{\mathcal{M}} \sum_j r_{ij}^\top g_j,\, p_i\right) \tag{2}$$

subject to $g_j \geq 0$ and $|g_j|_1 = 1$, where d is a distance function between two probability values. The choice of d is arbitrary; for example, we can choose the Euclidean distance

$$d\left(\frac{1}{\mathcal{M}} \sum_j r_{ij}^\top g_j,\, p_i\right) = \left(\frac{1}{\mathcal{M}} \sum_j r_{ij}^\top g_j - p_i\right)^2,$$

or the KL-divergence, which also seems a reasonable choice:

$$d\left(\frac{1}{\mathcal{M}} \sum_j r_{ij}^\top g_j,\, p_i\right) = -\, p_i \log\left(\frac{1}{\mathcal{M}} \sum_j r_{ij}^\top g_j\right) - (1 - p_i) \log\left(1 - \frac{1}{\mathcal{M}} \sum_j r_{ij}^\top g_j\right) + \text{const.}$$

4.2 Ranking-based Constraint Relaxation

Obviously, the worker accuracy is not exactly the same between the qualification questions and the main questions, and therefore, there is a risk that the performance constraint becomes harmful when the performance estimate is wrong. Our solution is that, instead of using the constraint based on the distance, we relax the constraint by the relative order of worker accuracies. More precisely, if a worker i_1 has higher accuracy than that of another worker i_2 on the qualification questions, we impose the constraint that i_1 has higher accuracy than i_2 on the main questions. We implement this idea as another relaxed objective function instead of Eq. (2) as

$$\{g_j^*\}_j = \operatorname*{arg\,min}_{\{g_j\}_j} \sum_{i_1, i_2} I(i_1 > i_2) \log\left(1 + e^{-\mathcal{T} \cdot J(i_1, i_2)}\right),$$

$$J(i_1, i_2) = \frac{1}{\mathcal{M}} \sum_j r_{i_1 j}^\top g_j - \frac{1}{\mathcal{M}} \sum_j r_{i_2 j}^\top g_j. \tag{3}$$

The constant $\mathcal{T} > 0$ controls how strict the constraint is.

All of the above choices are convex functions; hence, the optimal solutions are found by the simple gradient-based update, followed by the projection to the constraints $g_j \geq 0$ and $|g_j|_1 = 1$ (normalizing g_j so as to be a probability vector). The computation cost of the rank-based constraint is higher than the distance-based constraints, but it does not give a significant impact on the scalability of the entire optimization.

4.3 Performance Regularization

The performance constraint itself is model agnostic; that is, it works independently of the underlying probabilistic generative models of worker answers. Now, we combine the constraint with the existing statistical methods; namely, we include the performance constraint as a regularization term.

For example, let us consider a simple generative model. Assume that the probability of worker i giving a correct answer is a_i and that when a worker fails to give a correct answer, the worker chooses a wrong answer uniformly at random, namely,

$$P(\{r_{ij}\}_{i,j} \mid \{g_j\}_j) = \prod_{i,j} a_i^{r_{ij}^\top g_j} \left(\frac{1-a_i}{\mathcal{K}-1}\right)^{1-r_{ij}^\top g_j}.$$

The log-likelihood of the worker answers is given as

$$\begin{aligned} &L(\{r_{ij}\}_{i,j} \mid \{g_j\}_j) \\ &= \sum_{i,j} \left(r_{ij}^\top g_j \log a_i + (1 - r_{ij}^\top g_j) \log \left(\frac{1-a_i}{\mathcal{K}-1} \right) \right). \end{aligned} \tag{4}$$

In practice, we apply the following re-parameterization to ensure the positiveness of a_i:

$$a_i = \frac{1}{1+e^{-\alpha_i}}. \tag{5}$$

By including the distance d in (2) as a regularization term, the overall objective function is given as

$$\begin{aligned} &F(\{a_i\}_i, \{g_j\}_j) \\ &= -\sum_{i,j} \left(r_{ij}^\top g_j \log a_i + (1 - r_{ij}^\top g_j) \log \left(\frac{1-a_i}{\mathcal{K}-1} \right) \right) \\ &\quad + \lambda \sum_i d\left(\frac{1}{\mathcal{M}} \sum_j r_{ij}^\top g_j,\, p_i \right), \end{aligned} \tag{6}$$

where $\lambda > 0$ is the regularization constant. Instead, we can use the ranking constraint term in Eq. (3) as the regularization term. We call this *performance regularization*.

Performance regularization is also applicable to more sophisticated generative models. For example, if we use GLAD [Whitehill *et al.*, 2009] as the underlying model, we use the following re-parameterization instead of (5): $a_i = 1/\left(1+e^{-\alpha_i d_j}\right)$, where d_j denotes the easiness of task j. With the help of proper underlying models, the performance regularization method can incorporate both worker performance constraint and other types of useful factors for truth estimation such as task difficulty. The latent variables in the performance regularization term are only the estimated truth g_j. Therefore, it is easy to include the performance regularization term in the objective function of an underlying model to iteratively estimate the truths and other potential factors such as worker expertise on each question. One example is to use the MinMax [Zhou *et al.*, 2012] approach as the underlying model. Note that, if the performance of only a subset of workers is known, we can use the performance constraint or regularization only for such workers.

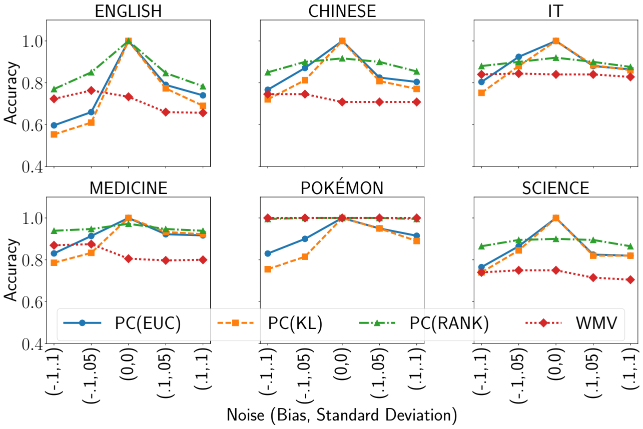


Figure 3: How robust is the performance constraint when we do not know the true accuracy? The distance-based performance constraints (PC(EUC) and PC(KL)) are moderately tolerant of a small amount of noise; however, the performance degrades fast when the noise level increases, in contrast with the weighted majority voting (WMV). The ranking-based constraint (PC(RANK)) is relatively robust to such error.

5 Experiments

We investigate the following issues in the experiments: (1). How powerful is the performance constraint (PC) alone? (2). How robust is the performance constraint (PC) when we do not know the true worker performance? (3). Does the performance regularization (PR) improve aggregation results?

We use the six real datasets collected by Li et al. (2017), which consist of multiple-heterogeneous-answer questions on *Chinese language*, *English language*, *Information technology (IT)*, *Medicine*, *Pokémon*, and *Science*. They are relatively difficult questions in which the simple majority voting performs poorly (See the 'MV' column in Table 1.)

5.1 How Powerful is PC Alone?

The first question we investigate is whether the performance constraint is informative in guessing the ground truths. In this experiment, we assume that we somehow know the true worker accuracy. We test the performance constraint using two different divergence measures, i.e., the Euclid distance and the KL divergence, and the ranking-based constraint. We compare the proposed method with some baselines that are applicable to heterogeneous multiple-choice questions: (i) the simple majority voting (MV), (ii) the weighted majority voting (WMV) [Kuncheva and Rodríguez, 2014] which is the solution that maximizes Eq. (4) with respect to $\{g_j\}_j$ (with $a_i = p_i$), and (iii) several existing unsupervised aggregation

Datasets	PC(EUC)	PC(KL)	PC(RANK)	WMV	MV	D&S	GLAD	DARE	MinMax
ENGLISH	1.000	1.000	1.000	0.733	0.467	0.400	0.533	0.600	0.567
CHINESE	1.000	1.000	0.917	0.708	0.625	0.542	0.625	0.625	0.667
IT	1.000	1.000	0.920	0.840	0.760	0.680	0.8	0.800	0.840
MEDICINE	1.000	1.000	0.972	0.806	0.667	0.694	0.806	0.861	0.972
POKÉMON	1.000	1.000	1.000	1.000	0.650	0.600	1.000	1.000	0.950
SCIENCE	1.000	1.000	0.900	0.750	0.550	0.550	0.650	0.600	0.650

Table 1: How powerful is the performance constraint alone? Quite powerful. The proposed methods (PC(EUC) and PC(KL)) achieve perfect aggregations, while the ranking-based constraint (PC(RANK)) is suboptimal. They outperform another way of using the worker performance (WMV) as well as various standard unsupervised methods (MV, D&S, GLAD, DARE, MinMax).

Dataset	MV	PC (alone)			WMV				GLAD				MinMax			
						+PR				+PR				+PR		
		EUC	KL	RANK		EUC	KL	RANK		EUC	KL	RANK		EUC	KL	RANK
ENGLISH	0.461	0.483	0.439	0.535	0.500	0.600	0.617	0.591	0.509	0.578	0.570	0.587	0.504	**0.626**	0.570	0.596
CHINESE	0.622	0.628	0.594	0.728	0.633	0.661	0.683	0.644	0.644	0.667	0.650	0.711	0.633	0.700	0.661	**0.772**
IT	0.768	0.695	0.690	0.779	0.753	0.821	0.821	0.811	0.768	0.800	0.800	**0.832**	0.816	0.816	0.779	0.826
MEDICINE	0.667	0.782	0.782	0.874	0.770	0.796	0.796	0.852	0.811	0.830	0.852	**0.967**	0.915	0.952	0.922	0.956
POKÉMON	0.647	0.820	0.760	0.987	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.920	**1.000**	0.907	0.960	0.893	**1.000**
SCIENCE	0.547	0.540	0.527	0.627	0.567	0.640	0.653	0.620	0.640	0.647	0.553	**0.667**	0.633	0.600	0.533	0.653

Table 2: Does the performance regularization improve aggregation results? The performance regularization ('+PR') consistently boosts the accuracy of its unsupervised analogue, especially with the ranking constraint.

methods (D&S [Dawid and Skene, 1979], GLAD [Whitehill *et al.*, 2009], DARE [Bachrach *et al.*, 2012], and the min-max entropy method (MinMax) [Zhou *et al.*, 2012]). The weighted majority voting method corresponds to the optimal and standard approach when we know the worker performance. We use the entire datasets in the experiments.

Table 1 shows the accuracy achieved by the performance constraint with two different distances (PC(EUC) and PC(KL)), the one with the ranking-based constraint (PC(RANK)), as well as the WMV and other existing unsupervised aggregation methods. Surprisingly, the distance-based constraint achieves almost perfect aggregation, in contrast with the inferior performance achieved by WMV, which is the standard usage of the performance information. The ranking-based constraint shows sub-optimal performance since it cannot exploit the full performance information, but it still significantly outperforms WMV. This shows that our method exploits the performance information very differently from the standard approaches. It is noteworthy that the proposed method even outperforms DARE, one of the most powerful unsupervised Bayesian aggregation methods, and MinMax, a powerful minimax entropy method.

5.2 How Robust is PC When We Do Not Know True Worker Performance?

As discussed earlier, it is just an ideal case that we know the accurate worker performance; such assumption does not necessarily hold in practice, and hence all we can do is only to "guess" the worker performance from past work or qualification questions. In order to test the robustness to inaccurate estimates of worker performance, we add noise with the bias to the estimated worker accuracy parameters $\{p_i\}_i$ and test the robustness against the noise. We add a noise generated by a Gaussian distribution $\mathcal{N}(\eta, \sigma^2)$ with bias η of estimated worker performance to true worker performance as the mean and fit into $[0, 1]$, and we use it as the performance estimate.

Figure 3 shows the results when we change $\eta \in \{-0.1, 0.1\}$ and $\sigma \in \{0.05, 0.1\}$. We compare the performance constraint with the weighted majority voting (4). WMV is not quite sensitive to the inaccurate estimation of worker performance, but always not quite good. The performance of the proposed methods drops as the noise level increases. The general tendency of the performance degrade of each PC variation looks similar among different datasets. When the noise level is not very high ($\sigma = 0.05$), in most of the cases, the proposed methods (PC) keep their performance relatively high. Among the three variations, the ranking constraint (PC(RANK)) fights relatively better against the noise, while the other two (PC(EUC) and PC(KL)) suffer from the increasing noise. From these observations, the robustness to high-level noise in the performance estimate is one of the weaknesses of the proposed method; but the ranking-based constraint relaxation alleviates this problem for practical use.

5.3 Does PR Improve Aggregations?

Finally, we test our approach in a more realistic scenario, where we do not know the true worker performance and use several qualification questions to estimate it. Our primary interest here is to find out if the performance regularization improves the aggregation accuracy when combined with ordinary statistical aggregation methods. To test the hypothesis, we combine the proposed method ('PR') with the WMV, GLAD and MinMax. In this experiment, a subset (25%) of the questions in each dataset is considered as a set of qualification questions; we estimate the worker performance using their ground truths. The others are used as a test set. We iterate the split and test ten times, and obtain the average aggregation accuracy. For tuning the hyper-parameter λ in the performance regularization, we utilize the idea of the KL cost annealing method [Bowman *et al.*, 2016] to avoid running the optimization many times for the candidate hyper-parameters. We first gradually and linearly increase λ from zero during optimization until the perplexity on the test subset reaches zero. After that, we fix λ and optimize until convergence. The perplexity is defined as $P(\{g_j^k\}_{j,k}) = \sum_j \sum_k -g_j^k \log g_j^k$, where g_j^k is the estimated value of g for question j and its k-th candidate answer. It only needs to run the optimization once, which speeds up the hyper-parameter tuning considerably. Note that we do not use the ground truth labels at all.

Table 2 summarizes the results. In contrast with the previous where we knew the exact performance of the workers, the overall accuracy of the performance constraint ('PC(alone)') degrades as implied by Figure 3. However, it is still comparable with WMV, the direct competitor, as well as the other methods such as MV, GLAD and MinMax. Among the three choices of the performance constraint, the one with the ranking-based constraint performs the best, which is also consistent with the results of the previous experiment. When combined with the unsupervised aggregation methods, i.e., WMV, GLAD and MinMax, the performance regularization ('+PR') consistently boosts the accuracy of its unsupervised analogue, especially with the ranking constraint.

The performance regularization exploits the qualification questions in a very different way than the semi-supervised methods; it does not rely on detailed worker labels for the qualification questions, while most of the semi-supervised approaches assume they are available. In practice, it is sometimes difficult to obtain the detailed worker labels in past tasks, for example, due to privacy reasons, even in the cases where we know the overall past performance of a worker.

6 Related Work

Majority voting is probably the simplest answer aggregation method. Since it assigns equal weights to all workers, the aggregation performance is not stable due to the wide variety of crowd workers in their ability and diligence; therefore researchers have studied more sophisticated statistical aggregation methods that allow diverse worker abilities and other uncertainties. Some approaches jointly estimate worker ability and true answers using the expectation-maximization (EM) algorithm [Dawid and Skene, 1979], bipartite models [Karger *et al.*, 2011], the maximum entropy principle [Zhou *et al.*, 2012], and Bayesian inference [Liu *et al.*, 2012; Venanzi *et al.*, 2014]. More sophisticated models that incorporate task difficulty [Whitehill *et al.*, 2009], worker-task affinity [Welinder *et al.*, 2010], and their Bayesian treatments [Wauthier and Jordan, 2011; Bachrach *et al.*, 2012] have also been proposed. Unsupervised quality control is a quite active area, so we do not exhaustively enumerate all the work, but just to name a few, Zhou and He (2016), Yin et al. (2017), Kawase et al. (2019) are such examples. For truth discovery on the Web, Zheng et al. (2017) reviewed and compared various existing aggregation methods. Some of those methods are applicable to the heterogeneous multiple-choice questions.

Another practical and typical way of crowdsourcing quality control is to perform qualification tests [Downs *et al.*, 2010; Ipeirotis *et al.*, 2010; Mortensen *et al.*, 2013]. A qualification test is a set of questions whose ground truth answers are known only to the requester, and are performed before proceeding to the main tasks to appraise the workers or to filter out incapable workers. Instead of such pre-screening, qualification questions are also injected into the main tasks to perform post-screening. If the workers were engaged in previous tasks, their results and evaluations by past requesters are also considered as qualification results. Besides directly using the qualification results for pre- or post-worker screening, they are also used in semi-supervised aggregation as the seeds or local constraints for latent ground truths. Ipeirotis et al. (2010) implemented a variant of D&S [Dawid and Skene, 1979] by using injected qualification questions. Qualification tests are also used for targeting the expert workers [Downs *et al.*, 2010]; however, it is not always guaranteed that well-designed qualification tests are available for requesters. Mortensen et al. (2013) utilized crowdsourcing for constructing biomedical ontologies and selected knowledgeable workers using qualification tests. Unlike the previous usages of qualification questions, we use them as a global constraint or a regularizer to infer the ground truths. Recently, although in a different context and with a different motivation, Whitehill (2019) discussed the possibility of guessing the ground truth labels from AUC values of machine learning classifiers, whose settings are somewhat similar to ours.

Most semi/un-supervised approaches strengthen the opinions of the majority workers, and therefore do not perform well in cases where the number of capable workers (i.e., experts) is smaller than that of incapable workers. There are a few studies addressing such few-expert scenarios. Li et al. (2014) used worker profiles such as demographic information to predict their abilities. Ma et al. (2015) focused on task information such as the words appearing in task descriptions, and combined a topic model and a worker ability model to characterize worker expertise. Kazai et al. (2012) used demographics information and personality traits. Ipeirotis and Gabrilovich (2014) used sponsored search associated with medical terms to guide medical experts to medical question tasks. Prelec et al. (2017) used the response for an additional question asking the answer distribution of other workers. Different from those approaches exploiting auxiliary information, Li et al. (2017) proposed a method not relying on side information by bundling questions into "hyper questions".

7 Conclusion

We addressed the crowd aggregation problem when we know the worker performance. Unlike the previous semi-supervised approaches, we use worker performance as the constraint to infer ground truth labels. We also proposed performance regularization, which can be combined with ordinary aggregation methods. The experiments showed that the proposed method is quite powerful when we know the exact worker performance. The aggregation accuracy degrades when the performance estimates are not accurate, but the ranking-based constraint alleviates the weakness. When combined with statistical aggregation methods, the performance regularizer boosted the aggregation performance.

Dataset	MV	PC (alone)			WMV				GLAD				MinMax			
						+PR				+PR				+PR		
		EUC	KL	RANK		EUC	KL	RANK		EUC	KL	RANK		EUC	KL	RANK
ENGLISH	0.400	0.439	0.391	0.439	0.435	0.535	0.500	**0.539**	0.404	0.522	0.478	0.522	0.361	0.457	0.417	0.457
CHINESE	0.583	0.561	0.533	0.606	0.606	0.656	0.578	0.633	0.600	0.644	0.583	**0.672**	0.600	0.567	0.517	0.611
IT	0.753	0.642	0.642	0.637	0.768	0.774	0.779	0.763	0.768	0.763	0.705	**0.800**	0.721	0.721	0.700	0.674
MEDICINE	0.715	0.752	0.730	0.759	0.763	0.811	0.830	0.800	0.807	0.822	0.811	**0.859**	0.774	0.841	0.815	0.819
POKÉMON	0.533	0.720	0.673	0.847	0.753	0.927	0.740	0.840	0.867	0.933	0.713	**0.947**	0.653	0.767	0.707	0.867
SCIENCE	0.513	0.500	0.460	0.533	0.540	0.593	0.433	0.547	0.513	0.527	0.513	**0.613**	0.480	0.520	0.493	0.527

Table 3: The results of the missing value case for the experiments in Table 2. 50% of the answers to each test question are removed. The performance regularization still improves the performance.

Acknowledgments

This work was partially supported by JSPS KAKENHI Grant Number 19K20277 and 15H01704.

A Proofs

Here, we provide omitted proofs from the proof of Theorem 1.

Claim 1. *For any $s' > s$ and $\rho < \mathcal{M} - 1$, we have $a_{s'}^\top \notin \mathrm{span}\{a_1^\top, \dots, a_s^\top\}$ with probability at least $(\mathcal{M} - \rho - 1)/\mathcal{M}$.*

Proof. We regard $a_{s'}$ a vector of random variables but $a_1, \dots, a_s$ are fixed. We define $\mathcal{E}$ as the event that $a_{s'}^\top \notin \mathrm{span}\{a_1^\top, \dots, a_s^\top\}$.

Let $b_1^\top, \dots, b_\rho^\top$ be a basis of $\mathrm{span}\{a_1^\top, \dots, a_s^\top\}$, and let J be the index set of the rows that form a basis of $(b_1^\top, \dots, b_\rho^\top)$. Note that $|J| = \rho$. In addition, we write A_t for the set of vectors in $\{1, -1\}^J$ with exactly t ones (i.e., $A_t := \{\alpha \in \{1, -1\}^J \mid \sum_{j \in J} \alpha_j = 2t - \rho\}$) and T for the set of t such that $(a_{s'j})_{j \in J}$ contains exactly t ones and $(a_{s'j})_{j \notin J}$ contains both one and minus one for some realization of $a_{s'}$ (i.e., $T := \{t \mid \max\{0,\ \rho - \mathcal{M}(1 - p_{s'}) + 1\} \le t \le \min\{\rho,\ \mathcal{M}p_{s'} - 1\}\}$ since $a_{s'}$ contains exactly $\mathcal{M}p_{s'}$ ones). Then, for any $\alpha \in A_t$ with $t \in T$, we have

$$P(\mathcal{E} \mid (a_{s'j})_{j \in J} = \alpha) \ge \frac{\mathcal{M} - \rho - 1}{\mathcal{M} - \rho}$$

since $a \in \mathrm{span}\{a_1^\top, \dots, a_s^\top\}$ with $(a_j)_{j \in J} = \alpha$ is uniquely determined but $a_{s'}$ with $(a_{s'j})_{j \in J} = \alpha$ has at least $\mathcal{M} - \rho$ possibilities and they occur with the same probability. Moreover, we have

$$\sum_{\alpha \in \bigcup_{t \in T} A_t} P((a_{s'j})_{j \in J} = \alpha) \ge 1 - \frac{\binom{\mathcal{M}p_{s'}}{\mathcal{M}-\rho} + \binom{\mathcal{M}(1-p_{s'})}{\mathcal{M}-\rho}}{\binom{\mathcal{M}}{\mathcal{M}-\rho}}$$

$$\ge 1 - \frac{\binom{\mathcal{M}-1}{\mathcal{M}-\rho}}{\binom{\mathcal{M}}{\mathcal{M}-\rho}} = \frac{\mathcal{M} - \rho}{\mathcal{M}}$$

since $0 < \mathcal{M}p_{s'} < \mathcal{M}$. Hence, we have

$$P(\mathcal{E}) \ge \sum_{\alpha \in \bigcup_{t \in T} A_t} P(\mathcal{E} \mid (a_{s'j})_{j \in J} = \alpha) P((a_{s'j})_{j \in J} = \alpha)$$

$$\ge \frac{\mathcal{M} - \rho - 1}{\mathcal{M} - \rho} \cdot \frac{\mathcal{M} - \rho}{\mathcal{M}} = \frac{\mathcal{M} - \rho - 1}{\mathcal{M}}. \qquad \square$$

Claim 2. *For any $s' > s$ and $\rho = \mathcal{M} - 1$, we have $a_{s'}^\top \notin \mathrm{span}\{a_1^\top, \dots, a_s^\top\}$ with probability at least $1/\mathcal{M}$.*

Proof. We regard $a_{s'}$ a vector of random variables but $a_1, \dots, a_s$ are fixed. Let $\ell = \mathcal{M}p_{s'}$ and let $\tilde{x}$ be a solution of IP(s) that is not the all-one vector nor the all-minus-one vector. Such an $\tilde{x}$ must exist since IP(s) has multiple solutions and the all-minus-one vector is not feasible by $p_1 \neq 1/2$. It is sufficient to prove that $\sum_j a_{s'j}\tilde{x}_j = (2\ell - \mathcal{M})$ happens with probability at most $(\mathcal{M} - 1)/\mathcal{M}$. Let t be the number of minus ones in $\tilde{x}$. Since $\tilde{x}$ is not the all-one vector nor the all-minus-one-vector, we have $1 \le t \le \mathcal{M} - 1$.

The probability that $\sum_j a_{s'j}x_j = (2\ell - \mathcal{M})$ is $\frac{\binom{t}{t/2}\binom{\mathcal{M}-t}{\ell-t/2}}{\binom{\mathcal{M}}{\ell}}$ since it occurs when $(a_{s'j}, \tilde{x}_j) = (-1, -1)$ holds for exactly $t/2$ choices of j. This probability is equal to the probability that exactly $t/2$ balls are red out of ℓ balls drawn one after another without replacement from a bag that containing t red balls and $\mathcal{M} - t$ black balls. If $\ell > \mathcal{M} - t/2$, the probability is clearly zero. If $\ell = \mathcal{M} - t/2$, the probability is at most

$$\frac{\binom{t}{t/2}\binom{\mathcal{M}-t}{\ell-t/2}}{\binom{\mathcal{M}}{\ell}} = \frac{\binom{t}{t/2}\binom{\mathcal{M}-t}{\mathcal{M}-t}}{\binom{\mathcal{M}}{\mathcal{M}-t/2}} = \frac{\binom{t}{t/2}}{\binom{\mathcal{M}}{t/2}}$$

$$\le \frac{\binom{t}{t/2}}{\binom{t+1}{t/2}} = \frac{t/2 + 1}{t + 1} \le \frac{\mathcal{M} - 1}{\mathcal{M}}$$

since $t \le \mathcal{M} - 1$. For the case when $\ell < \mathcal{M} - t/2$, consider the situation that we have drawn $\ell - 1$ balls and about to draw the final one ball from the bag. We have two cases to draw $t/2$ red balls: (i) the number of drawn red balls is $t/2 - 1$ and the final ball is red or (ii) the number of drawn red balls is $t/2$ and the final ball is black. Thus the event occurs with probability at most

$$\max\left\{\frac{t/2 + 1}{\mathcal{M} - \ell + 1},\ 1 - \frac{t/2}{\mathcal{M} - \ell + 1}\right\} \le \frac{\mathcal{M} - 1}{\mathcal{M}}$$

since $\ell < \mathcal{M} - t/2$ and $t/2 \ge 1$. $\square$

B Additional Experiments

In practice, not all workers answer all questions. We randomly remove 50% of the answers of each test question for each trial in the experiment of Table 2. Table 3 shows that the performance regularization still contributes to the performance gain.

References

[Bachrach *et al.*, 2012] Yoram Bachrach, Tom Minka, John Guiver, and Thore Graepel. How to grade a test without knowing the answers: A bayesian graphical model for adaptive crowdsourcing and aptitude testing. In *ICML*, pages 819–826, 2012.

[Baker *et al.*, 2004] Kevin Baker, Anthony Esgate, David Groome, David Heathcote, Richard Kemp, Moira Maguire, and Corriene Reed. *An introduction to applied cognitive psychology*. Psychology Press, 2004.

[Bowman *et al.*, 2016] Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. In *CoNLL*, pages 10–21, 2016.

[Dawid and Skene, 1979] Alexander P. Dawid and Allan M. Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1), 1979.

[Downs *et al.*, 2010] Julie S. Downs, Mandy B. Holbrook, Steve Sheng, and Lorrie Faith Cranor. Are your participants gaming the system?: Screening mechanical turk workers. In *CHI*, pages 2399–2402, 2010.

[Ipeirotis and Gabrilovich, 2014] Panagiotis G. Ipeirotis and Evgeniy Gabrilovich. Quizz: Targeted crowdsourcing with a billion (potential) users. In *WWW*, pages 143–154, 2014.

[Ipeirotis *et al.*, 2010] Panagiotis G. Ipeirotis, Foster Provost, and Jing Wang. Quality management on amazon mechanical turk. In *HCOMP*, pages 64–67, 2010.

[Karger *et al.*, 2011] David R. Karger, Sewoong Oh, and Devavrat Shah. Iterative learning for reliable crowdsourcing systems. In *NIPS*, pages 1953–1961, 2011.

[Kawase *et al.*, 2019] Yasushi Kawase, Yuko Kuroki, and Atsushi Miyauchi. Graph mining meets crowdsourcing: Extracting experts for answer aggregation. In *IJCAI*, pages 1272–1279, 2019.

[Kazai *et al.*, 2012] Gabriella Kazai, Jaap Kamps, and Natasa Milic-Frayling. The face of quality in crowdsourcing relevance labels: Demographics, personality and labeling accuracy. In *CIKM*, pages 2583–2586, 2012.

[Kuncheva and Rodríguez, 2014] Ludmila I Kuncheva and Juan J Rodríguez. A weighted voting framework for classifiers ensembles. *Knowledge and Information Systems*, 38(2):259–275, 2014.

[Law and Ahn, 2011] Edith Law and Luis von Ahn. Human computation. *Synthesis lectures on artificial intelligence and machine learning*, 5(3):1–121, 2011.

[Li *et al.*, 2014] Hongwei Li, Bo Zhao, and Ariel Fuxman. The wisdom of minority: Discovering and targeting the right group of workers for crowdsourcing. In *WWW*, pages 165–176, 2014.

[Li *et al.*, 2017] Jiyi Li, Yukino Baba, and Hisashi Kashima. Hyper questions: Unsupervised targeting of a few experts in crowdsourcing. In *CIKM*, pages 1069–1078, 2017.

[Liu *et al.*, 2012] Qiang Liu, Jian Peng, and Alexander Ihler. Variational inference for crowdsourcing. In *NIPS*, pages 692–700, 2012.

[Ma *et al.*, 2015] Fenglong Ma, Yaliang Li, Qi Li, Minghui Qiu, Jing Gao, Shi Zhi, Lu Su, Bo Zhao, Heng Ji, and Jiawei Han. Faitcrowd: Fine grained truth discovery for crowdsourced data aggregation. In *KDD*, pages 745–754, 2015.

[Mortensen *et al.*, 2013] Jonathan M Mortensen, Mark A Musen, and Natalya F Noy. Crowdsourcing the verification of relationships in biomedical ontologies. In *AMIA Annual Symposium Proceedings (AMIA)*, page 1020, 2013.

[Motwani and Raghavan, 1995] Rajeev Motwani and Prabhakar Raghavan. *Randomized Algorithms*. Cambridge University Press, 1995.

[Prelec *et al.*, 2017] Dražen Prelec, H Sebastian Seung, and John McCoy. A solution to the single-question crowd wisdom problem. *Nature*, 541(7638):532–535, 2017.

[Venanzi *et al.*, 2014] Matteo Venanzi, John Guiver, Gabriella Kazai, Pushmeet Kohli, and Milad Shokouhi. Community-based bayesian aggregation models for crowdsourcing. In *WWW*, pages 155–164, 2014.

[Wauthier and Jordan, 2011] Fabian L. Wauthier and Michael I. Jordan. Bayesian bias mitigation for crowdsourcing. In *NIPS*, pages 1800–1808, 2011.

[Welinder *et al.*, 2010] Peter Welinder, Steve Branson, Pietro Perona, and Serge J. Belongie. The multidimensional wisdom of crowds. In *NIPS*, pages 2424–2432, 2010.

[Whitehill *et al.*, 2009] Jacob Whitehill, Paul Ruvolo, Tingfan Wu, Jacob Bergsma, and Javier Movellan. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *NIPS*, pages 2035–2043, 2009.

[Whitehill, 2019] Jacob Whitehill. How does knowledge of the auc constrain the set of possible ground-truth labelings? In *AAAI*, pages 5425–5432, 2019.

[Yin *et al.*, 2017] Li'ang Yin, Jianhua Han, Weinan Zhang, and Yong Yu. Aggregating crowd wisdoms with label-aware autoencoders. In *IJCAI*, pages 1325–1331, 2017.

[Zheng *et al.*, 2017] Yudian Zheng, Guoliang Li, Yuanbing Li, Caihua Shan, and Reynold Cheng. Truth inference in crowdsourcing: Is the problem solved? *Proceedings of the VLDB Endowment*, 10(5):541–552, 2017.

[Zhou and He, 2016] Yao Zhou and Jingrui He. Crowdsourcing via tensor augmentation and completion. In *IJCAI*, pages 2435–2441, 2016.

[Zhou *et al.*, 2012] Dengyong Zhou, John C. Platt, Sumit Basu, and Yi Mao. Learning from the wisdom of crowds by minimax entropy. In *NIPS*, pages 2195–2203, 2012.