18th International Conference on Information Fusion
Washington, DC - July 6-9, 2015

Boosting Crowdsourcing With Expert Labels: Local vs. Global Effects

Qiang Liu
CSAIL
MIT
Email: qliu1@csail.mit.edu

Alexander Ihler
Computer Science
University of California, Irvine
Email: ihler@ics.uci.edu

John Fisher III
CSAIL
MIT
Email: fisher@csail.mit.edu

***Abstract*—Crowdsourcing provides a cheap but efficient approach for large-scale data and information collection. However, human judgments are inherently noisy, ambiguous and sometimes biased, and should be calibrated by additional (usually much more expensive) expert or true labels. In this work, we study the optimal allocation of the true labels to best calibrate the crowdsourced labels. We frame the problem as a submodular optimization, and propose a greedy allocation strategy that exhibits an interesting trade-off between a *local effect*, which encourages acquiring true labels for the most uncertain items, and a *global effect*, which favors the true labels of the most "influential" items, whose information can propagate to help the prediction of other items. We show that exploiting and monitoring the global effect yields a significantly better selection strategy, and also provides potentially valuable information for other tasks such as designing stopping rules.**

I. Introduction

Crowdsourcing has emerged as a powerful approach for collecting data and information at large scales. The idea is to outsource tasks that are easy for humans but difficult for machines, sending them to online "crowd" workers who are given a relatively small monetary incentive. Crowdsourcing has been widely used in many application and scientific domains, including machine learning, human-computer interaction and social forecasting, to name only a few.

A major challenge in crowdsourcing is quality control. The (often anonymous) crowd workers have unknown and highly diverse levels of expertise, making it a critical problem to evaluate workers' performance and optimally combine their labels. In addition, human judgments are inherently noisy, often with significant individual biases; this is especially common in the estimates of continuous quantities, such as probabilities, product prices, and point spreads in sports, where people tend to give under- or over-estimates based on their personal experience. In these cases, it is necessary to calibrate the crowdsourcing results by incorporating some ground truth information or accurate labels from domain experts.

Because the expert or true labels are often much more expensive than the labels from the crowd, this raises an important problem of understanding the values of these valuable resources and hence making optimal use of them. In this work, we study the optimal allocation of the true labels to best calibrate the crowd labels for estimating continuous quantities. We frame the problem into a minimization of a conditional variance criterion, and establish its monotonic submodularity, enabling efficient approximation via greedy selection. We observe that our greedy selection rule decomposes into two terms that reflect a trade-off between a local effect and a global effect, where the local effect encourages acquiring true labels from the most uncertain items, which improve the performance in a local, myopic fashion, while the global effect favors the most "influential" items, whose true labels provide valuable information for decreasing the uncertainty on their associated workers' performance, and significantly improve the prediction of all the other items via a snowball effect. We show that it is critical to consider the global effect when allocating the true labels, especially in the initial stage when the number of acquired true labels is small, and the uncertainty on the workers' performance is relatively large.

The remainder of the paper is organized as follows. Section I-A discusses related work. Sections II, III and IV frame the optimization problem and establish its submodularity. Due to the intractability of the objective function, a Laplace approximation is used (Section III). Section V derives a greedy update rule and discusses the related concepts of global and local effects. We present our experimental results in Section VI and conclude the paper in Section VII.

A. Related Works

Some aspects of the value of ground truth information in crowdsourcing were recently studied in [1], which uses a set of *control items* with pre-labeled true answers,

and confidentially "seeds" them into workers' task sets to evaluate workers' performance and correct their biases. [1] studied the problem of determining the optimal number of control items to minimize the error rate, while assuming the control items are randomly assigned to the workers. Our setting is significantly different: we assume the true labels are acquired *after* the crowd labels are collected, providing the flexibility to optimize the assignments of the true labels to improve performance.

There exists a large body of literature on probabilistic modeling in crowdsourcing (e.g., [2], [3], [4], [5], [6]), which are able to jointly estimate workers' reliabilities and the item labels without any ground truth information. See also, for example, [7], [8], [6] for related theoretical discussions. However, these methods mostly work on discrete (such as binary) labeling problems, on which it is reasonable to make the assumption that the majority of workers (overall) will be correct. Our work focuses on crowdsourcing for continuous quantities, for which incorporating ground truth information is much more critical due to the inherent bias effects in human judgments.

There also exists a large body of work on optimal source allocation and online decision making in crowdsourcing (see e.g., [9], [10] and references therein); these works mostly focus on optimal assignments between the crowd workers and items, which differs from our setting of adding and allocating ground truth information to calibrate and improve the crowdsourced labels.

II. Problem Setting

Assume we have a set of items (or questions) $I = \{i\}$, each of which has a continuous quantity μ_i with an unknown true value of μ_i^* that we want to estimate (e.g., price, point spreads, GDP). Let $J = \{j\}$ be a set of crowd workers that we hire to estimate $\{\mu_i\}$. Let $\mathcal{G} = (I, J, E)$ be the bipartite assignment graph between the workers and items, that is, $(ij) \in E$ iff the jth worker answers the ith item. For $(ij) \in E$, let x_{ij} be the estimate of μ_i given by the jth worker.

We assume the workers' labels are generated by a model $p(x_{ij}|\mu_i, \nu_j)$, where in addition to μ_i, we have a parameter ν_j for each worker j, characterizing her expertise, bias, or any other relevant features. In this work, we assume $\{x_{ij}\}$ are generated by

$$x_{ij} = \mu_i + b_j + \sigma_j \xi_{ij}, \qquad \xi_{ij} \sim \mathcal{N}(0, 1), \tag{1}$$

where $\nu_j = \{b_j, \sigma_j\}$ with b_j denoting the bias and σ_j^2 the variance of the jth worker, respectively.

Note that the biases $\{b_j\}$ are not identifiable solely from the crowdsourced labels $\{x_{ij}\}$, and should be calibrated using additional expert or ground truth information. To be specific, we assume we have the option of checking the true labels of a set of items $C \subseteq I$ of size $|C| = K$ *after* the crowdsourcing labels $\{x_{ij}\}$ are collected. Our goal is to find the best set C such that the prediction error on the remaining items is minimized; using Bayesian inference, this is framed as

$$\min_{C\colon |C|\le K} \Big\{ \mathbb{E}(\mathrm{var}(\mu_{\neg C}|X, \mu_C) \mid X) \equiv \int \sum_{i\in\neg C} \big[\mu_i - \mathbb{E}(\mu_i|X, \mu_C)\big]^2 p(\mu \mid X) d\mu d\nu \Big\} \tag{2}$$

where $\neg C = I \setminus C$; this corresponds to minimizing the mean square error (MSE) when predicting the vector μ, and is known as A-optimality ("average" or trace) in the experiment design literature (e.g., [11]).

Another common objective is the conditional entropy

$$\min_{C\colon |C|\le K} H(\mu_{\neg C}|X, \mu_C), \tag{3}$$

which is equivalent to maximizing the marginal entropy $H(\mu_C|X)$. However, as we show later, the entropy objective (3) gives a myopic selection strategy, and performs significantly worse than the conditional variance objective.

III. Laplace Approximation

The posterior $p(\mu|X)$ is generally non-Gaussian, making it difficult to even evaluate the objective function in (2). We address this problem using a Laplace approximation. To be specific, we calculate the posterior mode

$$[\hat{\mu}, \hat{\nu}] = \arg\max_{[\mu,\nu]} \log p(\mu, \nu|X),$$

and approximate the posterior by a normal distribution,

$$p(\mu, \nu|X) \approx \mathcal{N}([\hat{\mu}, \hat{\nu}], H^{-1}),$$

where H is the negative Hessian matrix of the log-likelihood evaluated at $[\hat{\mu}, \hat{\nu}]$, that is,

$$\begin{bmatrix} H_{\mu\mu} & H_{\mu\nu} \\ H_{\nu\mu} & H_{\nu\nu} \end{bmatrix} = - \begin{bmatrix} \dfrac{\partial^2 \log p(\mu,\nu|X)}{\partial \mu^2} & \dfrac{\partial^2 \log p(\mu,\nu|X)}{\partial \mu \partial \nu} \\ \dfrac{\partial^2 \log p(\mu,\nu|X)}{\partial \nu \partial \mu} & \dfrac{\partial^2 \log p(\mu,\nu|X)}{\partial \nu^2} \end{bmatrix}.$$

Marginalizing over ν, we obtain

$$p(\mu|X) \approx \mathcal{N}(\hat{\mu}, Q^{-1}), \text{ where } Q = H_{\mu\mu} - H_{\mu\nu} H_{\nu\nu}^{-1} H_{\nu\mu},$$

and hence the value of the conditional variance,

$$\mathrm{var}(\mu_{\neg C}|X, \mu_C) \approx \mathrm{tr}(Q[\neg C]^{-1}), \qquad \forall \mu_C \in \mathbb{R}^{|C|},$$

where $Q[\neg C]$ is the submatrix of Q formed by the rows and columns in subset C. Correspondingly, the minimum conditional variance problem in (2) is approximated by

$$\max_{C:\,|C|\leq K} \{f_Q(C) \equiv -\mathrm{tr}(Q[\neg C]^{-1})\}, \qquad (4)$$

which is completely determined by the matrix Q.

IV. Submodularity and Greedy Selection

The subset selection problem in (4) is still intractable. In this section, we establish the submodularity and monotonicity of $f_Q(C)$, allowing efficient approximation by a greedy algorithm. Interestingly, we remark that $f_Q(C)$ is not always submodular for general semi-definite matrices Q, but is always submodular for the Q in our problem once the model $p(x|\mu,\nu)$ is not ill-posed (i.e., the negative Hessian H is positive definite). Our results are summarized as follows.

Proposition IV.1. *(1). For any positive semi-definite (PSD) matrix Q, the $f_Q(C)$ in (4) is non-decreasing, that is, $f_Q(C) \geq f_Q(C')$ if $C' \subseteq C$.*

(2). If Q is PSD and $Q_{ij} \leq 0$ for $i \neq j$ (i.e., it is a Stieltjes matrix, equivalently a symmetric M-matrix), then $\Sigma = Q^{-1}$ is element-wise nonnegative, and $f_Q(C)$ is a submodular function.

(3). For $Q = H_{\mu\mu} - H_{\mu\nu}H_{\nu\nu}^{-1}H_{\nu\mu}$ as used in our model, $f_Q(C)$ is submodular if H is positive semi-definite and $H_{\nu\nu}$ is non-singular.

Proof. (1) is an immediate result of the monotonicity of conditional variance, i.e., $\mathrm{var}(Y) \geq \mathbb{E}[\mathrm{var}(Y|Y')]$. (2) follows a more general result in Theorem 3 by [12]. For (3), we just need to note that $H_{\mu\mu}$ is a diagonal matrix, and hence $Q_{ij} = -H_{\mu_i\nu}H_{\nu\nu}^{-1}H_{\nu\mu_j} \leq 0$, $\forall i \neq j$. □

Remark 1. Proposition IV.1(2) suggests that $f_Q(C)$ is sub-modular if $\{\mu_i\}$ are non-negatively correlated with each other. Intuitively, this means that the prediction of μ_i should not hurt that of the others; more specifically, an improvement on predicting μ_i (corresponding to a decrease of the residual $\mu_i - \mu^*$) does not hurt the prediction of $\mu_{i'}$ for $i \neq i'$. Similar conditions for the submodularity of conditional variance-type objectives have been discussed by, for example, [13], [14], [15], in which similar *suppressor-free conditions* are derived.

Remark 2. The condition in Proposition IV.1(2) is a necessary one, as illustrated by the following counterexample by [12]:

$$Q = \begin{bmatrix} 5 & -12 & 9 \\ -12 & 33 & -24 \\ 9 & -24 & 19 \end{bmatrix},$$

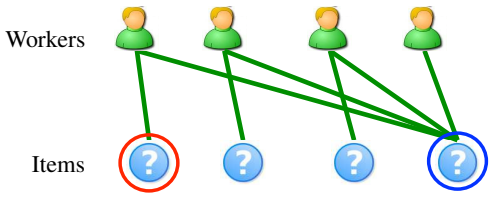

Fig. 1. Illustration of the local and global effects. Red circled: a very uncertain item (since answered by only one worker), whose true label has a large local effect (decreasing its own error), but a small global effect (in helping the prediction of the other items). Blue Circled: the most "influential" item (answered by many workers), whose true label has a large global effect (via propagating information through the workers associated with it and helping predict the other items), but a small local effect (since it is already relatively accurate).

where one can check that Q is positive definite but $f_Q(C)$ is not submodular.

V. Local vs. Global Effects

The monotonicity and submodularity of $f_Q(C)$ guarantee a $(1-1/e)$ approximation using a greedy algorithm. In this section, we derive the greedy update for (4) and show that it reflects an important trade-off between two effects: a *local effect* that encourages us to select the most uncertain items, and a *global effect* that encourages us to select the most "influential" items – those that can significantly help the predictions of other items.

Proposition V.1. *Let $\Sigma = \{\sigma_{ij}\} = Q[\neg C]^{-1}$, then for any $i \notin C$, we have*

$$f_Q(C \cup \{i\}) - f_Q(C) = \sigma_{ii} + \frac{\sum_{k\neq i}\sigma_{ik}^2}{\sigma_{ii}}.$$

Therefore, the optimal i^* for the greedy selection is

$$\begin{aligned} i^* &= \arg\max_i \{\sigma_{ii} + \frac{\sum_{k\neq i}\sigma_{ik}^2}{\sigma_{ii}}\} \\ &= \arg\max_i \{\sigma_{ii} + \sum_{k\neq i}\rho_{ik}^2\sigma_{kk}\}, \end{aligned} \qquad (5)$$

where $\rho_{ik} = \frac{\sigma_{ik}}{\sqrt{\sigma_{ii}\sigma_{kk}}}$ is the correlation between i and j.

The selection criterion $(\sigma_{ii} + \sum_{k\neq i}\rho_{ik}^2\sigma_{kk})$ gives an intuitive interpretation: the first term σ_{ii} counts the uncertainty of the ith item itself (a *local effect*), while the second term $\sum_{k\neq i}\rho_{ik}^2\sigma_{kk}$ assesses how much it would help in estimating other items if the true label of the ith item were acquired (a *global effect*). The second term will be large if the ith item has strong correlations with other difficult (high variance) items, and hence can help reduce their errors. This reflects a "propagation effect", that knowing the true labels of the ith item helps improve the estimates of the parameters $\{\nu_j : (ij) \in E\}$ for

workers that answered the ith item, which then helps improve the estimates of the other items answered by that set of workers; this information can then propagate on to more items, yielding a snowball effect.

The local and global effects are illustrated in the small graph shown in Figure 1. Assuming all the workers perform equally well, then the items labeled by the smallest number of workers (e.g., the one circled by red) are most uncertain, and have large σ_{ii}. On the other hand, the item labeled by the largest number of workers (blue circled) is more accurate itself (i.e., small σ_{ii}), but is most "influential" in that checking its answer can help evaluate the parameters of all the workers, and hence improve the estimates of all the other items. In this toy example, the different scores of the items are caused by their different degrees in the graph; in practice, items with the same degree may also have different scores if they are connected to workers that have different (posterior) parameter uncertainties.

Interestingly, as we demonstrate in our experiment in Section VI, the local and global effects dominate at different stages of the selection process. When the number of true labels acquired is small, the uncertainty on the workers' parameters (e.g., biases and variances) may be large, and one should select the most "influential" items to better evaluate the workers' parameters, exploiting the global effect. As more true labels are acquired, the uncertainty on the workers' parameters decreases, and the benefit of the global effects saturates; at this point, one should select the most uncertain items to exploit their local rewards.

The relative significance of the local vs. global effects can be quantified by the ratio of the two terms in (5),

$$\gamma = \frac{\sum_{k \neq i} \rho_{ik}^2 \sigma_{kk}}{\sigma_{ii}}, \tag{6}$$

which can be treated as a type of *value of information* (VOI) in the true labels to be acquired. Monitoring the ratio γ may also provide valuable guidance for designing optimal stopping rules or even more adaptive systems. For example, we may consider ceasing to collect more (expensive) true labels as the local effect begins to dominate (small γ), since the true labels will then essentially affect only the single item for which they are acquired, and may not be worth their cost. In addition, in this case, one may consider taking action to acquire more inexpensive labels from new crowd workers, which then makes the global effect return, making true labels more useful again. This provides a promising direction for constructing adaptive systems that automatically switch between crowd and expert labels and optimally trade off reliability and cost.

A. Entropy vs. Conditional Variance

An alternative item selection criterion is the conditional entropy shown in (3), which also yields a monotonic submodular optimization. However, a similar derivation shows that it corresponds to a more "myopic" greedy update of the form

$$i^* = \arg\max_i \{\sigma_{ii}\}, \tag{7}$$

which always selects the most uncertain item (*local effect*), while ignoring global effects. A similar myopic property of the entropy criterion was discussed in [16] for sensor network deployment, where they found that the entropy maximization tends to select the sensors at the border of the sensing field (which are locally most uncertain), leading to wasted sensing resources.

VI. Experiments

We illustrate our selection algorithms and the local vs. global effects on both simulated and real world datasets. We focus on a online setting throughout our experiments, where we update the Hessian matrix based on all the currently available information each time before we select and acquire a true label. This is tractable and should always be recommended in practice, since the updated Hessian provides a more accurate estimation, and is fast and cheap compared to the cost of acquiring true labels. Given the acquired true labels, we predict the remaining items using the posterior mode, and evaluate the result using the mean square error (MSE) w.r.t. the true values of the remaining items.

We compare the selection rule in (5) that includes both local and global effects (called `Local+Global`), with the more myopic selection rule in (7) that measures only the local effects (`Local`). We also implement a random selection rule where a random item is selected uniformly at each time (`Random`). We initialize all the selection rules with a common first item that has the largest degree in the assignment graph.

We start with a simulated dataset with 100 items and 100 workers, whose assignments are defined by a Bernoulli random bipartite graph in which each edge appears with probability 0.1, so that each item is labeled by 10 workers on average. We simulate the labels $\{x_{ij}\}$ according to the model in (1), where the true answers μ_i and the workers' biases b_j are drawn i.i.d. from standard normal distribution, and the workers' variances $\{\sigma_j\}$

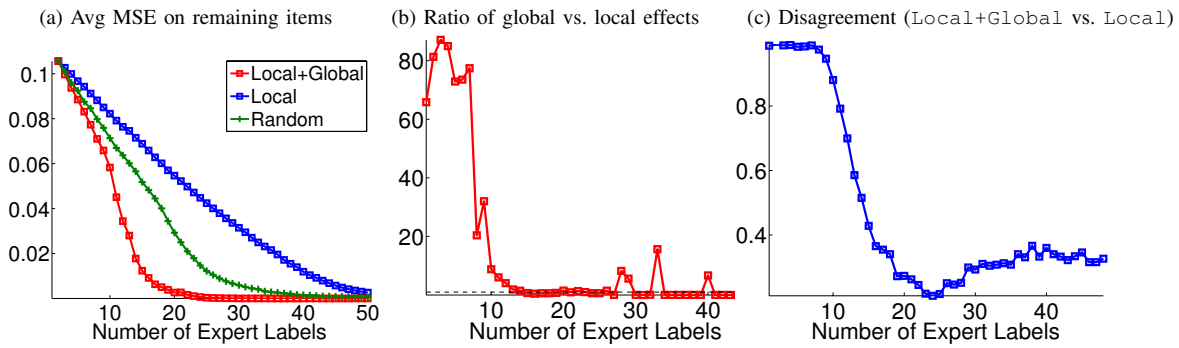

Fig. 2. Results on the simulated data. (a) The average MSE on the remaining items as the number of true labels increases; the monotonic decreasing of the average MSE is an indication of the "propagation" effect. (b) The ratio γ of the global and local effects as defined in (6). (c) The disagreement on the selected items between the `Local+Global` and `Local` rules.

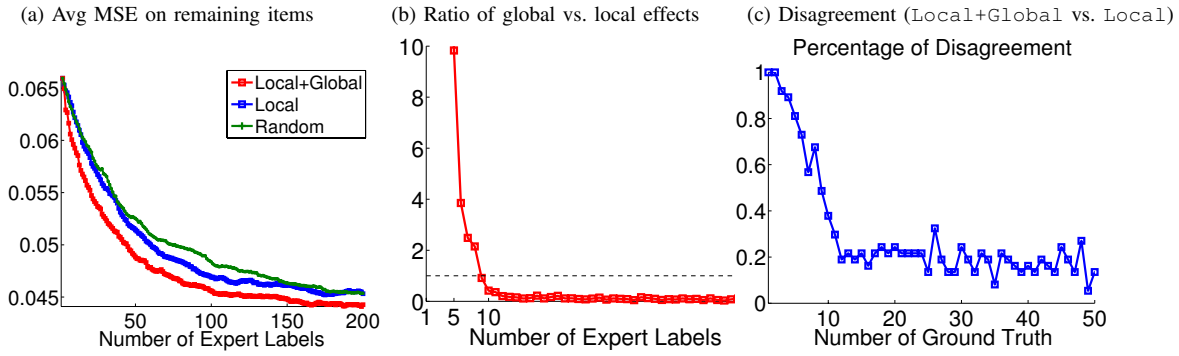

Fig. 3. Results on the real world dataset. The settings are the same as that in Fig 2.

are generated from an inverse Gamma distribution Inv-Gamma$(1, 1)$. The results are averaged over 100 random trials.

Fig 2 reports our results on the simulated data. Fig 2(a) shows that the average MSE on the *remaining items* decreases monotonically as we acquire more true labels; this is due to the *global effect* in which the true labels help evaluate the workers' parameters, and hence the prediction on the remaining items. We find that the `Local+Global` selection rule performs the best in terms of decreasing the MSE with a small number of true labels, while `Local` is even worse than `Random`, due to its myopic property.

In Fig 2(b), we show the ratio γ of the global and the local effects defined in (6) as we increase the number of true labels acquired by the `Local+Global` rule. As shown in Fig 2(b), γ starts at very high values when the number of true labels is small (corresponding to high global effect), but quickly reduces to small values as the number of true labels increases (corresponding to high local effect).

To assess the difference between the items selected by the `Local+Global` and the `Local` rules, we walk through the items selected by `Local+Global` sequentially, and calculate the next items returned by both `Local+Global` and `Local`. We then plot in Fig 2(c) the percentage of disagreement (across the 100 random trials) between `Local+Global` and `Local`. We find that the disagreement follows a similar trend to the ratio γ in Fig 2(b): it starts with 100% disagreement indicating that `Local+Global` and `Local` select completely different sets of items, and then switches to relatively low values as we increase the number of true labels, indicating that the difference between `Local+Global` and `Local` vanishes, due to the saturation of the propagation (global) effect.

We perform further experiments on a real world dataset collected by [17]. This dataset contains age estimates of 1002 face images given by 165 crowd workers on Amazon Mechanical Turk. We construct 100 random trials obtained by subsampling 500 images (items), and perform the same experiment as the simulated data. The results are shown in Fig 3(a)-(c), in which we observe a similar trend as in simulated data: the `Local+Global`

rule gives the best average MSE, and both the ratio γ and the disagreement between `Local+Global` and `Local` admit a similar transition from high values to low values.

VII. Conclusion

In this work, we study the optimal allocation of the true labels to best calibrate the crowdsourcing results for estimating continuous quantities. An important trade-off between a local effect and a global effect is discussed, and demonstrated to be critical for efficient and effective allocation of the true labels. Our observations provide critical insights for understanding the value of information of the true labels, raising the potential to make better use of both the (expensive) true labels and (inexpensive) crowd labels, and design more adaptive systems that could yield significantly better results.

VIII. Acknowledgement

This work is supported in part by VITALITE, which receives support from Army Research Office (ARO) Multidisciplinary Research Initiative (MURI) program (Award number W911NF-11-1-0391); NSF grants IIS-1065618 and IIS-1254071; and by the United States Air Force under Contract No. FA8750-14-C-0011 under the DARPA PPAML program.

References

[1] Q. Liu, M. Steyvers, and A. Ihler, "Scoring workers in crowdsourcing: How many control questions are enough?" in *Advances in Neural Information Processing Systems*, 2013, pp. 1914–1922.

[2] A. Dawid and A. Skene, "Maximum likelihood estimation of observer error-rates using the EM algorithm," *Applied Statistics*, pp. 20–28, 1979.

[3] Q. Liu, J. Peng, and A. Ihler, "Variational inference for crowdsourcing," in *Advances in Neural Information Processing Systems (NIPS)*, 2012, pp. 701–709.

[4] D. Zhou, J. Platt, S. Basu, and Y. Mao, "Learning from the wisdom of crowds by minimax entropy," in *Advances in Neural Information Processing Systems (NIPS)*, 2012, pp. 2204–2212.

[5] J. Whitehill, P. Ruvolo, T. Wu, J. Bergsma, and J. Movellan, "Whose vote should count more: Optimal integration of labels from labelers of unknown expertise," in *Advances in Neural Information Processing Systems (NIPS)*, 2009, pp. 2035–2043.

[6] Y. Zhang, X. Chen, D. Zhou, and M. I. Jordan, "Spectral methods meet EM: A provably optimal algorithm for crowdsourcing," in *Advances in Neural Information Processing Systems (NIPS)*, 2014, pp. 1260–1268.

[7] D. Karger, S. Oh, and D. Shah, "Iterative learning for reliable crowdsourcing systems," in *Advances in Neural Information Processing Systems (NIPS)*, 2011, pp. 1953–1961.

[8] C. Gao and D. Zhou, "Minimax optimal convergence rates for estimating ground truth from crowdsourced labels," *arXiv preprint arXiv:1310.5764*, 2013.

[9] X. Chen, Q. Lin, and D. Zhou, "Optimistic knowledge gradient policy for optimal budget allocation in crowdsourcing," in *Proceedings of the 30th International Conference on Machine Learning (ICML)*, 2013.

[10] A. Slivkins and J. W. Vaughan, "Online decision making in crowdsourcing markets: Theoretical challenges," *ACM SIGecom Exchanges*, vol. 12, no. 2, pp. 4–23, 2014.

[11] K. Chaloner and I. Verdinelli, "Bayesian experimental design: A review," *Statistical Science*, pp. 273–304, 1995.

[12] S. Friedland and S. Gaubert, "Submodular spectral functions of principal submatrices of a hermitian matrix, extensions and applications," *Linear Algebra and its Applications*, 2011.

[13] A. Das and D. Kempe, "Algorithms for subset selection in linear regression," in *Proceedings of the fortieth annual ACM symposium on Theory of computing*. ACM, 2008, pp. 45–54.

[14] Y. Ma, R. Garnett, and J. Schneider, "Submodularity in batch active learning and survey problems on gaussian random fields," in *NIPS 2012 Workshop on DISCML*, 2012.

[15] ——, "Σ-optimality for active learning on Gaussian random fields," in *Advances in Neural Information Processing Systems (NIPS)*, 2013, pp. 2751–2759.

[16] A. Krause and C. Guestrin, "Near-optimal nonmyopic value of information in graphical models," in *Conference on Uncertainty in Artificial Intelligence (UAI)*, July 2005.

[17] H. Han, C. Otto, X. Liu, and A. Jain, "Demographic estimation from face images: Human vs. machine performance," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014.