# A PAC Framework for Aggregating Agents' Judgments

**Hanrui Zhang**
Computer Science Department
Duke University
Durham, NC 27705
hrzhang@cs.duke.edu

**Vincent Conitzer**
Computer Science Department
Duke University
Durham, NC 27705
conitzer@cs.duke.edu

## Abstract

Specifying the objective function that an AI system should pursue can be challenging. Especially when the decisions to be made by the system have a moral component, input from multiple stakeholders is often required. We consider approaches that query them about their judgments in individual examples, and then aggregate these judgments into a general policy. We propose a formal learning-theoretic framework for this setting. We then give general results on how to translate classical results from PAC learning into results in our framework. Subsequently, we show that in some settings, better results can be obtained by working directly in our framework. Finally, we discuss how our model can be extended in a variety of ways for future research.

## Introduction

As AI systems are being broadly deployed in the world, the problem of specifying the *objective functions* that they pursue is becoming ever more complex. Simple objective functions that are sensible for the purpose of evaluating techniques in the lab often cause unwanted outcomes in practice. For example, one might try to simply minimize error rate in a speech recognition system, only to realize that the resulting system performs extremely well on the majority dialect but poorly on the minority dialect. This may socially not be acceptable.

More generally, AI systems increasingly need to make difficult tradeoffs. Should a self-driving car take an action that increases the risk for a nearby pedestrian, but reduces it for the occupant of the car? Should an algorithm that matches patients and donors in a kidney exchange prioritize younger patients even if this reduces the total number of matches? Given that today's AI systems do not have a broad understanding of the world, they cannot appropriately make these tradeoffs themselves; human input is required. But these problems are difficult for humans too, and they will not always agree.

One approach to doing so is the following (Conitzer et al. 2017; Noothigattu et al. 2018). Ask multiple human subjects what option they would choose in certain situations in the domain at hand; from this, learn a model for each of them, predicting what they would choose in other situations; and then, use techniques from *social choice* theory (Brandt et al. 2015) to aggregate this into a single decision policy for the AI system. For example, the "Moral Machine" website developed at MIT invites visitors to consider various scenarios in which the car will necessarily end up killing one group of people (and/or animals) or another, and ask them which they would choose.[1] Indeed, these responses have been aggregated into policies via a voting-based approach (Noothigattu et al. 2018). In a similar project, MTurkers were asked such comparison queries in the context of kidney exchanges and their responses were aggregated into a policy (Freedman et al. 2018).

Human subject responses are not freely available. They either require compensation (e.g., MTurkers) or a gamified design that makes it enjoyable to participate but may come at other costs.[2] This raises several related questions for the purpose of designing similar systems. How many responses should we aim to get? Do we need to recruit many distinct subjects, or does it suffice to have many responses from a few subjects? Do random queries suffice or should we actively design them? In this paper, we propose a formal learning-theoretic framework for this problem, building on the framework of Probably Approximately Correct (PAC) learning (Valiant 1984). In our model, there is a single *correct* concept $c^*$ that we attempt to learn.[3] Each human subject $j$ has her own concept $c^j$, which is a noisy estimate of the correct concept. When we ask subject $j$ query $x^{j,k}$, the subject will respond with $y^{j,k}$ according to her own concept $c^j$. From these responses, we aim to learn the correct concept $c^*$.



[1] This project has been much maligned for focusing on an unrealistic problem. But there is no doubt that self-driving cars will face *some* problems where they will have to make such tradeoffs, for example in deciding how aggressive to be in merging lanes (Sadigh et al. 2016).

[2] The Moral Machine team has been open about prioritizing the site going viral over other objectives.

[3] In the context of moral decision making, it is of course a matter of philosophical debate whether there is truly a *correct* concept, and we will not resolve this debate here. However, belief that such a correct concept exists is not necessary to use our methodology; one might also see it as merely a useful fiction to proceed with the analysis, as is sometimes done in voting theory as well (Elkind and Slinko 2015).

In the remainder of the paper, we first formally introduce the framework. We then give general results on how to translate classical results from PAC learning into results in our framework. Subsequently, we show that in some settings, better results can be obtained by working directly in our framework. Finally, we discuss how our model can be extended in a variety of ways for future research.

### Related Work

There is a rich body of research on learning from multiple entities, among which most closely related to ours are *collaborative learning* (Blum et al. 2017; Qiao 2018), *domain adaptation* (Mansour, Mohri, and Rostamizadeh 2009a; 2009b), and *learning from nearby sources* (Crammer, Kearns, and Wortman 2006; 2008). Collaborative learning concerns a setting in which multiple agents try to learn the same concept with respect to their own distinct distributions of data points; in contrast, in our setting, agents draw data points from the same global distribution, but label based on their own noisy concepts. Domain adaptation concerns a problem in which, given concepts with small errors w.r.t. different distributions, the goal is to aggregate the given concepts into a new one, with small error w.r.t. any mixture of the distributions. The setting is intrinsically different from ours, in the sense that (1) in our setting, all data points are from the same distribution (but are labeled according to different noisy concepts), and (2) our goal is to recover the ground truth instead of combining existing slightly erroneous ones into another slightly erroneous one. Learning from nearby sources considers a problem in which, given labeled data sets w.r.t. different sources and the "similarity" between these sources, the goal is to select a subset of data that will result in the best model for each of the sources. This problem, while being similar to ours in terms of the existence of multiple sources, studies estimation of many potentially unrelated concepts, instead of one ground truth.

Another line of work is *noise-tolerant learning* (Kearns 1998; Blum, Kalai, and Wasserman 2003; Natarajan et al. 2013). This problem is similar to ours in the sense that the goal is also to learn a target concept with noisy data points. However, in noise-tolerant learning, noise models are usually less structured: normally, all data points are corrupted in a more or less homogeneous way, at random (e.g., flipping each label independently with a fixed probability) or adversarially. Relatedly, Michael (2010) considers PAC learning when part of each data point is hidden, and Blum and Chalasani (1992) consider a setting where whenever a data point is labeled, it is labeled according to some concept drawn randomly or adversarially from a fixed set. In our model, in contrast, it is known which data points are from which agent, and each agent labels according to a single concept; the aggregation algorithm can, and should, make use of this information.

Judgment aggregation is a central topic in social choice theory, to which a significant body of research has been devoted (e.g., (Endriss, Grandi, and Porello 2012; Endriss 2016)). There, the problem is how to aggregate diverse judgments into a single outcome. In contrast, in this paper, we focus less on pure social-choice-theoretic aspects and more on statistical learning aspects.

| Agent | $x_1$ | $x_2$ | $x_3$ | $y$ |
|---|---|---|---|---|
| Alice | 1 | 0 | 0 | 1 |
| Alice | 1 | 0 | 1 | 1 |
| Alice | 1 | 1 | 0 | 1 |
| Bob | 1 | 0 | 0 | 0 |
| Bob | 1 | 0 | 1 | 1 |
| Bob | 0 | 0 | 1 | 0 |
| Charlie | 1 | 0 | 0 | 0 |
| Charlie | 1 | 1 | 0 | 1 |
| Charlie | 0 | 0 | 1 | 0 |

Table 1: Reports by the agents.

## The PAC Judgment Aggregation Model

Before formally introducing our model, first we discuss a concrete example.

**Example 1.** Suppose we want to learn a Boolean conjunction formula from the agents. Each of the 3 agents, Alice, Bob, and Charlie, has responded to 3 queries (see Table 1). Observe that every agent is consistent, in the sense that there is indeed a Boolean conjunction that conforms with her/his reports. In particular, Alice's formula is $x_1$, Bob's is $x_1 \wedge x_3$, and Charlie's is either $x_2$ or $x_1 \wedge x_2$. But our goal is to find an *aggregate* conjunction.

One way to aggregate the responses of the 3 agents is to find the Boolean conjunction that conflicts with as few reports as possible. The resulting aggregate conjunction is $x_1$, which conflicts only with Bob's first report and Charlie's first report. One may check that any other conjunction conflicts with more data points.

An alternative way of aggregating is to first compute the *majority label* on every feature vector, and then find the formula that conflicts with as few of the majority labels as possible. Among the 9 data points, there are only 4 distinct feature vectors: $(1, 0, 0)$, $(1, 0, 1)$, $(1, 1, 0)$ and $(0, 0, 1)$, and the majority judgments on these feature combinations are 0, 1, 1, and 0, respectively. Given these majority labels, there are 4 formulas that conflict with only 1 label: $x_1$, $x_2$, $x_1 \wedge x_2$, and $x_1 \wedge x_3$. Depending on the tie-breaking rule, any one of these 4 might be the aggregate formula.

From the above example, we see that the aggregate result depends heavily on the method used. To decide which method is best, we consider the following statistical model of the problem. As described in the introduction, this model assumes that every agent responds to data points (feature vectors) according to her own noisy estimate of the correct concept. Formally:

**Definition 1** (PAC Judgment Aggregation)**.** Given a distribution of data points $\mathcal{D}$, a concept class $\mathcal{C}$ where each $c \in \mathcal{C}$ maps each data point in the support of $\mathcal{D}$ to a label in $\{-1, 1\}$, and a noisy mapping $\nu : \mathcal{C} \to \Delta(\mathcal{C})$[4] which maps each concept to a distribution over concepts, the PAC

[4] $\Delta(\mathcal{C})$ denotes the family of distributions over $\mathcal{C}$.

judgment aggregation problem (with parameters $\varepsilon$ and $\delta$) is defined as follows:

1. A ground truth concept $c^* \in \mathcal{C}$ is chosen.
2. $m$ agents' concepts, denoted by $A = (c^1, \ldots, c^m)$ are generated by drawing independent and identically distributed (i.i.d.) samples from $\nu(c^*)$.
3. $\ell$ data points are generated for each agent $j$, by drawing i.i.d. samples $(x^{j,k})_{k\in[\ell]}$[5] and attaching to each sample $x^{j,k}$ the label $y^{j,k} = c^j(x^{j,k})$, together forming $((x^{j,k}, y^{j,k}))_{k\in[\ell]}$.
4. Based on the data points $(((x^{j,k}, y^{j,k}))_{k\in[\ell]})_{j\in[m]}$, a learning algorithm computes a hypothesis $h \in \mathcal{C}$. We are interested in algorithms that, with probability at least $1-\delta$, compute a hypothesis $h$ that satisfies:

$$\Pr_{x\sim\mathcal{D}}[c^*(x) \neq h(x)] \leq \varepsilon.$$

Definition 1 can be viewed as a natural generalization of the PAC learning model in the presence of noise in agents' judgments. In fact, if the noise vanishes, i.e., if the noisy mapping satisfies that for any $c \in \mathcal{C}$, $\Pr_{c_\nu\sim\nu(c)}[c = c_\nu] = 1$, then with probability 1, every agent has the same concept—the ground truth. In this special case, it does not help to query more than 1 agent, so that PAC aggregation coincides with PAC learning.

In the rest of this paper, we consider *finite* concept classes, and focus on *exact recovery* of the ground truth. That is, we consider algorithms that with probability $1-\delta$ output the correct concept, $h = c^*$. This is natural when $\mathcal{C}$ is finite: fixing a distribution $\mathcal{D}$, for any $\varepsilon$ smaller than the minimum probability that two concepts differ, the algorithm must output exactly the ground truth with high probability.[6]

## General Algorithms for PAC Aggregation

In this section, we present two paradigms to extend general sample complexity upper bounds from the traditional PAC learning setting to our PAC aggregation setting.

Before proceeding to the paradigms, we define the *distance* between two concepts, which helps simplify the notation. The intuition is straightforward: fixing the distribution $\mathcal{D}$ over data points, the probability that two concepts differ at a random data point induces a metric over the concept class. Formally, for two concepts $c_1$ and $c_2$, we define the distance between $c_1$ and $c_2$ to be

$$d(c_1, c_2) = \Pr_{x\sim\mathcal{D}}[c_1(x) \neq c_2(x)].$$

One can show that $d(\cdot,\cdot)$ is indeed a metric over $\mathcal{C}$; in particular, it satisfies the triangle inequality.

With the above definition of distance between concepts, in Definition 1, one may equivalently require the output hypothesis $h$ to satisfy: with probability at least $1-\delta$, $d(c^*, h) \leq \varepsilon$. In the rest of the paper, we will use this simpler notation whenever possible.

[5] $[\ell]$ denotes the set $\{1, 2, \ldots, \ell\}$.

[6] We assume that any two distinct concepts differ with positive probability, i.e., the probability of drawing a data point on which they disagree is positive.

### Occam's Razor

Recall the following basic result in PAC learning for when the concept class is finite:

**Proposition 1** (Occam's Razor for PAC learning (Folklore))**.** *With $O\left(\frac{\log(|\mathcal{C}|/\delta)}{\varepsilon}\right)$ data points, a concept $c$ that minimizes empirical error, which by definition is consistent with $c^*$ on all data points, satisfies $d(c, c^*) \leq \varepsilon$ with probability at least $1-\delta$.*

For exact recovery, when the minimum gap between any two concepts is $\alpha$ (i.e., $\min_{c_1\neq c_2} d(c_1, c_2) = \alpha$), setting $\varepsilon = \alpha/2$, Occam's Razor guarantees that $O\left(\frac{\log(|\mathcal{C}|/\delta)}{\alpha}\right)$ data points are sufficient to recover $c^*$. Note that the number of required data points increases as the size of the concept class $\mathcal{C}$ increases and the gap $\alpha$ decreases. In the rest of this subsection, we prove an extension of Occam's Razor to the PAC aggregation model, with an additional parameter characterizing the dependence on the strength of the noise (which is always 0 in PAC learning).

**Theorem 1** (Occam's Razor for PAC Aggregation)**.** *Suppose the noisy mapping $\nu$ satisfies the following two conditions:*

1. *For any $c$ and $c' \neq c$,*

$$\mathbb{E}_{c_\nu\sim\nu(c)}[d(c, c_\nu)] \leq \mathbb{E}_{c_\nu\sim\nu(c)}[d(c', c_\nu)] - \alpha_{\mathrm{OR}}.$$

2. *For any $c$ and $c'$ (where possibly $c = c'$), letting $c_\nu \sim \nu(c)$, $d(c', c_\nu)$ is sub-Gaussian with parameter $\beta$, i.e., for any $\lambda \in \mathbb{R}$,*

$$\mathbb{E}[\exp(\lambda(d(c', c_\nu) - \mathbb{E}[d(c', c_\nu)]))] \leq \exp(\beta^2\lambda^2/2).$$

*Then, using $m = O\left(\frac{\log(|\mathcal{C}|/\delta)\beta^2}{\alpha_{\mathrm{OR}}^2}\right)$ agents and $\ell m = O\left(\frac{\log(|\mathcal{C}|/\delta)}{\alpha_{\mathrm{OR}}^2}\right)$ data points in total, with probability at least $1-\delta$, a concept $c$ that minimizes the empirical error, defined as*

$$\mathrm{err}_{A,S}(c) = \frac{1}{\ell m}\sum_{j\in[m],k\in[\ell]} \mathbb{I}[c(x^{j,k}) \neq y^{j,k}],$$

*is the ground truth.*

Before proving the theorem, we briefly discuss the implications of its components:

**Dependence on $|\mathcal{C}|$.** The main message of the theorem is similar to that of the classical Occam's Razor result for PAC learning: with sufficiently many data points, with high probability, an empirical error minimizer recovers the ground truth. Both $m$ and $\ell m$ are required to increase proportionally to $\log(|\mathcal{C}|)$. In other words, if the concept class is small, then few agents and data points suffice to recover the ground truth. This is why the result is named after Occam's Razor—simpler models tend to explain the world better.

**Dependence on $\alpha_{\mathrm{OR}}$.** Condition 1 of the theorem states that in expectation, the noisy version $c_\nu$ of a concept $c$ is closer to $c$ than to any other concept. The gap $\alpha_{\mathrm{OR}}$ by which $c_\nu$ is closer to $c$ than to any other concept determines the number of agents and samples required to recover $c$. The larger the gap is, the fewer agents and samples are required. In particular, when the noise vanishes, for $c$, $c' \neq c$, and $c_\nu \sim \nu(c)$, we have $c = c_\nu$, and therefore $d(c, c_\nu) = d(c, c) = 0$, and $d(c', c_\nu) = d(c, c')$. In such cases, $\alpha_{\mathrm{OR}}$ is the minimum distance between two distinct concepts.

**Dependence on $\beta$.** Condition 2 of the theorem states that the distribution of noisy judgments is well concentrated, in the sense that the distance between $c_\nu$ and any concept is a sub-Gaussian random variable with parameter $\beta$. The smaller $\beta$ is, the fewer agents are required. This condition may appear rather strong at first sight. Nevertheless, since $d(c', c_\nu) \in [0, 1]$, it is always true that $d(c', c_\nu)$ is sub-Gaussian with parameter $1/2$ (i.e., $\beta \leq 1/2$). In natural noise models, the sub-Gaussian parameter $\beta$ goes to 0 as the noise vanishes, in which case the number of agents required is significantly reduced. For example, if the noisy version $c_\nu$ of $c$ is never too far away from $c$—i.e., for any $c$, $c_\nu \sim \nu(c)$, we have $d(c, c_\nu) \leq \gamma$ with probability 1—then we have $\beta \leq \min\{\gamma, 1/2\}$. This is because for any $c'$, by the triangle inequality $d(c', c_\nu) \in [d(c, c') - d(c, c_\nu), d(c, c') + d(c, c_\nu)] \subseteq [d(c, c') - \gamma, d(c, c') + \gamma]$, and any random variable with a support of length bounded by $2\gamma$ is sub-Gaussian with parameter $\gamma$. In particular, when there is no noise (i.e., $c_\nu = c$ with probability 1), we have $\beta = 0$, and $m = 1$ agent suffices[7]—corresponding to the traditional PAC learning setting.

**Tradeoff between $m$ and $\ell$.** When $\beta = o(1)$, the required number of agents $m$ is asymptotically smaller than the total number of data points $\ell m$. As a result, a tradeoff between $m$ and $\ell$ emerges: while keeping the probability of failure $\delta$ the same, one may use more data points per agent in order to decrease the number of agents needed (and vice versa), as long as the number of agents meets the minimum requirement.

**Learnability with $\ell = 1$ data point per agent.** Setting $\beta = 1/2$ in the second condition of Theorem 1, we immediately obtain the following simpler version:

**Corollary 1** (Occam's Razor, Simplified)**.** *Suppose the noisy mapping $\nu$ satisfies: for any $c$ and $c' \neq c$,*

$$\mathbb{E}_{c_\nu \sim \nu(c)}[d(c, c_\nu)] \leq \mathbb{E}_{c_\nu \sim \nu(c)}[d(c', c_\nu)] - \alpha_{\mathrm{OR}},$$

*Then, using $m = O\left(\frac{\log(|\mathcal{C}|/\delta)}{\alpha_{\mathrm{OR}}^2}\right)$ agents and $\ell = 1$ data point per agent, with probability $1 - \delta$, a concept that minimizes the empirical error is the ground truth.*

In other words, with sufficiently many agents, one data point per agent suffices to recover the ground truth.

[7] Here, we abuse notation to allow $1 = O(0)$.

**Computational efficiency.** Like the classical Occam's Razor result for PAC learning, Theorem 1 does not guarantee computational efficiency. This is because it is sometimes hard to compute a minimizer of the empirical error. Indeed, it is impossible to design general algorithms for PAC learning/aggregation that are "computationally efficient" without specifying the representation of the particular problem. In a later section, we present computationally efficient algorithms in several specific settings. We now turn back to prove Theorem 1.

*Proof of Theorem 1.* First we define some notation. Recall that $A = (c^1, \ldots, c^m)$ is the (ordered) set of agents. Let $S = ((x^{j,k})_k)_j$ be the (ordered) set of data points. Let

$$\mathrm{err}_A(c) = \mathbb{E}_{S \sim \mathcal{D}^{\ell m}}[\mathrm{err}_{A,S}(c)]$$
$$\mathrm{err}(c) = \mathbb{E}_{A \sim (\nu(c^*))^m}[\mathrm{err}_A(c)].$$

denote the expected empirical error of a concept $c$, with the latter expression also taking the expectation over the agents' concepts. For each $c \in \mathcal{C}$, we show that with high probability, the empirical error of $c$ is close to its expectation. Fixing $c$, consider two events:

- $\mathcal{E}_1$: $|\mathrm{err}_{A,S}(c) - \mathrm{err}_A(c)| \leq \frac{1}{6}\alpha_{\mathrm{OR}}$.
- $\mathcal{E}_2$: $|\mathrm{err}_A(c) - \mathrm{err}(c)| \leq \frac{1}{6}\alpha_{\mathrm{OR}}$.

We show that both events happen with high probability, so with high probability they happen simultaneously, in which case we have:

$$|\mathrm{err}_{A,S}(c) - \mathrm{err}(c)| \leq \frac{1}{3}\alpha_{\mathrm{OR}}.$$

First we bound the probability that $\mathcal{E}_1$ does not happen. Note that conditioned on $A$, $\mathrm{err}_{A,S}(c)$ is the average of $\ell m$ independent r.v.'s in $[0, 1]$, namely $\{\mathbb{I}[c(x^{j,k}) = y^{j,k}]\}_{j,k}$, with mean $\mathrm{err}_A(c)$. Applying the Hoeffding bound to $\mathrm{err}_{A,S}(c)$ conditioned on $A$, we have

$$\Pr\left[|\mathrm{err}_{A,S}(c) - \mathrm{err}_A(c)| > \frac{1}{6}\alpha_{\mathrm{OR}}\right] \leq 2e^{-\ell m \alpha_{\mathrm{OR}}^2/18} \leq \frac{\delta}{2|\mathcal{C}|}.$$

Now consider $\mathcal{E}_2$. Observe that $\mathrm{err}_A(c) = \frac{1}{m}\sum_j d(c, c^j)$ and $\mathrm{err}(c) = \mathbb{E}_{c_\nu \sim \nu(c^*)}[d(c, c_\nu)]$. It follows that $\mathrm{err}_A(c)$ is sub-Gaussian with parameter $\frac{\beta}{\sqrt{m}}$. Applying the Hoeffding bound for sub-Gaussian variables to $\mathrm{err}_A(c)$ gives

$$\Pr\left[|\mathrm{err}_A(c) - \mathrm{err}(c)| > \frac{1}{6}\alpha_{\mathrm{OR}}\right] \leq 2e^{-m\alpha_{\mathrm{OR}}^2/72\beta^2} \leq \frac{\delta}{2|\mathcal{C}|}.$$

Now taking the union bound over the two events and all concepts, with probability $1 - \delta$, for any $c \in \mathcal{C}$,

$$|\mathrm{err}_{A,S}(c) - \mathrm{err}(c)| \leq \frac{1}{3}\alpha_{\mathrm{OR}}.$$

In that case, for any $c \neq c^*$, we have

$$\mathrm{err}_{A,S}(c^*) \leq \mathrm{err}(c^*) + \frac{1}{3}\alpha_{\mathrm{OR}} \leq \mathrm{err}(c) - \frac{2}{3}\alpha_{\mathrm{OR}}$$
$$\leq \mathrm{err}_{A,S}(c) - \frac{1}{3}\alpha_{\mathrm{OR}} < \mathrm{err}_{A,S}(c).$$

In other words, $c^*$ minimizes the empirical error. □

## Majority Voting

When the support of $\mathcal{D}$, denoted by $|\mathcal{D}|$, is finite, another way to aggregate agents' judgments is to first estimate their individual concepts, and then let these estimated concepts take a majority vote for each $x$ in the support of $\mathcal{D}$ (denoted by $x \in \mathcal{D}$). Formally, we have:

**Theorem 2.** *When $|\mathcal{D}|$ is finite, if*

1. *the minimum distance between concepts is $\alpha_{\mathrm{MV}} > 0$, and*
2. *the noisy mapping satisfies, for some $\theta > 0$, that for any $x \in \mathcal{D}$, $\Pr_{c_\nu \sim \nu(c)}[c(x) = c_\nu(x)] \geq \frac{1}{2} + \theta$,*

*then with $m = O\left(\frac{\log(|\mathcal{D}|/\delta)}{\theta^2}\right)$ agents and $\ell = O\left(\frac{\log(|\mathcal{C}|\log|\mathcal{D}|/\theta\delta)}{\alpha_{\mathrm{MV}}}\right)$ data points per agent, the pointwise majority concept $h$ of the empirical error minimizers $h^j$ of each agent's data points, i.e., the concept $h$ such that for any $x \in \mathcal{D}$, $h(x) = \operatorname{sgn}\left(\sum_j h^j(x)\right)$, is the ground truth with probability $1 - \delta$.*

Again, before proceeding to the proof, we discuss the implications of the various components of the theorem.

**Dependence on $|\mathcal{C}|$.** As in the Occam's Razor result, the number of data points required by the majority-voting approach depends logarithmically on $|\mathcal{C}|$. However, unlike in Occam's Razor, the number of agents required for majority voting does not explicitly depend on $|\mathcal{C}|$. This can be partially explained by the fact that since $|\mathcal{D}| < \infty$, there are at most $2^{|\mathcal{D}|}$ possible concepts in total.

**Dependence on $|\mathcal{D}|$.** Although, in contrast to Occam's Razor, majority voting requires $\mathcal{D}$ to be finite, one may observe that the dependence on $|\mathcal{D}|$ is rather mild, i.e., logarithmic for $m$ and doubly logarithmic for $\ell$. This means majority voting remains relatively efficient even when $|\mathcal{D}|$ grows exponentially fast.

**Dependence on $\alpha_{\mathrm{MV}}$.** The parameter $\alpha_{\mathrm{MV}}$ here is similar to $\alpha_{\mathrm{OR}}$ in Occam's Razor, in the sense that when the noise vanishes, i.e., $\Pr_{c_\nu \sim \nu(c)}[c_\nu = c] = 1$, the two parameters are exactly the same. The key difference is that, for majority voting, $\alpha_{\mathrm{MV}}$ does not depend on the strength or form of the noise.

**Dependence on $\theta$.** Majority voting requires that the noisy concepts are more likely to agree with the ground truth. The parameter $\theta$ can be viewed as the distance between the expected judgment and a random guess, which in some sense characterizes the strength of the noise. The larger $\theta$ is, the easier to distinguish the noisy judgments from random guesses and extract the ground truth, and therefore fewer agents and data points are required.

Now we prove Theorem 2.

*Proof of Theorem 2.* We first show that with probability $1 - \frac{\delta}{2}$, the empirical error minimizers for the agents' data points coincide exactly with the agents' concepts. According to Proposition 1, with $\ell = O\left(\frac{\log(|\mathcal{C}|m/\delta)}{\alpha_{\mathrm{MV}}}\right) = O\left(\frac{\log(|\mathcal{C}|\log|\mathcal{D}|/\theta\delta)}{\alpha_{\mathrm{MV}}}\right)$ data points, $h^j = c^j$ with probability at least $1 - \frac{\delta}{2m}$. Taking the union bound over the $m$ agents, we conclude that with probability at least $1 - \frac{\delta}{2}$, it is the case that for every $j$, $h^j = c^j$.

Now, conditioning on the event that $h^j = c^j$ for any $j$, we show that majority voting recovers $c^*$ exactly with probability at least $1 - \frac{\delta}{2}$. Consider some $x \in \mathcal{D}$ where w.l.o.g. $c^*(x) = 1$. Since $\Pr_{c^j \sim \nu(c^*)}[c^j(x) = 1] \geq \frac{1}{2} + \theta$, the Chernoff bound gives

$$\Pr\left[\frac{1}{m}\sum_{j\in[m]} h^j(x) \leq 0\right] \leq \exp(-\theta^2 m/2).$$

With $m = O\left(\frac{\log(|\mathcal{D}|/\delta)}{\theta^2}\right)$, the right hand side is at most $\frac{\delta}{2|\mathcal{D}|}$. Taking the union bound over the support of $\mathcal{D}$, we have: with probability $1 - \frac{\delta}{2}$, for all $x \in \mathcal{D}$, $h(x) = c^*(x)$.

Recall that the immediately preceding argument conditioned on $h^j = c^j$ for all $j$, which happens with probability at least $1 - \frac{\delta}{2}$. Taking the union bound over the two parts, it follows that with probability at least $1 - \delta$, the pointwise majority concept $h$ is exactly the ground truth $c^*$.

Finally, we note that the definition of the pointwise majority concept does not guarantee its membership in $\mathcal{C}$. This is not a problem, because when the algorithm succeeds, we do have $h \in \mathcal{C}$ since $c^* \in \mathcal{C}$. When it fails and $h \notin \mathcal{C}$, we can let the algorithm output an arbitrary hypothesis in $\mathcal{C}$. □

## Occam's Razor vs. Majority Voting

Occam's Razor (Theorem 1) and majority voting (Theorem 2) rely on different parameters. As these parameters vary across settings, either approach may be preferred over the other. Two key differences are:

1. Theorem 2 requires both $|\mathcal{D}|$ and $|\mathcal{C}|$ to be finite, while Theorem 1 depends only on $|\mathcal{C}|$. This means majority voting may fail when the support of $\mathcal{D}$ is infinite.
2. The parameter $\alpha_{\mathrm{MV}}$ in Theorem 2 is always positive, but Theorem 1 puts a strong requirement on the gap $\alpha_{\mathrm{OR}}$ that in principle does not always hold. As an extreme (if unrealistic) example, suppose that the noisy mapping $\nu$ is actually a deterministic permutation of the concept space $\mathcal{C}$. In this case, the noisy version of a concept is clearly closer in expectation to another concept, namely the concept to which it is deterministically mapped.

# Tighter Bounds in Restricted Settings

In this section, we consider a linear model, where each concept $c$ is an $n$-dimensional vector in $\mathbb{R}^n$. For a data point $x$ (also in $\mathbb{R}^n$), the label that $c$ assigns to $x$ is determined by the sign of $c \cdot x$, denoted by $\operatorname{sgn}(c \cdot x)$.[8]

[8]For simplicity, we consider only $\mathcal{D}$ such that for any $c$, $\Pr_{x\sim\mathcal{D}}[c \cdot x = 0] = 0$. Alternatively, we may say when $c \cdot x = 0$, the label is 0. Our results are consistent with either convention.

**ALGORITHM 1:** Aggregation algorithm for binary judgments.

**Input :** $\ell$ labeled samples each from $m$ agents $\{(x^{j,k}, y^{j,k})\}_{j,k}$.
**Output:** $h$ which with high probability is equal to $c^*$.
**for** $i \in [n]$ **do**
  Let $h_i \leftarrow \operatorname{sgn}\left(\sum_{j,k} \operatorname{sgn}\left(x_i^{j,k}\right) y^{j,k}\right)$
**end**
Output $h = (h_i)_i$

One example of the linear model is pass/fail exams in which there are $n$ true-or-false questions. If concept $c^*$ is the ground truth solution, then the correct answer to question $i$ is $\operatorname{sgn}(c_i^*)$ and that question is worth $|c_i^*|$ points, i.e., the right answer to the $i$-th question increases the score by $|c_i^*|$, and the wrong one decreases the score by $|c_i^*|$. The support of $\mathcal{D}$ is $\{-1, 1\}^n$, where $x \in \mathcal{D}$ corresponds to a completed exam with a true or false (1 or $-1$) answer to every question. Hence, indeed, the *correct* score for completed exam $x$ is $c^* \cdot x$. A completed exam should get a grade of "Pass" if $c^* \cdot x > 0$ and "Fail" otherwise. Unfortunately, there are only imperfect evaluators—the $m$ judges—available to grade the exams, where judge $j$'s estimate of the ground truth is $c^j$. Thus, when judge $j$ evaluates an exam, $j$ will output "Pass" if and only if $c^j \cdot x > 0$. Each judge $j$ makes P/F decisions about $\ell$ sampled completed exams $\{x^{j,k}\}_{k \in [\ell]}$. Given these samples and labels, our aim is to recover the ground truth $c^*$. (In real applications, an "exam" can be any test with a vector of binary outcomes, and a "judge" any individual that makes judgments about what the overall result should be.)

We show that in certain settings, while the general Occam's Razor (Theorem 1 and Corollary 1) and majority voting (Theorem 2) results yield nontrivial sample complexity bounds, both the required number of agents $m$ and the total number of data points $\ell m$ can be significantly reduced using setting-specific methods. We further prove nearly tight lower bounds on $m$ and $\ell m$.

## Efficient Learning Algorithms

We give efficient algorithms for two restricted settings in the linear model. First, we consider the following setting: the attributes of instances are i.i.d., agents only make binary (positive/negative) judgments about (i.e., place binary weights on) each attribute, and each agent's judgment about an attribute $c_i^j$ is obtained independently by flipping the corresponding judgment in the ground truth, $c_i^*$, with some fixed probability. We show that as long as the distribution of the attributes is not extremely heavy-tailed and the noisy mapping preserves some information about the ground truth, then we can efficiently recover the ground truth. Formally, we prove the following theorem:

**Theorem 3** (Binary Judgments, I.I.D. Symmetric Distributions)**.** *Suppose that $\mathcal{C} = \{-1, 1\}^n$; for each $i \in [n]$, $\mathcal{D}_i = \mathcal{D}_0$ is a non-degenerate[9] symmetric distribution with bounded absolute third moment; and the noisy mapping with noise rate $\eta$ satisfies*

$$\nu(c)_i = \begin{cases} c_i, & \text{w.p. } 1-\eta \\ -1, & \text{w.p. } \eta/2 \\ 1, & \text{w.p. } \eta/2 \end{cases},$$

*Then, Algorithm 1 with $m = O\left(\frac{\ln(n/\delta)}{(1-\eta)^2}\right)$ agents and $\ell m = O\left(\frac{n \ln(n/\delta)}{(1-\eta)^2}\right)$ data points in total outputs the correct concept $h = c^*$ with probability at least $1-\delta$, in $O(\ell m n)$ time.*

The proof of Theorem 3, as well as those of Propositions 2 and 3 and Theorems 4, 5 and 6, are available in the appendix of the full version of the paper. To demonstrate the power of setting-specific algorithms, we compare Theorem 3 with the guarantees provided by Theorems 1 and 2.

**Performance of Occam's Razor.** Recall that the bounds of Theorem 1 rely on 3 parameters: $|\mathcal{C}|$, $\alpha_{\mathrm{OR}}$ and $\beta$. When $\mathcal{C} = \{-1, 1\}^n$, $|\mathcal{C}| = 2^n$, and one may show:

**Proposition 2.** *When $\mathcal{C}$ and $\nu$ satisfy the conditions in Theorem 3, and $\mathcal{D}$ is the uniform distribution over $\{-1, 1\}^n$, we have $\alpha_{\mathrm{OR}} = \Theta\left(\frac{1-\eta}{\sqrt{n}}\right)$.*

So, applying Theorem 1 to the above setting, we obtain that with $\ell m = O\left(\frac{n^2 \log(1/\delta)}{(1-\eta)^2}\right)$ data points (and $m \leq \ell m$ agents[10]), the empirical error minimizer is the ground truth with probability $1-\delta$. That is, even with the further restriction that the distribution $\mathcal{D}$ is uniform, this number is larger roughly by a factor of $n$ than the corresponding requirement of Theorem 3.

**Performance of majority voting.** Now we consider Theorem 2. The parameters involved are $|\mathcal{C}| = 2^n$, $|\mathcal{D}| = 2^n$, $\alpha_{\mathrm{MV}}$ and $\theta$. One may show:

**Proposition 3.** *When $\mathcal{C}$ and $\nu$ satisfy the conditions in Theorem 3, and $\mathcal{D}$ is the uniform distribution over $\{-1, 1\}^n$, we have $\alpha_{\mathrm{MV}} = \Theta(n^{-1/2})$ and $\theta = \Theta\left(\frac{1-\eta}{\sqrt{n}}\right)$.*

So, applying Theorem 2, majority voting requires $m = O\left(\frac{n^2 \log(1/\delta)}{(1-\eta)^2}\right)$ agents and $\ell m = O\left(\frac{n^{7/2} \log(n/(1-\eta)\delta)}{(1-\eta)^2}\right)$ data points in total. The number of agents required is at least the same as that of Occam's Razor, and the number of data points is even larger than that of Occam's Razor by a factor of $n^{3/2}$. To summarize, for the setting above, the setting-specific algorithm outperforms the general ones, and Occam's Razor outperforms majority voting.

We now proceed to an efficient algorithm for another setting, where the distributions of the attributes are independent (but not necessarily identical) Gaussian distributions, agents have discrete (instead of binary) judgments about each attribute, and the noisy mapping flips the sign of the judgment of the attribute independently with some fixed probability. In

[9] $\mathcal{D}$ is non-degenerate if for $X \sim \mathcal{D}$, $\Pr[X = 0] \neq 1$.

[10] The parameter $\beta$ appears hard to sharply estimate in this setting, which is why we do not compare the requirements on $m$ — it would be less meaningful to compare based on a loose estimation of $\beta$.

**ALGORITHM 2:** Aggregation algorithm for discrete judgments.

**Input :** $\ell$ labeled samples each from $m$ agents $\{(x^{j,k}, y^{j,k})\}_{j,k}$.
**Output:** $h$ which with high probability is equal to $c^*$.
**for** $i \in [n]$ **do**
  Let $s_i^2 \leftarrow \frac{1}{\ell m} \sum_{j,k} (x^{j,k})^2$
  Let $a_i \leftarrow \frac{1}{s_i} \frac{1}{\ell m} \sum_{j,k} \operatorname{sgn}\left(x_i^{j,k}\right) y^{j,k}$
**end**
Let $a_0 \leftarrow \min_i a_i$
**for** $i \in [n]$ **do**
  Let $h_i = p_i/q_i$ be the closest fraction to $a_i/a_0$, where $p_i, q_i \in \mathbb{Z} \cap [-d, d]$ and $p_i > 0$.
**end**
Output $h = (h_i)_i$

words, at the cost of further restricting the shape of the distribution $\mathcal{D}$, this setting allows more heterogeneity across the attributes, in the sense that the sizes of both the attributes' values and the judgments can vary. Formally, we prove:

**Theorem 4** (Discrete Judgments, Independent Gaussian Distributions)**.** *For constant $d$, suppose that $\mathcal{C} = (\mathbb{Z} \cap [-d, d])^n$; for each $i \in [n]$, $\mathcal{D}_i = \mathcal{N}(0, \sigma_i^2)$ is a Gaussian distribution with mean $0$ and standard deviation $\sigma_i \in [1, d]$; and the noisy mapping with noise rate $\eta$ satisfies*

$$\nu(c)_i = \begin{cases} c_i, & \text{w.p. } 1-\eta \\ -|c_i|, & \text{w.p. } \eta/2 \\ |c_i|, & \text{w.p. } \eta/2 \end{cases},$$

*Then, Algorithm 2 with $m = O\left(\frac{\ln(n/\delta)}{(1-\eta)^2}\right)$ agents and $\ell m = O\left(\frac{n \ln(n/\delta)}{(1-\eta)^2}\right)$ data points in total outputs a correct concept $h$, in the sense that there is some $\alpha \in \mathbb{R}^+$ such that $h = \alpha c^*$, with probability at least $1-\delta$, in $O(\ell m n)$ time.*

## Lower Bounds

Now that we have provided some efficient algorithms, we discuss the information-theoretic hardness of the linear model. We show that, even in highly restricted settings, the numbers of agents and data points required to recover the ground truth are roughly the same as the upper bounds obtained in Theorems 3 and 4.

In our first lower bound, we restrict attention to binary judgments. We show that even if the algorithm is given each agent $j$'s exact concept $c_j$—which is the most it could hope to learn for any number $\ell$ of samples per agent—it still needs $\Omega\left(\frac{\log(n/\delta)}{(1-\eta)^2}\right)$ agents to recover the ground truth. Formally:

**Theorem 5** (Lower Bound: Number of Agents)**.** *If $\mathcal{C} = \{-1, 1\}^n$ and $\mathcal{D}_i$ is the uniform distribution over $\{-1, 1\}$, then any algorithm that outputs the correct concept with probability $1-\delta$ requires $m = \Omega\left(\frac{\log(n/\delta)}{(1-\eta)^2}\right)$ agents.*

We note that this lower bound extends immediately to the discrete-judgments setting in Theorem 4. It follows that the requirement on the number of agents in Theorems 3 and 4 (i.e., $m = O\left(\frac{\log(n/\delta)}{(1-\eta)^2}\right)$) is tight up to a constant factor.

Next, we prove a lower bound on the total number of data points (or agents) when each agent reports only 1 data point.

**Theorem 6** (Lower Bound: Total Number of Data Points)**.** *If $\mathcal{C} = \{-1, 1\}^n$, $\mathcal{D}_i$ is the uniform distribution over $\{-1, 1\}$ or the standard Gaussian distribution $\mathcal{N}(0, 1)$, the number of data points per agent is $\ell = 1$, and $1 - \eta = \Omega(n^{-1/2})$, then any algorithm that outputs the correct concept with constant probability requires $\ell m = \Omega\left(\frac{n}{(1-\eta)^2}\right)$ data points in total.*

Given Theorem 6, the requirement in Theorems 3 and 4 (i.e., $\ell m = O\left(\frac{n \log(n/\delta)}{(1-\eta)^2}\right)$) is tight up to a factor of $\log(n/\delta)$, when $\ell = 1$. We suspect that the loss of the $\log(n/\delta)$ factor is due to an intrinsic limitation of the mutual information argument used to prove the theorem.

## Discussion

Our objective in this paper has been to introduce a general framework for learning from multiple agents' judgments, and to illustrate the types of results that can be obtained in this framework. Much work remains to be done.

While we have provided some results that are quite general, we have also shown that sometimes stronger results can be obtained for more specific settings, with specific assumptions on the distribution of instances, concept class, and noise in the agents' perceived concepts. This raises the question of whether similar results can be obtained under different assumptions that may fit particular applications better. It should be noted that in many cases, an efficient PAC aggregation algorithm also implies an efficient algorithm in the simpler traditional PAC learning model for the same concept class. However, PAC learning of some natural concept classes is known to be computationally hard. For example, efficient PAC learning algorithms for 3-term DNFs do not exist unless RP = NP (see (Kearns, Vazirani, and Vazirani 1994)); hence, the same is true for efficient PAC aggregation algorithms in this setting.

A natural variant would allow *active learning*, where instead of receiving the labels for random instances, the algorithm can choose which instances it would like to have labeled (and by which agents).[11] Indeed, if we consider settings such as that of the MIT Moral Machine, this is potentially a sensible model, because those who designed the website control which instances are shown. It would thus be desirable to extend the theory of active learning to this setting. For example, the majority voting paradigm, which we use to translate passive learning algorithms to our setting, works for active learning too. One may show:

**Theorem 7** (informal)**.** *Given a PAC active learning algorithm for concept class $\mathcal{C}$, under distribution $\mathcal{D}$ of data points, where $|\mathcal{C}|$ and $|\mathcal{D}|$ are finite and the conditions in Theorem 2 are satisfied, with sufficiently many agents and data points, the algorithm that (1) runs the PAC active learning algorithm on each individual agent's data points, and*

[11] There are, in fact, multiple variants; for example, one variant would require that every agent sees the same instances.

*then (2) outputs the concept that is the the pointwise majority of the concepts learned in the first step, recovers the ground truth with high probability.*

One downside of the active learning approach in this context is that the precise instances shown would depend on (for example) the concept class, so that one could not simply re-use the same data for a different concept class. If gathering data from human subjects is costly (in terms of both finances and effort), then using random instances would allow for better re-use (also by other researchers) of the data. Also, of course any upper bound in the passive model still applies in the active model.

One could imagine other modes of interaction yet. For example, we may show an instance to three agents and ask them to come to a consensus judgment (cf. (Goel and Lee 2016; Fain et al. 2017)). The space of possible designs is extremely broad, and there is much potential for both theoretical and empirical work.

## Acknowledgements

We are thankful for support from NSF under awards IIS-1814056 and IIS-1527434. We also thank anonymous reviewers for helpful comments.

## References

Berry, A. C. 1941. The accuracy of the gaussian approximation to the sum of independent variates. *Transactions of the american mathematical society* 49(1):122–136.

Blum, A., and Chalasani, P. 1992. Learning switching concepts. In *Proceedings of the fifth annual workshop on Computational learning theory*, 231–242. ACM.

Blum, A.; Haghtalab, N.; Procaccia, A. D.; and Qiao, M. 2017. Collaborative pac learning. In *Advances in Neural Information Processing Systems*, 2392–2401.

Blum, A.; Kalai, A.; and Wasserman, H. 2003. Noise-tolerant learning, the parity problem, and the statistical query model. *Journal of the ACM (JACM)* 50(4):506–519.

Brandt, F.; Conitzer, V.; Endriss, U.; Lang, J.; and Procaccia, A. D. 2015. *Handbook of Computational Social Choice*. Cambridge University Press.

Conitzer, V.; Sinnott-Armstrong, W.; Borg, J. S.; Deng, Y.; and Kramer, M. 2017. Moral decision making frameworks for artificial intelligence. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 4831–4835.

Crammer, K.; Kearns, M.; and Wortman, J. 2006. Learning from data of variable quality. In *Advances in Neural Information Processing Systems*, 219–226.

Crammer, K.; Kearns, M.; and Wortman, J. 2008. Learning from multiple sources. *Journal of Machine Learning Research* 9(Aug):1757–1774.

Elkind, E., and Slinko, A. 2015. Rationalizations of voting rules. In Brandt, F.; Conitzer, V.; Endriss, U.; Lang, J.; and Procaccia, A. D., eds., *Handbook of Computational Social Choice*. Cambridge University Press. chapter 8.

Endriss, U.; Grandi, U.; and Porello, D. 2012. Complexity of judgment aggregation. *Journal of Artificial Intelligence Research* 45:481–514.

Endriss, U. 2016. Judgment aggregation. In F. Brandt, V. Conitzer, U. E. J. L., and Procaccia, A., eds., *Handbook of Computational Social Choice*. Cambridge University Press.

Esseen, C.-G. 1945. Fourier analysis of distribution functions. A mathematical study of the Laplace-Gaussian law. *Acta Mathematica* 77(1):1–125.

Fain, B.; Goel, A.; Munagala, K.; and Sakshuwong, S. 2017. Sequential deliberation for social choice. In *Proceedings of the Thirteenth Conference on Web and Internet Economics (WINE-17)*, 177–190.

Freedman, R.; Borg, J. S.; Sinnott-Armstrong, W.; Dickerson, J. P.; and Conitzer, V. 2018. Adapting a kidney exchange algorithm to align with human values. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*.

Goel, A., and Lee, D. T. 2016. Towards large-scale deliberative decision-making: Small groups and the importance of triads. In *Proceedings of the Seventeenth ACM Conference on Economics and Computation (EC)*, 287–303.

Kearns, M. J.; Vazirani, U. V.; and Vazirani, U. 1994. *An introduction to computational learning theory*. MIT press.

Kearns, M. 1998. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM (JACM)* 45(6):983–1006.

Laurent, B., and Massart, P. 2000. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics* 1302–1338.

Mansour, Y.; Mohri, M.; and Rostamizadeh, A. 2009a. Domain adaptation: Learning bounds and algorithms. In *22nd Conference on Learning Theory, COLT 2009*.

Mansour, Y.; Mohri, M.; and Rostamizadeh, A. 2009b. Domain adaptation with multiple sources. In *Advances in neural information processing systems*, 1041–1048.

Michael, L. 2010. Partial observability and learnability. *Artificial Intelligence* 174(11):639–669.

Natarajan, N.; Dhillon, I. S.; Ravikumar, P. K.; and Tewari, A. 2013. Learning with noisy labels. In *Advances in neural information processing systems*, 1196–1204.

Noothigattu, R.; Gaikwad, S. N. S.; Awad, E.; D'Souza, S.; Rahwan, I.; Ravikumar, P.; and Procaccia, A. D. 2018. A voting-based system for ethical decision making. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*.

Qiao, M. 2018. Do outliers ruin collaboration? *arXiv preprint arXiv:1805.04720*.

Sadigh, D.; Sastry, S.; Seshia, S. A.; and Dragan, A. D. 2016. Planning for autonomous cars that leverage effects on human actions. In *Robotics: Science and Systems (RSS-16)*.

Valiant, L. G. 1984. A theory of the learnable. *Communications of the ACM* 27(11):1134–1142.

# Omitted Proofs

## Proof of Theorem 3

The proof relies on the following technical lemma, which essentially says the "sensitivity" of the sign of the sum of $n$ i.i.d. r.v.'s to each summand is about $\Omega(n^{-1/2})$.

**Lemma 1** (Sensitivity Lemma)**.** *Given i.i.d. copies $X_1, \ldots, X_n$ of any non-degenerate symmetric random variable with bounded absolute third moment (i.e., $\mathbb{E}[|X_i|^3] < \infty$),*

$$\Pr\left[|X_1| \geq \left|\sum_{2\leq i\leq n} X_i\right|\right] = \Omega(n^{-1/2}).$$

*And as a corollary,*

$$\Pr\left[\operatorname{sgn}\left(\sum_{1\leq i\leq n} X_i\right) \neq \operatorname{sgn}\left(\sum_{2\leq i\leq n} X_i\right)\right] = \Omega(n^{-1/2}).$$

*Proof.* To prove the lemma, note that Central Limit Theorem guarantees that $\sum_{2\leq i\leq [n]} X_i$ tends to a normal distirbution, and the convergence of the CDF is uniform. We refer to the following lemma regarding the speed of this convergence.

**Lemma 2** (Convergence Rate of CLT, Theorem 7, Chapter IV (Esseen 1945))**.** *Consider i.i.d. copies $Y_1, \ldots, Y_m$ of any nonzero (i.e., $\Pr[Y_i = 0] = 0$) symmetric random variable with unit second moment (i.e., $\mathbb{E}[Y_i^2] = 1$) and bounded absolute third moment (i.e., $\mathbb{E}[|Y_i|^3] < \infty$). Let $\Phi(x)$ be the CDF of a standard Gaussian distribution $\mathcal{N}(0,1)$, $F_m$ be the CDF of $m^{-1/2}\sum_{i\in[m]} Y_i$. We have*

$$\lim_{m\to\infty} \max_x \sqrt{m}|F_m(x) - \Phi(x)| \leq \frac{1}{\sqrt{2\pi}}.$$

*There is equality iff $Y_i$ is a symmetric Bernoulli variable.*

We now prove the sensitivity lemma. First, if $\Pr[X_i = 0] > 0$, we partially realize the randomness of whether each $X_i$ being 0 or not for $2 \leq i \leq n$. Suppose $m$ out of the $n-1$ variables are not 0. For $i \in [m]$, let $Y_i$ be an r.v. which has the distribution of $X_i$ conditioned on $X_i \neq 0$. Note that with probability $\Omega(1)$, $m \geq \frac{1}{2}\Pr[X_i \neq 0]n$. We argue the inequality for each partial realization, and take expectation (ignoring the contribution when $m = o(n)$) to establish the lemma.

For notational simplicity we condition on the partial realization implicitly when bounding probabilities involving $\{Y_i\}$. Observe that $\Pr[Y_i = 0] = 0$. Also w.l.o.g. we may assume $\mathbb{E}[Y_i^2] = 1$, since $X_i$ has bounded absolute third moment and therefore second moment. According to Lemma 2, when $Y_i$ is not Bernoulli, there is $\varepsilon > 0$ such that for $x \geq 0$,

$$\Pr\left[\sum_{i\in[m]} Y_i \leq x\right] \geq \frac{1}{\sqrt{2\pi m}}\int_{-\infty}^{x} e^{\frac{-t^2}{2m}}\,\mathrm{d}t - \left(\frac{1}{\sqrt{2\pi m}} - \frac{\varepsilon}{\sqrt{m}}\right) - o(m^{-1/2}).$$

And so,

$$\begin{aligned}
\Pr\left[\left|\sum_{i\in[m]} Y_i\right| \leq 1\right] &\geq \frac{1}{\sqrt{2\pi m}}\int_{-1}^{1} e^{\frac{-t^2}{2m}}\,\mathrm{d}t - \frac{2}{\sqrt{2\pi m}} + \frac{2\varepsilon}{\sqrt{m}} - o(m^{-1/2}) \\
&= \frac{2}{\sqrt{2\pi m}} + o(m^{-1/2}) - \frac{2}{\sqrt{2\pi m}} + \frac{2\varepsilon}{\sqrt{m}} - o(m^{-1/2}) \\
&= \frac{2\varepsilon}{\sqrt{m}} - o(m^{-1/2}).
\end{aligned}$$

On the other hand, since $\mathbb{E}[Y_i^2] = 1$, $\Pr[|Y_i| \geq 1] > 0$, and therefore $\Pr[|X_1| \geq 1] > 0$. It follows immediately that

$$\begin{aligned}
\Pr\left[|X_1| \geq \left|\sum_{2\leq i\leq n} X_i\right|\right] &= \Pr\left[|X_1| \geq \left|\sum_{i\in[m]} Y_i\right|\right] \\
&\geq \Pr[|X_1| \geq 1] \cdot \Pr\left[\left|\sum_{i\in[m]} Y_i\right| \leq 1\right] \\
&\geq \Pr[|X_1| \geq 1] \cdot \left(\frac{2\varepsilon}{\sqrt{m}} - o(m^{-1/2})\right) \\
&= \Omega(m^{-1/2}).
\end{aligned}$$

Now suppose $Y_i$ is symmetric Bernoulli. It is well known that for any integer $x$,

$$\Pr\left[-1 \le \sum_{i\in[m]} Y_i \le 0\right] \ge \sqrt{\frac{2}{\pi m}} - o(m^{-1/2}).$$

It follows that

$$\begin{aligned}
\Pr\left[|X_1| \ge \left|\sum_{2\le i\le n} X_i\right|\right] &= \Pr\left[|X_1| \ge \left|\sum_{i\in[m]} Y_i\right|\right] \\
&\ge \Pr[X_1 = 1] \cdot \Pr\left[-1 \le \sum_{i\in[m]} Y_i \le 0\right] \\
&\ge \Pr[X_1 = 1] \cdot \left(\sqrt{\frac{2}{\pi m}} - o(m^{-1/2})\right) \\
&= \Omega(m^{-1/2}).
\end{aligned}$$

Now with constant probability $m \ge \frac{1}{2}\Pr[X_i \ne 0]n$, in which case $\Omega(m^{-1/2}) = \Omega(n^{-1/2})$ uniformly. Taking expectation over the partial realization of 0's yields the lemma immediately.

□

Now we turn back to prove Theorem 3. W.l.o.g. assume $c^* = (1, 1, \ldots, 1)$. For some $i \in [n]$, consider the probability that $c_i^* = 1$ is correctly estimated, i.e., $\Pr[h_i = c_i^*]$. We first bound the probability that the agents $\{c_i^j\}_{j\in[m]}$ together constitute a good approximation of $c_i^*$. Let $\mathcal{E}_1$ be the event, that

$$\frac{1}{m}\sum_{j\in[m]} c_i^j \ge 1 - \eta - t_1,$$

where $t_1$ is a parameter to be determined later. By Hoeffding bound,

$$\Pr[\mathcal{E}_1] \ge 1 - e^{-mt_1^2/2}.$$

Now for fixed $\{c_j\}_j$, we consider how well $\frac{1}{\ell m}\sum_{j,k} x_i^{j,k} y^{j,k}$ approximates $\frac{1}{m}\sum_{j\in[m]} c_i^j$. Let $\{-i\} = [n] \setminus \{i\}$. First note that for any $j$, $k$, $\mathbb{E}[x_i^{j,k} y^{j,k} \mid \{c^j\}] = \frac{\alpha c_i^j}{\sqrt{n}} - o(n^{-1/2})$, where $\alpha = \alpha(n) = \Omega(1)$. In fact, w.l.o.g. assuming $c_i^j = 1$, Lemma 1 implies

$$\begin{aligned}
\mathbb{E}[x_i^{j,k} y^{j,k} \mid \{c^j\}] &= \Pr\left[\operatorname{sgn}\left(\sum_{u\in\{-i\}} x_u^{j,k} c_u^j\right) \ne \operatorname{sgn}\left(\sum_{u\in[n]} x_u^{j,k} c_u^j\right)\right] \\
&\quad \times \mathbb{E}\left[\operatorname{sgn}\left(\sum_{u\in[n]} x_u^{j,k} c_u^j\right) - \operatorname{sgn}\left(\sum_{u\in\{-i\}} x_u^{j,k} c_u^j\right) \middle| \operatorname{sgn}\left(\sum_{u\in\{-i\}} x_u^{j,k} c_u^j\right) \ne \operatorname{sgn}\left(\sum_{u\in[n]} x_u^{j,k} c_u^j\right)\right] \\
&= \Omega(n^{-1/2}) \cdot \Theta(1) = \frac{\alpha c_i^j}{\sqrt{n}}.
\end{aligned}$$

Summing over $j$ and $k$, we have

$$\mathbb{E}\left[\frac{1}{\ell m}\sum_{j,k} x_i^{j,k} y^{j,k} \middle| \{c_j\}\right] = \frac{1}{m}\sum_{j\in[m]} \frac{\alpha c_i^j}{\sqrt{n}}.$$

Let $\mathcal{E}_2$ be the event, that

$$\frac{1}{\ell m}\sum_{j,k} x_i^{j,k} y^{j,k} \ge \mathbb{E}\left[\frac{1}{\ell m}\sum_{j,k} x_i^{j,k} y^{j,k} \middle| \{c^j\}\right] - t_2,$$

where $t_2$ is a parameter to be determined later. Again by Hoeffding bound,

$$\Pr[\mathcal{E}_2] \ge 1 - e^{-\ell m t_2^2/2}.$$

Our goal is to show that for small enough $t_1$ and $t_2$, when $\mathcal{E}_1$ and $\mathcal{E}_2$ happen simultaneously, the estimation

$$h_i = \operatorname{sgn}\left(\sum_{j,k} x_i^{j,k} y^{j,k}\right) = c_i^* = 1.$$

In other words,

$$\frac{1}{\ell m}\sum_{j,k} x_i^{j,k} y^{j,k} > 0.$$

For this reason, we choose

$$t_1 = \frac{1-\eta}{3},\ t_2 = \frac{\alpha(1-\eta)}{3\sqrt{n}}.$$

Now when both events happen, for $n$ large enough, we have

$$\frac{1}{m}\sum_{j\in[m]} c_i^j \geq \frac{2(1-\eta)}{3},$$

and

$$\begin{aligned}
\frac{1}{\ell m}\sum_{j,k} x_i^{j,k} y^{j,k} &\geq \mathbb{E}\left[\frac{1}{\ell m}\sum_{j,k} x_i^{j,k} y^{j,k} \,\middle|\, \{c^j\}\right] - t_2 \\
&= \frac{1}{m}\sum_{j\in[m]} \frac{\alpha c_i^j}{\sqrt{n}} - \frac{\alpha(1-\eta)}{3\sqrt{n}} \\
&\geq \frac{2\alpha(1-\eta)}{3\sqrt{n}} - \frac{\alpha(1-\eta)}{3\sqrt{n}} \\
&\geq \frac{\alpha(1-\eta)}{3\sqrt{n}} \\
&> 0.
\end{aligned}$$

At the same time, we need to ensure that the probabilities of both events are large enough. To be precise, we need $e^{-mt_1^2/2} \leq \frac{\delta}{2n}$ and $e^{-\ell m t_2^2/2} \leq \frac{\delta}{2n}$. Plugging in $t_1$ and $t_2$ and taking $\ln$, we get

$$\begin{aligned}
\frac{m(1-\eta)^2}{18} &\geq \ln(2n/\delta), \\
\frac{\ell m \alpha^2 (1-\eta)^2}{18n} &\geq \ln(2n/\delta).
\end{aligned}$$

It suffices to set $m = \frac{18\ln(2n/\delta)}{(1-\eta)^2}$ and $\ell = \frac{n}{\alpha^2}$.

Applying union bound on $\mathcal{E}_1$ and $\mathcal{E}_2$, we see that for large enough $m$ and $\ell$, with probability at most $\frac{n}{2\delta} + \frac{n}{2\delta} = \frac{n}{\delta}$ we make a mistake for the $i$-th bit in the estimation. A union bound over the $n$ bits immediately upper bounds the probability of failure by $\delta$, thereby concluding the proof.

## Proof of Proposition 2

W.l.o.g. fix $c = (1, \ldots, 1)$. The closest concept to $c_\nu \sim \nu(c)$ is obtained by flipping one coordinate of $c$. We therefore consider $c' = (-1, 1, \ldots, 1)$, and show that

$$\mathbb{E}_{c_\nu \sim \nu(c)}[d(c', c_\nu) - d(c, c_\nu)] = \Theta((1-\eta)/\sqrt{n}).$$

We couple $d(c, c_\nu)$ and $d(c', c_\nu)$ in the following way:

1. With probability $\eta$, the first entry of $c_\nu$ is uniformly random. Conditioned on this, the expectations of $d(c, c_\nu)$ and $d(c', c_\nu)$ are the same. This part therefore contributes nothing to the difference.
2. With probability $1-\eta$, the first entry of $c_\nu$ is 1. Conditioned on this, following the argument in Lemma 1, the expected difference is precisely

$$\frac{1}{2}\Pr_{x\sim\mathcal{D}}\left[\left|\sum_{2\leq i\leq n} x_i\right| = 1\right] = \Theta(n^{-1/2}).$$

The proposition follows by taking the total expectation.

## Proof of Proposition 3

The bound on $\alpha$ follows from Proposition 2 by setting $\eta = 0$. For the bound on $\theta$, clearly the worst case happens when $c \cdot x = 1$. Suppose $c = (1, 1, \ldots, 1)$, and $x$ has $k$ $-1$'s and $k+1$ 1's, where $2k+1 = n$. For simplicity assume $x_1 = \cdots = x_k = x_n = 1$, and $x_{k+1} = \cdots = x_{2k} = -1$. Observe that

$$\Pr_{c_\nu \sim \nu(c)} \left[ \sum_{1 \le i \le n-1} c_\nu x > 0 \right] = \Pr_{c_\nu \sim \nu(c)} \left[ \sum_{1 \le i \le n-1} c_\nu x < 0 \right],$$

and

$$\Pr_{c_\nu \sim \nu(c)} \left[ \sum_{1 \le i \le n-1} c_\nu x = 0 \right] = \Theta(n^{-1/2}).$$

When $\sum_{1 \le i \le n-1} c_\nu x \neq 0$, $\operatorname{sgn}(c_\nu \cdot x)$ is independent from $(c_\nu)_n$. When $\sum_{1 \le i \le n-1} c_\nu x = 0$, $\operatorname{sgn}(c_\nu \cdot x) = 1$ with probability $\frac{1}{2} + \frac{1}{2}(1-\eta)$. Overall, $\Pr[c_\nu \cdot x > 0] = \frac{1}{2} + \Theta\left(\frac{1-\eta}{\sqrt{n}}\right)$.

## Proof of Theorem 4

First we prove an analog of Lemma 1 in the Gaussian case.

**Lemma 3.** *Given independent Gaussian variables $X_1, \ldots, X_n$ with means $0$ and variances $\sigma_1^2, \ldots, \sigma_n^2$ where $\sigma_i \in [1, d]$ for some constant $d$,*

$$\Pr\left[ |X_1| \ge \left| \sum_{2 \le i \le n} X_i \right| \right] = \frac{2\sigma_1}{\pi s} + o(n^{-1/2}).$$

*where $s^2 = \sum_{2 \le i \le n} \sigma_i^2 = \Theta(\sqrt{n})$. And as a corollary,*

$$\Pr\left[ \operatorname{sgn}\left( \sum_{1 \le i \le n} X_i \right) \neq \operatorname{sgn}\left( \sum_{2 \le i \le n} X_i \right) \right] = \frac{\sigma_1}{\pi s} + o(n^{-1/2}).$$

*Proof.* Let $X = X_1$ and $Y = \sum_{2 \le i \le n} X_i$. Note that $X$ and $Y$ are two independent Gaussian variables with means $0$ and variances $\sigma_1^2$ and $s^2$ respectively.

$$\begin{aligned}
\Pr[|X| \ge |Y|] &= \int_x \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{x^2}{2\sigma_1^2}} \int_{|y| \le |x|} \frac{1}{\sqrt{2\pi}s} e^{-\frac{y^2}{2s^2}} \,\mathrm{d}x\mathrm{d}y \\
&= \int_{-n^{1/3}}^{n^{1/3}} \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{x^2}{2\sigma_1^2}} \int_{|y| \le |x|} \frac{1}{\sqrt{2\pi}s} e^{-\frac{y^2}{2s^2}} \,\mathrm{d}x\mathrm{d}y + o(n^{-1/2}) \\
&= \int_{-n^{1/3}}^{n^{1/3}} \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{x^2}{2\sigma_1^2}} \frac{1}{\sqrt{2\pi}s} \cdot 2|x| \,\mathrm{d}x + o(n^{-1/2}) \\
&= -2 \int_0^{n^{2/3}} \frac{\sigma_1}{\pi s} e^{-\frac{x^2}{2\sigma_1^2}} \,\mathrm{d}\left( -\frac{x^2}{2\sigma_1^2} \right) + o(n^{-1/2}) \\
&= \frac{2\sigma_1}{\pi s} + o(n^{-1/2}).
\end{aligned}$$

□

Now we turn back to prove Theorem 4. Recall that when proving the binary version, Theorem 3, for the algorithm to succeed for the $i$-th entry, we require two events to hold simultaneously. Here we extend this scheme. More specifically, for $i \in [n]$ we consider the following events:

- $\mathcal{E}_1$:

$$\left| \frac{1}{m} \sum_{j \in [m]} c_i^j - (1-\eta) c_i^* \right| \le \frac{1-\eta}{10(d+1)^3}.$$

- $\mathcal{E}_2$:

$$\left|\frac{1}{\ell m}\sum_{j,k}\operatorname{sgn}(x_i^{j,k})y^{j,k}-\mathbb{E}\left[\frac{1}{\ell m}\sum_{j,k}\operatorname{sgn}(x_i^{j,k})y^{j,k}\,\middle|\,\{c^j\}\right]\right|\le\frac{1-\eta}{10\pi(d+1)^3\Sigma},$$

where

$$\Sigma^2=\sum_{i\in[n]}\sigma_i^2(c_i^*)^2.$$

- $\mathcal{E}_3$:

$$|s_i^2-\sigma_i^2|=\left|\frac{1}{\ell m}\sum_{j,k}(x_i^{j,k})^2-\sigma_i^2\right|\le\frac{1}{10(d+1)^4},$$

and as a result,

$$|s_i-\sigma_i|=\frac{|s_i^2-\sigma_i^2|}{s_i+\sigma_i}\le\frac{1}{10(d+1)^4}\le\frac{\sigma_i}{10(d+1)^4}.$$

We claim that if each of the three events fails with probability at most $\frac{\delta}{3n}$, then the algorithm fails with probability at most $\delta$. We first prove the above claim, and then show the probability of failure is indeed small.

Note that with probability at least $1-\delta$, all three events happen for all $n$ entries simultaneously. We show that for any $i_1,i_2\in[n]$, $h_{i_1}/h_{i_2}=c_{i_1}^*/c_{i_2}^*$. W.l.o.g. assume $i_1=1$ and $i_2=2$. Consider $a_1$ and $a_2$ first. Observe that

$$\begin{aligned}
\left|\frac{1}{2}a_i-\frac{(1-\eta)c_i^*}{\pi\Sigma}\right| &= \left|\frac{1}{s_i}\frac{1}{2\ell m}\sum_{j,k}\operatorname{sgn}(x_i^{j,k})y^{j,k}-\frac{(1-\eta)c_i^*}{\pi\Sigma}\right| \\
&\le \left|\frac{1}{s_i}\frac{1}{2\ell m}\mathbb{E}\left[\sum_{j,k}\operatorname{sgn}(x_i^{j,k})y^{j,k}\,\middle|\,\{c^j\}\right]-\frac{(1-\eta)c_i^*}{\pi\Sigma}\right|+\frac{1-\eta}{10\pi(d+1)^3\Sigma} &(1)\\
&= \left|\frac{1}{s_i}\frac{1}{m}\sum_j\frac{\sigma_i c_i^j}{\pi\Sigma}-\frac{(1-\eta)c_i^*}{\pi\Sigma}\right|+\frac{1-\eta}{10\pi(d+1)^3\Sigma}+o(n^{-1/2}) &(2)\\
&\le \left(1+\frac{1}{10(d+1)^4}\right)\frac{1}{\pi\Sigma}\left|\frac{1}{m}\sum_j c_i^j-(1-\eta)c_i^*\right|+\frac{1}{10(d+1)^4}\left|\frac{(1-\eta)c_i^*}{\pi\Sigma}\right| \\
&\quad+\frac{1-\eta}{10\pi(d+1)^3\Sigma}+o(n^{-1/2}) &(3)\\
&\le \frac{1-\eta}{9\pi(d+1)^3\Sigma}+\frac{1-\eta}{10\pi(d+1)^3\Sigma}+\frac{1-\eta}{10\pi(d+1)^3\Sigma}+o(n^{-1/2}) &(4)\\
&\le \frac{1-\eta}{3\pi(d+1)^3\Sigma}+o(n^{-1/2}). &(5)
\end{aligned}$$

(1) is because event $\mathcal{E}_1$ happens; (2) follows from Lemma 3 and the fact that (when w.l.o.g. $c_i^j>0$)

$$\mathbb{E}[\operatorname{sgn}(x_i^{j,k})y^{j,k}\mid\{c^j\}]=2\Pr\left[\operatorname{sgn}\left(\sum_{u\in\{-i\}}x_u^{j,k}c_u^j\right)\ne\operatorname{sgn}\left(\sum_{u\in[n]}x_u^{j,k}c_u^j\right)\right];$$

(3) is because event $\mathcal{E}_3$ happens; (4) is because event $\mathcal{E}_2$ happens.

Now consider $|a_1/a_2 - c_1^*/c_2^*|$. If for any $i$ and $t > 0$ small enough, $\left|\frac{\pi\Sigma}{2(1-\eta)}a_i - c_i^*\right| \le t$, then

$$\begin{aligned}\left|\frac{a_1}{a_2} - \frac{c_1^*}{c_2^*}\right| &= \left|\left(\frac{\pi\|c^*\|}{1-\eta}a_1\right) / \left(\frac{\pi\|c^*\|}{1-\eta}a_2\right) - \frac{c_1^*}{c_2^*}\right| \\ &\le \left|\frac{c_1^* + t}{c_2^* - t} - \frac{c_1^*}{c_2^*}\right| \\ &= \left|\frac{t(c_1^* + c_2^*)}{(c_2^* - t)c_2^*}\right| \\ &\le \left|\frac{t(d+1)}{1-t}\right| \\ &\le \frac{3t(d+1)}{4}.\end{aligned}$$

Putting (5) into the above inequality (i.e., $t = \frac{1}{3(d+1)^3} + o(1)$),

$$\left|\frac{a_1}{a_2} - \frac{c_1^*}{c_2^*}\right| \le \frac{1}{4(d+1)^2} + o(1).$$

On the other hand, for any $p_1, q_1, p_2, q_2 \in \mathbb{Z} \cap [-d, d]$, as long as $p_1/q_1 \neq p_2/q_2$, we have $|p_1/q_1 - p_2/q_2| \ge \frac{1}{d(d-1)}$. In other words, for large enough $n$ and any $i_1, i_2 \in [n]$, when the three events happen, the closest feasible fraction to $a_{i_1}/a_{i_2}$ is exactly $c_{i_1}^*/c_{i_2}^*$. Also note that $a_i > 0 \iff c_i^* > 0$, which means $h$ recovers $c^*$ up to a positive scalar.

Next we show that indeed, with large enough $m$ and $\ell$, $\mathcal{E}_1$, $\mathcal{E}_2$ and $\mathcal{E}_3$ happen with high probability. Consider $\mathcal{E}_1$ first. Note that $c_i^j \in [-d, d]$ and $\mathbb{E}[c_i^j] = (1-\eta)c_i^*$. Hoeffding bound gives immediately

$$\Pr\left[\left|\frac{1}{m}\sum_{j\in[m]} c_i^j - (1-\eta)c_i^*\right| \ge \frac{1-\eta}{10(d+1)^3}\right] \le 2\exp\left(-\frac{m\left(\frac{1-\eta}{(d+1)^3}\right)^2}{2d^2}\right).$$

When $m \ge 10^3 \frac{d^8 \log(n/\delta)}{(1-\eta)^2}$, the probability above is smaller than $\frac{\delta}{3n}$. For $\mathcal{E}_2$, note that $\Sigma^2 = \sum_{i\in[n]} \sigma_i^2 (c_i^*)^2 \le \sum_{i\in[n]} d^2 \cdot d^2 = d^4 n$, so $\Sigma \le d^2\sqrt{n}$. Hoeffding bound gives

$$\begin{aligned}\Pr\left[\left|\frac{1}{\ell m}\sum_{j,k} \operatorname{sgn}(x_i^{j,k}) y^{j,k} - \mathbb{E}\left[\frac{1}{\ell m}\sum_{j,k} \operatorname{sgn}(x_i^{j,k}) y^{j,k} \,\middle|\, \{c^j\}\right]\right| \ge \frac{1-\eta}{10\pi(d+1)^3\Sigma}\right] &\le 2\exp\left(-2\ell m\left(\frac{1-\eta}{10\pi(d+1)^3\Sigma}\right)^2\right) \\ &\le 2\exp\left(-2\ell m\left(\frac{1-\eta}{10\pi(d+1)^3 d^2\sqrt{n}}\right)^2\right).\end{aligned}$$

When $\ell m \ge 10^5 \frac{d^{10} n \log(n/\delta)}{(1-\eta)^2}$, the above probability is smaller than $\frac{\delta}{3n}$. Now consider $\mathcal{E}_3$. Note that $\frac{1}{\ell m}\sum_{j,k}(x_i^{j,k})^2$ is a chi-squared variable (scaled by a factor of $\frac{\sigma_i^2}{\ell m}$) with $\ell m$ degrees of freedom. The following lemma establishes concentration for chi-squared distributions.

**Lemma 4** (Concentration of $\chi^2$ Distributions (Laurent and Massart 2000))**.** *Let $U$ be a $\chi^2$ statistic with $D$ degrees of freedom. For any positive $x$,*

$$\begin{aligned}\Pr[U - D \ge 2\sqrt{Dx} + 2Dx] &\le \exp(-x), \\ \Pr[D - U \le 2\sqrt{Dx}] &\le \exp(-x).\end{aligned}$$

*As a corollary, for any $t > 0$,*

$$\Pr[|U - D| \ge 2\sqrt{Dt} + 2Dt] \le 2\exp(-t).$$

Applied to $\frac{1}{\ell m}\sum_{j,k}(x_i^{j,k})^2$ with $t = 10\log(n/\delta)$, when $\ell m \ge 10^5 \frac{d^7 n \log(n/\delta)}{(1-\eta)^2}$, Lemma 4 yields

$$\Pr\left[|s_i^2 - \sigma_i^2| \ge \frac{1}{10(d+1)^4}\right] \le \Pr\left[|s_i^2 - \sigma_i^2| \ge 2\frac{1-\eta}{100d^{3.5}\sqrt{n}} + 2\frac{(1-\eta)^2}{10^4 n d^7}\right] \le \frac{\delta}{3n}.$$

We conclude that for large enough $n$, Algorithm 2 succeeds with probability at least $1-\delta$ as long as $m \ge 10^3 \frac{d^5 \log(n/\delta)}{(1-\eta)^2}$ and $\ell m \ge 10^5 \frac{d^7 n \log(n/\delta)}{(1-\eta)^2}$.

## Proof of Theorem 5

Consider the uniform prior over $\mathcal{C}$. Assume, for the sake of a lower bound, that we know exactly the noisy classifier $c^j$ assigned to each agent $j$. Since Coordinates of the classifiers are independent, consider the problem of estimating $c_i^*$ with probability $1-p$. Clearly the MLE is to take the majority vote of all agents, i.e., $h_i = \operatorname{sgn}\left(\sum_j c_i^j\right)$. Consider the binomial r.v. $X = \sum_j c_i^j$. The estimation $h_i$ is wrong iff $X < \frac{1}{2}m$. We now estimate the probability of a wrong estimation.

$$\begin{aligned}
p \geq \Pr\left[X < \frac{1}{2}m\right] &= \sum_{0 \leq k < \frac{1}{2}m} \binom{m}{k} \left(\frac{\eta}{2}\right)^k \left(1 - \frac{\eta}{2}\right)^{m-k} \\
&\geq \sum_{\frac{1}{2}m - \sqrt{m} \leq k < \frac{1}{2}m} \binom{m}{k} \left(\frac{\eta}{2}\right)^k \left(1 - \frac{\eta}{2}\right)^{m-k} \\
&\geq C_1 \sum_{\frac{1}{2}m - \sqrt{m} \leq k < \frac{1}{2}m} \sqrt{\frac{m}{k(m-k)}} \frac{m^m}{k^k (m-k)^{m-k}} \left(\frac{\eta}{2}\right)^k \left(1 - \frac{\eta}{2}\right)^{m-k} \\
&\geq C_2 \sum_{\frac{1}{2}m - \sqrt{m} \leq k < \frac{1}{2}m} \frac{1}{\sqrt{m}} 2^m \left(\frac{\eta}{2}\right)^k \left(1 - \frac{\eta}{2}\right)^{m-k} \\
&= C_2 \sum_{\frac{1}{2}m - \sqrt{m} \leq k < \frac{1}{2}m} \frac{1}{\sqrt{m}} \eta^k (2-\eta)^{m-k} \\
&\geq C_2 \sum_{\frac{1}{2}m - \sqrt{m} \leq k < \frac{1}{2}m} \frac{1}{\sqrt{m}} \eta^{m/2} (2-\eta)^{m/2} \\
&= C_2 \left(1 - (1-\eta)\right)^{m/2} \left(1 + (1-\eta)\right)^{m/2} \\
&= C_2 \left(1 - (1-\eta)^2\right)^{m/2} \\
&\geq C_3 \exp\left(-\frac{m(1-\eta)^2}{2}\right).
\end{aligned}$$

$C_1$, $C_2$ and $C_3$ above are positive constants. It follows that $m = \Omega\left(\frac{\log(1/p)}{(1-\eta)^2}\right)$. Now for $p$, we need

$$(1-p)^n \geq 1 - \delta.$$

For $p$ small enough,

$$1 - pn \sim (1-p)^n \geq 1 - \delta \Longrightarrow p = O\left(\frac{\delta}{n}\right).$$

Putting these together we get

$$m = \Omega\left(\frac{\log(n/\delta)}{(1-\eta)^2}\right).$$

## Proof of Theorem 6

Consider the mutual information between a new data point $(x, y)$ and $c^*$. We will show that given any prior on $c^*$ (in particular, when there is no randomness in $c^*$), $H(x, y\,;\, c^*) = O\left((1-\eta)^2\right)$. It follows immediately the number of data points needed given the uniform prior is $\ell m = \Omega\left(\frac{n}{(1-\eta)^2}\right)$.

Now we bound the mutual information. Observe that

$$H(x, y\,;\, c^*) = H(x, y) - H(x, y \mid c^*) \leq H(x) + 1 - H(x, y \mid c^*) = H(x) + 1 - (H(x \mid c^*) + H(y \mid x, c^*)) = 1 - H(y \mid x, c^*).$$

The problem then reduces to bounding $H(y \mid x, c^*)$. W.l.o.g. assume $c^* = (1, \ldots, 1)$. Consider first the case where $\mathcal{D}$ is the uniform distribution over $\{-1, 1\}^n$. For some $x_0$ with $k$ entries being 1 and $n-k$ being $-1$, we consider $H(y \mid x = x_0, c^*)$. Note that now the only source of randomness is in the noisy classifier $c$ which is independent to everything else. Consider the random variable $c \cdot x_0$. Observe that $c \cdot x_0$ is the sum of $n$ independent Bernoulli variables (whose value can be 1 or $-1$), consisting of two groups. The first group corresponds to the 1's of $x_0$, where each variable has mean $1 - \eta$, and the second group corresponds to the $-1$'s, with mean $\eta - 1$. We now apply Central Limit Theorem to analyze the distribution of $c \cdot x_0$. Berry-Esseen Theorem (Berry 1941) states that, with mild assumptions which are true for our purpose, the error of CLT is

$O(k^{-1/2})$ where $k$ is the number of i.i.d. r.v.'s summed over. It follows that for $\frac{1}{3}n \le k \le \frac{2}{3}n$, the CDF of $c \cdot x_0$ is uniformly approximated by $\mathcal{N}((2k-n)(1-\eta), n)$ with an error of $O(n^{-1/2})$. W.l.o.g. assume $2k \ge n$, so $(2k-n)(1-\eta) \ge 0$, and let $n$ be odd, so $c \cdot x_0$ is never 0. We have

$$\begin{aligned}\frac{1}{2} \ge \Pr[c \cdot x_0 \le 0] &\ge \frac{1}{2} - \Pr[0 \le c \cdot x_0 \le (2k-n)(1-\eta)] \\ &= \frac{1}{2} - \int_0^{(2k-n)(1-\eta)} \frac{1}{\sqrt{2\pi n}} \exp\left(-\frac{(t-(2k-n)(1-\eta))^2}{2n}\right) \mathrm{d}t - O(n^{-1/2}) \\ &\ge \frac{1}{2} - \frac{(2k-n)(1-\eta)}{\sqrt{2\pi n}} - O(n^{-1/2}).\end{aligned}$$

Let $k = \frac{1}{2}(n + t\sqrt{n})$ where $t \ge 1$, we have

$$\frac{1}{2} - C_1(1-\eta)t \le \frac{1}{2} - \frac{t(1-\eta)}{\sqrt{2\pi}} - O(n^{-1/2}) \le \Pr[c \cdot x_0 \le 0] \le \frac{1}{2},$$

where $C_1$ is a positive constant since $(1-\eta) = \Omega(n^{-1/2})$. In other words, conditioned on $x_0$, $y$ is 1 w.p. at most $1/2 + C_1(1-\eta)t$, and $-1$ w.p. at least $1/2 - C_1(1-\eta)t$.

Now recall that a Bernoulli variable with mean $1/2 + \varepsilon$ has entropy $1 - \varepsilon^2$ for small $\varepsilon$. Let $|x|$ be the number of 1's in $x$. It follows that

$$H(y \mid ||x| - 0.5n| \le t\sqrt{n}, c^*) \ge 1 - C_2(1-\eta)^2 t^2,$$

where $C_2$ is a positive constant. Also observe that for $t > 0$,

$$\Pr[||x| - 0.5n| \ge t\sqrt{n}] \le 2\exp(-2t^2).$$

Now we have

$$\begin{aligned}H(y \mid x, c^*) &\ge 1 - C_2(1-\eta)^2 - \int_1^\infty \left(C_2(1-\eta)^2 t^2\right)' \cdot 2\exp(-2t^2)\,\mathrm{d}t \\ &\ge 1 - C_2(1-\eta)^2 - C_3(1-\eta)^2 \\ &= 1 - O\left((1-\eta)^2\right).\end{aligned}$$

So

$$H(x, y\,;\, c^*) \le 1 - H(y \mid x, c^*) = O\left((1-\eta)^2\right),$$

which concludes the proof in the Bernoulli case.

Now consider the case where $\mathcal{D}$ is a standard $n$-dimensional Guassian distribution. Instead of the number of 1's and $-1$'s, we consider the sum of coordinates of $x_0$. Let $|x| = \sum_i x_i$. Standard concentration bound implies that for any $t > 0$,

$$\Pr[||x| - 0.5n| \ge t\sqrt{n}] \le 2\exp(-2t^2).$$

Also, for $x_0$ where $|x_0| = t\sqrt{n}$ where w.l.o.g. $t > 0$, we still have

$$\frac{1}{2} - C_4(1-\eta)t \le \Pr[c \cdot x_0 \le 0] \le \frac{1}{2},$$

where $C_4 > 0$ is a constant. This is because $c \cdot x_0$ is the sum of $n$ scaled Bernoulli variables where the sum of the scaling factors is $t$. This sum is more concentrated when the scaling factors are uniform, which is exactly what happens in the Bernoulli case. As a consequence, the probability $\Pr[c \cdot x_0 \le 0]$ is closer to $\frac{1}{2}$.[12] The rest of the argument is totally similar to the Bernoulli case.

[12] The same fact can also be seen by applying the Lyapunov CLT for heterogeneous random variables.