On the efficiency of data collection for multiple Naïve Bayes classifiers

Edoardo Manino[a], Long Tran-Thanh[a], Nicholas R. Jennings[b]

[a] *University of Southampton, University Road, Southampton SO17 1BJ*
[b] *Imperial College, Kensington, London SW7 2AZ*

Abstract

Many classification problems are solved by aggregating the output of a group of distinct predictors. In this respect, a popular choice is to assume independence and employ a Naïve Bayes classifier. When we have not just one but multiple classification problems at the same time, the question of how to assign the limited pool of available predictors to the individual classification problems arises. Empirical studies show that the policies we use to perform such assignment have a strong impact on the accuracy of the system. However, to date there is little theoretical understanding of this phenomenon. To help rectify this, in this paper we provide the first theoretical explanation of the accuracy gap between the most popular policies: the non-adaptive uniform allocation, and the adaptive allocation schemes based on uncertainty sampling and information gain maximisation. To do so, we propose a novel representation of the data collection process in terms of random walks. Then, we use this tool to derive new lower and upper bounds on the accuracy of the policies. These bounds reveal that the tradeoff between the number of available predictors and the accuracy has a different exponential rate depending on the policy used. By comparing them, we are able to quantify the advantage that the two adaptive policies have over the non-adaptive one for the first time, and prove that the probability of error of the former decays at more than double the exponential rate of the latter. Furthermore, we show in our analysis that this result holds both in the case where we know the accuracy of each individual predictor, and in the case where we only have access to a noisy estimate of it.

Keywords: classification, Naïve Bayes, crowdsourcing, active learning

1. Introduction

Classification is a common machine learning problem where the objective is to assign single or multiple objects to their correct classes [3]. Many real-world classification problems are solved by aggregating the output of a group of predictors, each one of them proposing a class they think the objects belong to. A classic example is that of a patient consulting several physicians before deciding to go under invasive surgery [6]. Another is that of a police commander, dispatching patrols on several locations to figure out whether some illegal activities are being carried out [20]. A further example is that of crowdsourcing, where a large number of unqualified workers are asked to solve some simple task that machines are not able to tackle yet [22].

All of these examples have a number of properties in common. First, the predictors are potentially inaccurate: physicians may have different level of expertise on a specific illness, patrols may fail to spot important pieces of evidence, crowdworkers may try and execute the tasks as fast as possible regardless of accuracy. Second, all these settings present not just one but multiple classification problems at the same time: either a medical centre needs to assess the health of a number of patients, or a police commander needs to monitor several different locations, or a crowdsourcing project needs to find the solution to a whole set of tasks. Third, the available resources are limited: physicians can only see a limited number of patients per day, patrols cannot visit all the locations during their shift, and crowdworkers will execute only a small number of tasks before moving on to something else. Finally, the predictors can be assumed to be independent: physicians do not know the opinions of their colleagues before visiting a patient, and neither do patrols and crowdworkers in a well-designed system.

In these settings with limited resources, it is important to aggregate the predictors' outputs as efficiently as possible. Solving this problem has a long tradition of scientific enquiry, which dates back to Marquis de Condorcet [17] and his analysis of the majority voting aggregation rule. In more modern times, researchers have explored how to reduce the number of classification errors by weighting the predictors according to their accuracy [18]. When the weights are proportional to the log-odds of the predictors' accuracy, this method is also known as a Naïve Bayes classifier [3], which is the main focus of this paper. When the predictors' accuracy is unknown, one of several methods can be used to estimate it, either by reinforcement learning [13] or by inferring it from the predictors' output itself [14, 9, 4].

While these works provide algorithms to solve the classification problem, they do not study the relationship between the number of available predictors and the probability of a classification error. Knowing the rate at which additional predictors will lower the probability of error can be vital in many practical applications, as it allows us to predict the accuracy of the system given the available resources. Recently, both Berend and Kontorovich [2] and Gao et al. [7] have independently addressed this issue. However, as we show in this paper, their results can be further refined to get tighter estimates of this error-resources tradeoff.

At the same time, the aforementioned research focuses solely on the aggregation part of the classification problem, which happens after all the data has been collected from the predictors. Since in our setting we have multiple classification problems, there is an opportunity to improve how the data is collected. More specifically, if we allow the predictors to report their predictions sequentially, we can strategise on how many additional data points to collect for each classification task in the future, given the current information. For example, a medical centre may notice that the health of some patients has been difficult to assess and schedule them for more visits, or a police commander may receive contradictory reports for one location and send more patrols there, or the crowdsourcing platform may check which tasks have not been clearly classified yet and decide to allocate more workers to them. This sort of *adaptive* collection policy has been proposed many times in the literature. In particular, Barowy et al. [1] collect more data on the tasks where a clear majority has not formed yet, Welinder and Perona [24] retrain their probabilistic aggregator after every new data point and collect more data on the tasks with larger uncertainty, and Simpson and Roberts [21] always choose the data point that brings the maximum expected information gain.

However, the advantage of employing such adaptive policies has been shown only empirically so far [24, 1, 21]. Indeed, a prominent theoretical result by Karger et al. [9] states that any collection policy exhibits the same asymptotic tradeoff between the probability of a classification error and the average number of predictors per task R in the form $\mathbb{P}(error) \leq \exp(-cR)$. This means that we expect the error rate of any policy to decrease exponentially as we observe the output of more and more predictors. Crucially, this includes the simple *non-adaptive* policy of employing the same number R of predictors for each task. Despite numerous attempts to prove that adaptive policies are superior nonetheless, the current literature only covers specific scenarios, but fails to address the general case. In particular, Chen et al. [5] study the case where we have full control over the predictors, and thus we can select a small subset of them and let them provide most of the data points. In contrast, both Ho et al. [8] and Khetan and Oh [10] study the case where the productivity of each predictor is not under our control, but manage to show a difference between adaptive and non-adaptive policies only when the classification problems have varying degrees of difficulty.

Against this background, in this paper we provide the first theoretical explanation of the impact of data collection on the accuracy of a multiple classifier system. We do so by analysing the error-resources tradeoff of the most popular collection policies for multiple Naïve Bayes classifiers: the non-adaptive uniform allocation, and the adaptive schemes based on uncertainty sampling and information gain maximisation. More specifically, we make the following contributions to the state of the art. First, we improve on the results in Berend and Kontorovich [2] and Gao et al. [7] and provide tighter bounds on the error-resource tradeoff of the uniform allocation policy. Second, we propose a new way to represent the runtime behaviour of an adaptive policy in terms of a random walk in the log-odds domain. In so doing, we are able to derive the first bounds on the error-resource tradeoff of the uncertainty sampling policy. Third, we study the behaviour of uncertainty sampling in the presence of *collisions*, i.e. when the policy cannot execute its best action because it would mean collecting the same data point multiple times, and show that in most cases its performance does not change significantly. Fourth, we prove that the information gain maximisation policy is equivalent to uncertainty sampling in this setting. Finally, we use the aforementioned results to quantify the gap between adaptive and non-adaptive policies, and prove that the probability of error of the former decays at more than double the exponential rate of the latter.

Since the performance of data collection policies is tightly linked to the aggregation method in place, we repeat our analysis twice to cover the following common scenarios. In the first, the accuracy of each single predictor is known, and thus can be used to weigh the influence of their output on the final classification. In this case, the classification problem reduces to computing the output of a set of independent Naïve Bayes classifiers (more on this in Section 2). In the second, the accuracy is unknown and must be estimated by subjecting each predictor to a series of trials. We show how much the noise introduced by these estimates affects the results derived for the noiseless case.

This paper is structured in the following way. In Section 2 we introduce a formal model of our setting with all the relevant data aggregation methods and collection policies. In Section 3 we study the error-resources tradeoff of the collection policies when the accuracy of the predictors is known. In Section 4 we repeat our analysis when the accuracy of the predictors is unknown. In Section 5 we draw our conclusions and outline possible future work.

2. Preliminaries

A classic way of describing a setting with multiple classifiers and shared predictors is the celebrated Dawid-Skene model [6]. This has received considerable attention from both empirically-oriented and theoretical literature [24, 21, 7]. For simplicity, we restrict ourselves to the one-coin variant of the Dawid-Skene model. This assumes that all the classifiers are binary, which allows for a cleaner derivation of our theoretical results. The same approach has also been used in Liu et al. [14], Karger et al. [9], Bonald and Combes [4].

We present here the specific assumptions of the one-coin Dawid-Skene model, and the corresponding notation we use in this paper. First, the objective of the system is to solve M distinct classification problems, which we call *tasks* in the remainder of the paper. These tasks are binary in nature and we denote their underlying ground-truth vector as $\boldsymbol{y}$, where $y_i \in \{\pm 1\}$ is the true class of task i, with $i \in [1, M]$. Second, we assume the presence of a group of N predictors that provide us with a set of data points $X \equiv \{x_{ij}\}$, or *labels*, over the course of the data collection process. In this regard, we model the collection process in sequential fashion, where at each time t one and only one predictor j^t becomes available, gets assigned to a task i, and provides a value for the corresponding label $x_{ij^t} \in \{\pm 1\}$. We assume that the order of arrival of the predictor is random, and each predictor can in general become available multiple times. We use $X^t \equiv \{x_{ij}^t\}$ to refer to the set of labels observed up to time t. Third, we assume that there is a maximum *budget* $B << MN$ of labels or, in other terms, the data collection process ends at time $t = B$. Since the final number of labels $|X^{t=B}| = B$ is typically much smaller than the number of all possible task-predictor pairs, we assign by convention the value $x_{ij} = 0$ to the labels of any missing pair i, j. Fourth, we assume that the probability of observing a correct label depends only on the accuracy of the individual predictors, which we denote by $\mathbb{P}(x_{ij} = y_i) \equiv p_j$. In other words, we introduce here the Naïve Bayes property that the predictors are independent (conditioned on the ground-truth label of the task), and their accuracy is not affected by the task they are assigned to. Finally, we assume that the accuracy of the predictors is extracted from a common distribution $p_j \sim f_p$, and that the prior probability on the ground-truth labels $\mathbb{P}(y_i = +1) = 0.5$ is uniform (extending our results to other asymmetric priors is trivial as detailed in Section 3.1).

In this setting, the two challenges of collecting and aggregating the labels remain. In Sections 2.1 and 2.2 we report the methods proposed by the current literature and the few existing theoretical results that pertain to our work.

2.1. Aggregation methods

Given any partial set of of labels X^t available during the data collection process, we need a way to aggregate them into a vector $\hat{\boldsymbol{y}}^t$ of predictions over the classification of the tasks (where $\hat{y}_i^t$ is the prediction on task i given the information available at time t). Here we consider two of the most common aggregators, one that assumes perfect knowledge over the predictors' accuracy $\boldsymbol{p}$, and a second one that estimates this quantity instead.

2.1.1. Weighted Majority Voting

The simple *majority voting* rule predicts the class of a task by assigning $\hat{y}_i^t$ to the class with the majority of the labels. This aggregation method has the major drawback that predictors with low accuracy ($p_j \leq 0.5$) have the same influence as very accurate ones on the aggregated outcome. When the predictors' accuracy $\boldsymbol{p}$ is known, it is possible to assign larger weights to the accurate predictors and negative weights to those with $p_j < 0.5$, hence improving the reliability of the system. Furthermore, Nitzan and Paroush [18] show that there exists a set of optimal weights $\boldsymbol{w}$ that produces the best aggregated accuracy possible. This set of weights is defined as follows:

$$w_j \equiv \log\left(\frac{p_j}{1-p_j}\right) \tag{1}$$

and the corresponding *weighted majority voting* rule forms its predictions as follows:

$$\hat{y}_i^t \equiv \operatorname{sign}\left\{\sum_{j=1}^{N} x_{ij}^t w_j\right\} \tag{2}$$

Note that the weights w_j correspond to the log-odds that predictor j is correct. Similarly, the argument of Equation 2 is equal to the log-odds that the ground-truth label on task i is positive:

$$z_i^t \equiv \sum_{j=1}^{N} x_{ij}^t w_j = \log\left(\prod_{j=1}^{N}\left(\frac{p_j}{1-p_j}\right)^{x_{ij}^t}\right) = \log\left(\frac{\mathbb{P}(y_i = +1|X^t)}{\mathbb{P}(y_i = -1|X^t)}\right) \tag{3}$$

2.1.2. Bayesian Estimation

When the predictors' accuracy $\boldsymbol{p}$ is unknown, we can estimate it by subjecting each predictor j to Q trials and recording how many times c_j they report the correct answer. These trials can be thought of as part of the prior information we have about each predictor or, like in the case of crowdsourcing, be a set of *golden* tasks hidden in the normal workflow of the crowdworkers [19]. Given the result of these trials, the estimated predictors' accuracy can be computed in the following way. Assume that the real accuracy is distributed according to $p_j \sim \text{Beta}(\alpha, \beta)$. Then, the expected posterior accuracy given the number of successes c_j is the following:

$$\hat{p}_j \equiv \mathbb{E}_{p_j}\Big\{\mathbb{P}(p_j|c_j, Q, \alpha, \beta)\Big\} = \frac{c_j + \alpha}{Q + \alpha + \beta} \tag{4}$$

These estimates can be plugged into the weighted majority voting rule to form predictions over the classification of the tasks. In this regard, this aggregation rule is not only Bayes-optimal [2], but also keeps the property that the weighted sum $\hat{z}_i^t = \sum_{j=1}^N x_{ij}^t \hat{w}_j$, where $\hat{w}_j = \log(\hat{p}_j/(1-\hat{p}_j))$, is equal to the log-odds that the correct label on task i is positive. This property is key to the results we present in Section 4.3.

2.2. Collection policies

Aggregation methods aim at minimising the probability of a classification error given a set of labels X^t. However, they do not give us any clue on how these labels have to be collected. In fact, during the course of the data collection process, we need to decide which task i^t we want to allocate the available predictor j^t to. The sequence of decisions that ensue can be formalised in terms of a *collection policy*, a rule or heuristic that selects the task i^t given the current set of labels X^{t-1} and the incoming predictor j^t. The existing literature provides us with the following three popular collection policies.

2.2.1. Uniform Allocation (`UNI`)

This non-adaptive policy assigns the same number of predictors $R = B/M$ to each task (rounded to the nearest integer). In our examples, this would be like having every patient i visited by R physicians, every location i checked by R patrols, or every task i executed by R crowdworkers. In this respect, this policy can be completely defined at time $t = 0$, since observing the labels in X^t does not have any influence on the allocation. With this view, the existing results on the error-resources tradeoff of a single Naïve Bayes classifier in Berend and Kontorovich [2] and Gao et al. [7] are valid for this policy. The former work states that the probability of an error when the predictors' accuracy $\boldsymbol{p}$ is known is bounded by (see Theorem 1 therein):

$$\mathbb{P}(\hat{y}_i \neq y_i) \leq \exp\Bigg(-\frac{1}{2}\sum_{j:x_{ij}\neq 0}\Big(p_j - \frac{1}{2}\Big)w_j\Bigg) \tag{5}$$

Whereas the latter provides two different bounds under the same conditions:

$$\mathbb{P}(\hat{y}_i \neq y_i) \leq \prod_{j:x_{ij}\neq 0} 2\sqrt{p_j(1-p_j)} \tag{6}$$

and

$$\mathbb{P}(\hat{y}_i \neq y_i) \geq \frac{1}{4} \prod_{j:x_{ij}\neq 0} 2\sqrt{p_j(1-p_j)} \exp\Big(-\sqrt{\sum_{j=1}^{R} w_j^2}\Big) \tag{7}$$

We will improve on these results by providing tighter bounds in Section 3.2.

2.2.2. Uncertainty Sampling (US)

This adaptive policy maintains a measure of uncertainty over the current set of predictions $\hat{\boldsymbol{y}}^t$, and collects new labels over the most uncertain tasks. In our examples, this would be like estimating the uncertainty over the patients' health, the status of the locations, or the predicted classification of the crowdsourcing tasks, and then redirecting more physicians, patrols and crowdworkers to the most uncertain ones. Proposed first in the active learning community [12], this approach has been empirically shown to be successful in the multiple classifiers setting by Welinder and Perona [24] and Barowy et al. [1]. We propose the first theoretical analysis of its performance in Sections 3.3 and 4.3.

2.2.3. Information Gain Maximisation (IG)

This adaptive policy aims at maximising the amount of information carried by every new label x_{ij^t} [15]. It does so by choosing the task with the largest expected information gain, given the current posterior distribution on the predictions $\hat{\boldsymbol{y}}^t$. In our examples, this would be like choosing which patient to visit next based on the information we gain on their health, choosing which location to patrol based on the information we gain on their status, or choosing which task to label based on the information we gain on their classification. The empirical performance of this policy in a multiple classifiers setting is described by Simpson and Roberts [21]. We provide the first theoretical analysis of this policy in Sections 3.4 and 4.4, where we prove that in this setting its behaviour is identical to that of uncertainty sampling.

3. Known predictors' accuracy

In this section, we analyse the performance of the collection policies when weighted majority voting is used to aggregate the labels. Recall that this aggregation method requires perfect knowledge of the predictors' accuracy $\boldsymbol{p}$. In other terms, this is the scenario where we know the probability of a physician assessing their patients' health correctly, a patrol checking the state of their target precisely, or a crowdworker identifying the ground-truth class of a task successfully.

More specifically, our contribution is threefold. First, we compare and improve on the existing results for the UNI policy. Second, we introduce novel upper and lower bounds on the accuracy of the adaptive `US` policy, and show that it achieves better performance over the non-adaptive `UNI` one. Finally, we prove that in this scenario the two policies `US` and `IG` are equivalent. All these results are derived under a random-walk interpretation of the aggregation process, which is introduced next.

3.1. Modeling weighted majority voting as a random walk

The weighted majority voting algorithm can be interpreted as a random-walk in the log-odds domain as follows. Under the assumption that the prior on the true value of each task is symmetric, i.e. $\mathbb{P}(y_i = \pm 1) = 0.5$, the corresponding log-odds in the absence on any label will be $z_i = 0$ (extension to other priors is trivial[1]). From this starting point, the value of z_i will change in discrete steps every time a predictor casts its vote on task i. We denote the contribution of predictor j to task i as $s_{ij} \equiv x_{ij} w_j$. At the end of the collection process, the log-odds on each task i will reach their final values $z_i = \sum_j x_{ij} w_j$, the sign of which informs the final decision of the weighted majority aggregator.

Key to our analysis is understanding the distribution of the steps s_{ij}. This can be easily derived from the probability density function of the predictors' accuracy $p_j \sim f_p$. First, let us write the distribution of the weights $w_j \sim f_w$ as follows:

$$f_w(w_j) = \sigma(w_j)[1 - \sigma(w_j)] f_p(\sigma(w_j)) \tag{8}$$

where $\sigma(w_j) = p_j$ is the inverse of the log-odds function, and $\sigma(w_j)[1 - \sigma(w_j)]$ is its first-order derivative in w_j.

Second, assume by convention that the true class of task i is $z_i^* = +1$. Since a step of magnitude s_{ij} can be taken either by experts with weight $w_j = s_{ij}$ or $w_j = -s_{ij}$, we can write the probability density function of the steps as follows:

$$f_s(s_{ij}) = \sigma(s_{ij})[f_w(s_{ij}) + f_w(-s_{ij})] \tag{9}$$

where the term $\sigma(s_{ij})$ factors in the probability of observing a step with that specific sign and magnitude.

It must be noted that there is a connection between Equation 9 and the concept of *committee potential* proposed by Berend and Kontorovich [2]. In that paper, the authors define the quantity $\Phi = \sum_{j=1}^{R}(p_j - \frac{1}{2})w_j$ as the classification potential of a subset of R predictors, and use it to bound the accuracy of their aggregated votes. This relates to f_s in Equation 9 in that $\mathbb{E}_{f_s}\{s_{ij}\} = \mathbb{E}_{p_j}\{(2p_j - 1)w_j\}$ is the average potential of the whole population of predictors. We show in the following discussion how taking the whole population into account allows us to derive tighter bounds on the accuracy of the aggregated votes.

In general, we have $\mathbb{E}_{f_s}\{s_{ij}\} \geq 0$ for any distribution of the predictors' accuracy f_p. Equality holds only if all the predictors have accuracy $p_j = 0.5$ or, in other terms, when the whole population provides only random answers. This

[1] In the general case, where the prior is $\mathbb{P}(y_i = +1) \equiv q \in (0, 1)$, the corresponding log-odds in the absence of any labels are $z_i = \log(q/(1-q))$.

is due to the fact that predictors with $p_j < 0.5$ get assigned negative weights, so that even their labels move the log-odds in the correct direction on average. As a consequence, the random walk on z_i will exhibit a drift towards the true class y_i in most cases. However, we show in the following sections how different policies capitalise on this drift at different rates.

3.2. Performance of the `UNI` *policy*

When the `UNI` policy is used to allocate the predictors, every task i receives $R = B/M$ labels. We assume here that the policy does not introduce any bias in the allocation, so that each worker is equally likely to label any task. This is equivalent to the assumption in Karger et al. [9] of a random worker-task graph with constant edge cardinality R. Under this condition, we can focus our attention on the classification performance of one single task, as it is indistinguishable from the others in expectation. Furthermore, we can drop the index t from our notation, and study the probability of error when all the labels $X^{t=B}$ have been already collected.

As is common, let us set the convention that the true class is positive, i.e. $y_i = +1$. We are interested in computing the probability of a classification error after observing R labels. Recall that an error happens when the weighted majority ends up on the wrong class, thus:

$$\mathbb{P}(\hat{y}_i \neq y_i) = \mathbb{P}(z_i < 0) + \frac{1}{2}\mathbb{P}(z_i = 0) \tag{10}$$

where the second term represents the fact that the aggregator takes a random guess when the final log-odds are zero.

Since the steps s_{ij} in the random walk are independent, the probability in Equation 10 can be computed by convolution and integration as follows:

$$\mathbb{P}(\hat{y}_i \neq y_i) = \int_{-\infty}^{0-} F_s^R(z_i) dz_i + \frac{1}{2} F_s^R(0) \tag{11}$$

where $F_s^R = \mathop{\ast}_{k=1}^{R} f_s$ is the probability density function f_s convolved with itself R times.

While Equation 11 provides us with the exact performance of the `UNI` policy, it suffers from three drawbacks. First, we need full access to the probability density function f_s in order to compute it. Second, in general no closed form of it exists, and thus we need to approximate it by numerical methods. Finally, it is not clear from Equation 11 what is the impact of the number of labels R on the probability of error, which makes it difficult to compare it with the performance of the other policies.

For these reasons, we exploit the techniques proposed by Gao et al. [7] and compute the following upper bound on the probability of error. Note that our bound is tighter by a factor of 2 with respect to a simple adaptation of their result (see Equation 6).

Theorem 1. *Given a subset R of predictors with known accuracies $p_1, \ldots, p_N$, the probability of an error is upper bounded by:*

$$\mathbb{P}(\hat{y}_i \neq y_i) \leq \frac{1}{2}\Bigg[\prod_{j:x_{ij}\neq 0} 2\sqrt{p_j(1-p_j)}\Bigg] \tag{12}$$

Proof. Define the halved log-odds on task i as $h_i \equiv \sum_{j:x_{ij}\neq 0} x_{ij}\frac{1}{2}w_j$. It is clear that the probability of an error given $\boldsymbol{p}$ is:

$$\mathbb{P}(\hat{y}_i \neq y_i) = \sum_{\boldsymbol{x}_i}\left(\frac{1}{2}\mathbb{I}\{h_i = 0\} + \mathbb{I}\{h_i < 0\}\right)\mathbb{P}(\boldsymbol{x}_i) \tag{13}$$

where $\boldsymbol{x}_i$ is the vector of labels on task i. Now, the probability of observing $\boldsymbol{x}_i$ can be rewritten as follows:

$$\begin{aligned}\mathbb{P}(\boldsymbol{x}_i) &= \prod_{j:x_{ij}\neq 0} \frac{\mathbb{P}(x_{ij})\exp(-x_{ij}\frac{1}{2}w_j)}{\sqrt{p_j(1-p_j)}}\frac{\sqrt{p_j(1-p_j)}}{\exp(-x_{ij}\frac{1}{2}w_j)} \\ &= \exp(h_i)\prod_{j:x_{ij}\neq 0}\sqrt{p_j(1-p_j)}\end{aligned} \tag{14}$$

since $\exp(-x_{ij}\frac{1}{2}w_j) \in \{\sqrt{(1-p_j)/p_j}, \sqrt{p_j/(1-p_j)}\}$ depending on the sign of $x_{ij} \in \{\pm 1\}$. As a consequence, Equation 13 can be upper bounded as follows:

$$\mathbb{P}(\hat{y}_i \neq y_i) \leq \sum_{\boldsymbol{x}_i}\left(\frac{1}{2}\mathbb{I}\{h_i = 0\} + \mathbb{I}\{h_i < 0\}\right)\prod_{j:x_{ij}\neq 0}\sqrt{p_j(1-p_j)} \tag{15}$$

since $\exp(h_i) \leq 1$ for all $h_i \leq 0$. However, notice that for any set of weights $\boldsymbol{w}$, labels $\boldsymbol{x}_i$ and corresponding h_i, there is an opposite vector of labels $\boldsymbol{x}'_i = -\boldsymbol{x}_i$ which induces a value of the halved log-odds equal to $h'_i = -h_i$. As a result, the sum in Equation 15 totals 2^{R-1} which, after some trivial manipulations, gives the result in Theorem 1. □

The result in Theorem 1 can be extended to our case where the $\boldsymbol{p} = (p_1, \ldots, p_N)$ are independently extracted from a common distribution f_p. By taking the expected value over all possible $\boldsymbol{p}$ we get:

$$\mathbb{E}_{\boldsymbol{p}}\big\{\mathbb{P}(\hat{y}_i \neq y_i)\big\} \leq \frac{1}{2}\Big[2\mathbb{E}_{f_p}\Big\{\sqrt{p_j(1-p_j)}\Big\}\Big]^R \tag{16}$$

Equation 16 makes the dependence between the number of predictors R and the probability of an error clear. In this respect, notice that $2\sqrt{p_j(1-p_j)} \leq 1$, $\forall p_j$, with equality holding only for uninformative predictors $p_j = 0.5$. As a consequence, allocating more predictors to a task can only improve the overall classification accuracy. The exact rate at which this occurs is given by the expectation over $f_{\boldsymbol{p}}$. As an example, in case of beta-distributed predictors with $f_{\boldsymbol{p}} \sim \text{Beta}(\alpha, \beta)$, this expectation simplifies to the following:

$$\mathbb{E}_{\boldsymbol{p}}\{\mathbb{P}(\hat{y}_i \neq y_i)\} \leq \frac{1}{2}\left[2\frac{\text{B}(\alpha+\frac{1}{2},\beta+\frac{1}{2})}{\text{B}(\alpha,\beta)}\right]^R \tag{17}$$

At the same time, it is interesting to compare Theorem 1 with the corresponding bound proposed in Berend and Kontorovich [2] (see Equation 5), as this last result is usually overlooked by the crowdsourcing literature (see for example Gao et al. [7]). In this regard, we can show that our bound in Equation 12 yields tighter estimates:

Theorem 2. *Given a subset R of predictors with known accuracies $p_1, \ldots, p_N$, we have:*

$$\frac{1}{2}\left[\prod_{j:x_{ij}\neq 0} 2\sqrt{p_j(1-p_j)}\right] \leq \exp\left[-\frac{1}{2}\sum_{j:x_{ij}\neq 0}(p_j - \frac{1}{2})\log\left(\frac{p_j}{1-p_j}\right)\right] \tag{18}$$

where the first term is our result in Theorem 1, and the second is Theorem 1 in Berend and Kontorovich [2].

Proof. By taking the logarithm of both sides, we are left with the following:

$$-\log(2) + \frac{1}{2}\sum_{j:x_{ij}\neq 0}\log(4p_j(1-p_j)) \leq \frac{1}{2}\sum_{j:x_{ij}\neq 0}(\frac{1}{2} - p_j)\log\left(\frac{p_j}{1-p_j}\right) \tag{19}$$

which is obviously true, since for any p_j the left-side term $\log(4p_j(1-p_j))$ is never greater than the right-side term $(1/2 - p_j)\log(p_j/(1-p_j))$. □

On a different note, we can also put a *lower* bound on the probability of an error under the `UNI` policy. This gives us a precise understanding on the performance of this policy in the best-case scenario. Again, we use the techniques first proposed by Gao et al. [7], but improve on their bound (see Equation 7) by an exponential factor:

Theorem 3. *Given a subset R of predictors with known accuracies $p_1, \ldots, p_N$, the probability of an error is lower bounded by:*

$$\mathbb{P}(\hat{y}_i \neq y_i) \geq 0.36\left[\prod_{j:x_{ij}\neq 0} 2\sqrt{p_j(1-p_j)}\right]\exp\left(-\frac{1}{2}\sqrt{\sum_{j:x_{ij}\neq 0} w_j^2}\right) \tag{20}$$

Proof. By combining Equations 13 and 14, we can write the probability of an error as follows:

$$\begin{aligned}\mathbb{P}(\hat{y}_i \neq y_i) = &\left[\prod_{j:x_{ij}\neq 0} 2\sqrt{p_j(1-p_j)}\right] \\ &\sum_{\boldsymbol{x}_i}\left(\frac{1}{2}\mathbb{I}\{h_i = 0\} + \mathbb{I}\{h_i < 0\}\right)\exp(h_i)\left(\frac{1}{2}\right)^R\end{aligned} \tag{21}$$

Now, we can put a lower bound on the second term in Equation 21 for any $D \geq 0$:

$$
\begin{aligned}
\sum_{\boldsymbol{x}_i} \left(\frac{1}{2}\mathbb{I}\{h_i = 0\} + \mathbb{I}\{h_i < 0\} \right) \exp(h_i) \left(\frac{1}{2}\right)^R & \\
&\geq \sum_{\boldsymbol{x}_i} \left(\frac{1}{2}\mathbb{I}\{h_i = 0\} + \mathbb{I}\{h_i < 0\} - \mathbb{I}\{h_i > -D\} \right) \exp(-D) \left(\frac{1}{2}\right)^R \\
&= \left(\frac{1}{2} - \sum_{\boldsymbol{x}_i} \mathbb{I}\{h_i > -D\} \left(\frac{1}{2}\right)^R \right) \exp(-D)
\end{aligned}
\tag{22}
$$

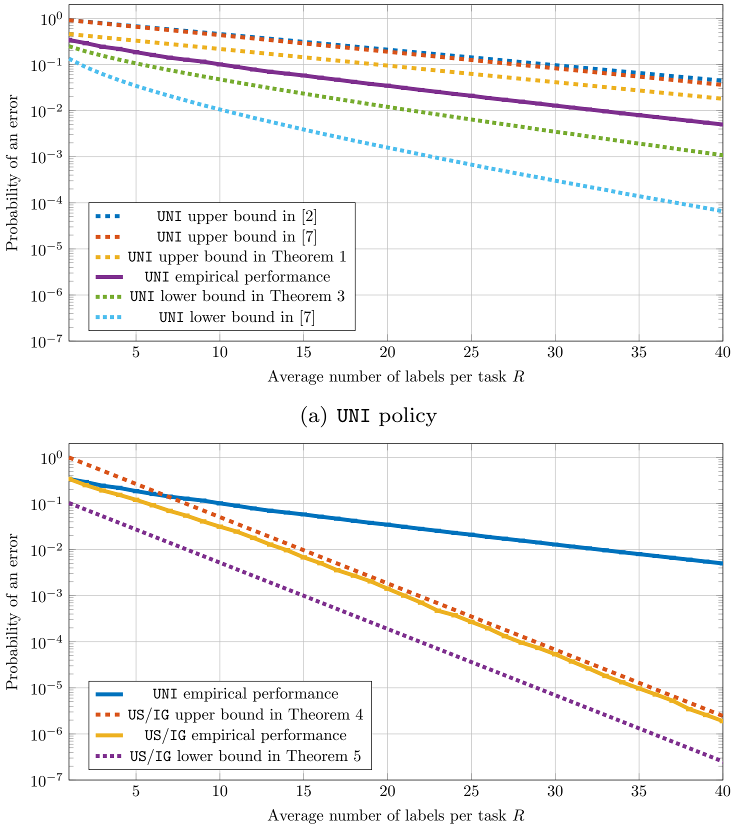
Then, we notice that the vector of labels $\boldsymbol{x}_i$ in Equation 22 can be interpreted as an observation of R independent Rademacher variables with probability $(1/2)^R$. Therefore, the sum over all $\boldsymbol{x}_i$ is equivalent to the probability of h_i exceeding $-D$, which can be bounded with Hoeffding's inequality:

$$
\mathbb{P}(h_i > -D) \leq \exp\left(- \frac{2D^2}{\sum_{j:x_{ij}\neq 0} \frac{1}{4} w_j^2} \right) \tag{23}
$$

Finally, by choosing $D = \frac{1}{2}\sqrt{\sum_{j:x_{ij}\neq 0} w_j^2}$, and substituting $0.5 - \exp(-2) \approx 0.36$ for simplicity, we get the result of the theorem. Note that, compared to the bound in Equation 7, this bound is tighter: not only is the constant factor outside the product 0.36 instead of 0.25, but also the magnitude of the argument of the exponential D is halved. □

Comparing the two results in Theorems 1 and 3, we can see that the upper and lower bound share the common term $\prod_j 2\sqrt{p_j(1-p_j)}$. At the same time, the two bounds may seem not to match at first glance because of the additional exponential term in Equation 20. However, this term depends on the square root of the variance of the weights, which means that its asymptotical contribution is in the form $\exp(-c_1\sqrt{R})$. On the contrary, the main term has a contribution in the form $\exp(-c_2 R)$ which quickly becomes dominant as R increases. As a result, we can conclude that the bounds in Theorem 1 and 3 do match asymptotically.

A visualisation of the results presented in this section is shown in Figure 1(a). There, we compare the empirical performance of the `UNI` policy with the respective theoretical bounds for increasing values of R. The empirical performance was evaluated on synthetic data with predictors distributed according to $f_p = \mathrm{Beta}(4,3)$. This represents a scenario with a fairly varied population of predictors whose average accuracy is not too distant from 0.5 (similar results can be obtained for different choices of f_p). As expected, the probability of an error decreases as the number of predictors per task R increases. Notice how the bounds in Theorem 1 and 3 closely match the error-resources tradeoff of the `UNI` policy, whereas the previous results happen to be slacker. In Section 3.5 we



(a) `UNI` policy

(b) `US` and `IG` policies

Figure 1: Comparison between the theoretical bounds and the empirical performance of the `UNI`, `US` and `IG` policies under the weighted majority voting aggregator, with $p_j \sim \text{Beta}(4,3)$.

will take these two bounds into consideration again, as we use them to compare the performance of the `UNI` policy with the other ones.

3.3. Performance of the `US` policy

We now move to our analysis of the performance of the `US` policy. This policy is adaptive, in the sense that its decisions on which task i^t to label next depend on the information collected so far. More specifically, this policy always chooses the task with the largest uncertainty, i.e. the one whose log-odds are closest to zero:

$$i^t = \underset{i:x_{ij^t}=0}{\text{argmin}}\{|z_i^t|\} \tag{24}$$

However, note that the `US` policy might not be able to perform its intended action at time t, if the available predictor j^t has already labelled the most uncertain task. In this case, Equation 24 will redirect the predictor on the next unlabelled task, in what we call a *collision* event. Since these events can reduce the performance of the `US` policy, but make our theoretical analysis more complicated, we study their impact separately in Section 3.3.1.

For now, let us assume that the `US` policy is always able to select the most uncertain task. Due to this specific behaviour, the magnitude of the log-odds $|z_i^t|$ tends to increase at roughly the same pace on all M tasks during the collection process. This phenomenon allows us to prove the following two bounds:

Theorem 4. *The probability of a classification error under the* `US` *policy is upper bounded by:*

$$\mathbb{P}(\hat{y}_i \neq y_i) \leq \exp\big(-\mathbb{E}_{f_s}\{s_{ij}\}(R-1)\big) \tag{25}$$

Proof. From the perspective of a single task i, the `US` policy operates in short bursts of activity, as i keeps receiving new labels until it is no longer the most uncertain one. We define $z_B \equiv \max_t\{\min_i\{|z_i|\}\}$ as the threshold that all tasks have crossed at some point of the collection process. In this respect, we can model the evolution of the log-odds z_i^t as a bounded random walk, which starts in $z_i^{t=0} = 0$ and ends when z_i^t leaves the interval $(-z_B, +z_B)$.

Given this, let us assume that we can fix the threshold z_B and then collect as many labels as needed in order to cross it. We denote the log-odds after crossing the threshold as z_i^r, where $z_i^r \notin (-z_B, +z_B)$, and the log-odds at the step before as z_i^{r-1}. According to this definition, r is a stopping time since it is uniquely defined by the information collected before step r. Thus, we can use Wald's equation [23] to link the expected value of z_i^r and the stopping time r:

$$\mathbb{E}\{z_i^r\} = \mathbb{E}\{r\}\mathbb{E}_{f_s}\{s_{ij}\} \tag{26}$$

Recall, however, that z_i^r is the sum of r i.i.d random variables, and that $z_i^{r-1} \in (-z_B, +z_B)$ by definition. As a consequence, we can further bound the expected value of z_i^r by:

$$\mathbb{E}\{z_i^r\} = \mathbb{E}\{z_i^{r-1}\} + \mathbb{E}_{f_s}\{s_{ij}\} < z_B + \mathbb{E}_{f_s}\{s_{ij}\} \tag{27}$$

Putting Equations 26 and 27 together, we can derive a bound for the threshold z_B:

$$z_B > \mathbb{E}_{f_s}\{s_{ij}\}\big(\mathbb{E}\{r\} - 1\big) \tag{28}$$

At the same time, we also know that the random walks on the M tasks are independent, and that the variance of r for a bounded random walk with i.i.d. steps is finite. Therefore, as $M \to \infty$ the total number of steps required to cross the threshold will converge to its expected value, i.e. $B \to M\mathbb{E}\{r\}$. This property allows us to substitute $\mathbb{E}\{r\} = R$.

However, we also know that the confidence in our prediction is $\mathbb{P}(\hat{y}_i = y_i) = \sigma(|z_i|)$. Thus, we can bound the final accuracy as:

$$\mathbb{P}(\hat{y}_i = y_i) \geq \sigma\big(\mathbb{E}_{f_s}\{s_{ij}\}(R-1)\big) \tag{29}$$

which yields Equation 25 after some simple algebraic manipulations. □

Theorem 5. *The probability of a classification error under the* `US` *policy is lower bounded by:*

$$\mathbb{P}(\hat{y}_i \neq y_i) \geq \frac{1}{2} \exp\big(- \mathbb{E}_{f_s}\{s_{ij}\}R - \mathbb{E}_{f_s}\{|s_{ij}|\} - 0.56\big) \tag{30}$$

Proof. Similarly to the proof of Theorem 4, we are interested in the value of z_i at the end of the random walk:

$$\mathbb{E}\{|z_i^r|\} = \mathbb{E}\{z_i^r\} - 2\int_{-\infty}^{-z_B} z\mathbb{P}(z_i^r = z)dz \tag{31}$$

where

$$\begin{aligned}\int_{-\infty}^{-z_B} z\mathbb{P}(z_i^r{=}z)dz &= -z_B\mathbb{P}(z_i^r{\leq}{-z_B}) + \int_{-\infty}^{0} u\mathbb{P}(z_i^r{=}u{-}z_B)du \\ &\geq -z_B\sigma(-z_B) - \frac{1}{2}\mathbb{E}_{f_s}\{|s_{ij}|\}\end{aligned} \tag{32}$$

since $\mathbb{P}(z_i^r \leq -z_B)$ is the probability of an error, which is never larger than $\sigma(-|z_B|)$, and $\mathbb{P}(z_i^r = u - z_B)$ is the probability of undershooting the threshold by u, which is never larger than the probability of overshooting it in expectation. Now, we can compute $\max_{z\geq 0}\{z\sigma(-z)\} \approx 0.28$. Additionally, we know that $\mathbb{E}\{z_i^r\} = R\mathbb{E}_{f_s}\{s_{ij}\}$. Thus, the following is true:

$$\mathbb{E}\{|z_i^r|\} \leq R\mathbb{E}_{f_s}\{s_{ij}\} + \mathbb{E}_{f_s}\{|s_{ij}|\} + 0.56 \tag{33}$$

By noting that $\mathbb{P}(\hat{y}_i \neq y_i) = \mathbb{E}\{\sigma(-|z_i|)\} \geq \sigma(-\mathbb{E}\{|z_i|\}) \geq \frac{1}{2}\exp(-\mathbb{E}\{|z_i|\})$ we obtain the result of the theorem. □

Notice that the bounds in Equations 25 and 30 match asymptotically and guarantee that the `US` policy has an exponential tradeoff with constant $c = \mathbb{E}_{f_s}\{s_{ij}\}$. This means that the policy fully exploits the drift in the random walk over the log-odds z_i to reduce the probability of an error.

A visualisation of the results presented in this section is shown in Figure 1(b). There, we compare the empirical performance of the `US` policy with the theoretical bounds derived in Equations 25 and 30. The empirical performance was evaluated with the same settings as Figure 1(a) (see the end of Section 3.2). Furthermore, we set the number of tasks to $M = 1000$ and a single label per predictor in order to avoid collisions (see Section 3.3.1 for more details). Notice how the upper bound closely matches the error-resources tradeoff of the `US` policy, and how the performance of this policy is increasingly better than that of the `UNI` policy as the average number of predictors per task R becomes larger. This performance gap has already been discovered empirically by Welinder and Perona [24] in a similar context. In Section 3.5 we study it from the theoretical perspective by using the results in Equations 25 and 30.

3.3.1. Impact of collisions

The results in Theorems 4 and 5 are based on the assumption that the `US` policy is always able to collect a new label on the most uncertain task $i^* = \operatorname{argmin}_i\{|z_i^t|\}$. In some cases this cannot be guaranteed, as the predictors can provide multiple labels, and thus the available predictor j^t might have already labelled task i^* in one of the previous time steps $t' \in [0, t-1]$. For example, in the medical scenario a physician visits multiple patients each day. Similarly, police patrols can check multiple locations during a single shift, and crowdworkers typically execute a number of tasks before logging out of the crowdsourcing platform.

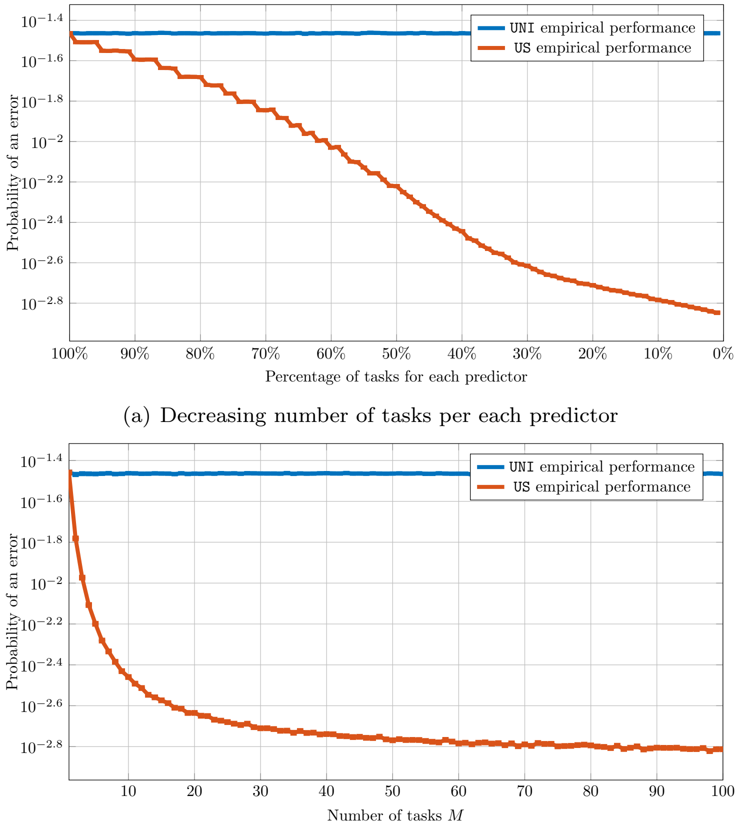
When this happens, the `US` policy is forced to choose the next best task according to Equation 24. As the number of these collisions grow, the collection process becomes less efficient. In the worst-case scenario, each predictor j labels all the M tasks, and thus the `US` policy becomes equivalent to the `UNI` policy (exactly $R = N$ predictors are allocated to each task i). In contrast, in the best-case scenario each predictor j labels only one task and thus no collision can occur. In order to study what happens in between these two extremes, we run the following experiments.

First, we let the number of labels per predictor vary from $Q = M$ (all the tasks) to $Q = 1$ (only one task). The results shown in Figure 2(a) were evaluated for predictors with accuracy distributed according to $f_p = \text{Beta}(4, 3)$ and with $R = 20$ labels per task (see Figure 1 for comparison). Note that similar results can be obtained for different choices of these parameters. As the figure shows, even with predictors that label up to 40% of the tasks, the `US` policy retains its ability to deliver an order of magnitude lower probability of error than the `UNI` policy. The reason behind this is that, even if a collision occurs at time t, the `US` policy is able to assign some other predictors to the most uncertain task i^* at a later stage. In broader terms, as long as there is enough scope to strategise during the collection process, the `US` policy is able to compensate for the presence of collisions.

Second, recall that the results in Theorems 4 and 5 are derived under the assumption that the number of tasks is large ($M \to \infty$). In this respect, we want to know how many tasks the `US` policy needs to reach its full potential. To this end, we fix $Q = 1$ and let the number of tasks increase from $M = 1$ to $M = 100$ (the other parameters are the same as Figure 2(a)). Note that for $M = 1$ the `US` and `UNI` policies are equivalent, as both allocate $N = R$ predictors on that single classification task. The results in Figure 2(b) show that it takes no more than $M = 10$ tasks to give the `US` policy an order of magnitude of advantage over the non-adaptive `UNI` policy. Furthermore, it takes as little as $M = 50$ tasks for the performance of the `US` policy to converge to the asymptotical performance with $M \to \infty$.

3.4. Performance of the `IG` policy

The `IG` policy always selects the task that yields the largest information gain. The information gain of a new label x_{ij^t} is measured in terms of the Kullback-Leibler divergence [11] between the current posterior distribution on the task



(a) Decreasing number of tasks per each predictor

(b) Increasing number of classification tasks

Figure 2: Impact of collisions on the performance of the `UNI` and `US` policies. Being non-adaptive, the `UNI` policy is not affected by them and its performance stays constant.

classification, and the one updated with the additional data point x_{ij^t}. Formally, we can write the information gain as follows:

$$\begin{aligned}\mathcal{I}\big(X^{t-1} \cup x_{ij^t} \big|\big| X^{t-1}\big) \\ = \sum_{y\in\{\pm 1\}} \mathbb{P}(y_i = y|X^{t-1} \cup x_{ij^t}) \log\left(\frac{\mathbb{P}(y_i = y|X^{t-1} \cup x_{ij^t})}{\mathbb{P}(y_i = y|X^{t-1}})\right)\end{aligned} \tag{34}$$

At the same time, we do not have access to the value of the next label x_{ij^t} in advance. Thus, the `IG` policy is forced to select the task i^t that yields the largest value of the information gain in expectation:

$$i^t = \operatorname*{argmax}_{i:x_{ij^t}=0} \Big\{\mathbb{E}_{x_{ij^t}}\big\{\mathcal{I}\big(X^{t-1} \cup x_{ij^t} \big|\big| X^{t-1}\big)\big\}\Big\} \tag{35}$$

Given this definition, we can prove that the policies `US` and `IG` are in fact equivalent:

Theorem 6. *Given the current set of labels X^{t-1} and a predictor with weight $w_{j^t} \neq 0$, the two policies* `US` *and* `IG` *select the same task i^t, except in the case of a tie.*

Proof. Let us write the information gain in closed form as follows:

$$\begin{aligned} &\mathcal{I}\big(X^{t-1} \cup x_{ij^t} \big|\big| X^{t-1}\big) \\ &\qquad = \sum_{y \in \{\pm 1\}} \sigma\big(y(z_i^{t-1} + x_{ij^t} w_{j^t})\big) \log \left(\frac{\sigma\big(y(z_i^{t-1} + x_{ij^t} w_{j^t})\big)}{\sigma[y z_i^{t-1}]} \right) \end{aligned} \tag{36}$$

since the log-odds vector $\boldsymbol{z}^t$ changes on task i only. The expected value of Equation 36 is thus the following:

$$\begin{aligned} &\mathbb{E}_{x_{ij^t}} \big\{ \mathcal{I}\big(X^{t-1} \cup x_{ij^t} \big|\big| X^{t-1}\big) \big\} \\ &\qquad = \sum_{x_{ij^t} \in \{\pm 1\}} \mathcal{I}\big(X^{t-1} \cup x_{ij^t} \big|\big| X^{t-1}\big) \mathbb{P}(x_{ij^t} | X^{t-1}, w_{j^t}) \end{aligned} \tag{37}$$

where

$$\mathbb{P}(x_{ij^t} | X^{t-1}, w_{j^t}) = \sigma(z_i^{t-1}) \sigma(x_{ij^t} w_{j^t}) + \sigma(-z_i^{t-1}) \sigma(-x_{ij^t} w_{j^t}) \tag{38}$$

since $\sigma(z_i^{t-1})$ is the current posterior probability on task i, and $\sigma(w_{j^t})$ is the probability that the new label x_{ij^t} is correct.

First, note that Equation 37 is even in z_i^{t-1} and w_{j^t} (this can be easily proven by substitution). Given this, we can focus on the positive intervals $z_i^{t-1}, w_{j^t} \in (0, \infty)$ and show that the function is monotonically decreasing. We can do so by taking the derivative of Equation 37 and proving that it is negative by contradiction:

$$\begin{aligned} &\frac{d}{d z_i^{t-1}} \Big\{ \mathbb{E}_{x_{ij^t}} \big\{ \mathcal{I}\big(X^{t-1} \cup x_{ij^t} \big|\big| X^{t-1}\big) \big\} \Big\} \\ &= \frac{\exp(z_i^{t-1})(\exp(w_{j^t}) - 1)}{(1 + \exp(z_i^{t-1}))^2 (1 + \exp(w_{j^t}))} \left[w_{j^t} + \log \left(\frac{1 + \exp(z_i^{t-1} - w_{j^t})}{1 + \exp(z_i^{t-1} + w_{j^t})} \right) \right] \geq 0 \end{aligned} \tag{39}$$

which yields:

$$\frac{1 + \exp(z_i^{t-1} - w_{j^t})}{1 + \exp(z_i^{t-1} + w_{j^t})} \geq \exp(-w_{j^t}) \tag{40}$$

which is false for any $z_i^{t-1}, w_{j^t} \in (0, \infty)$. Finally, we can see that both the objective function $|z_i^{t-1}|$ for the `US` policy in Equation 24 and the objective function for the `IG` policy in Equation 37 are even in z_i^{t-1}, monotonically increasing (decreasing) for $z_i^{t-1} \to \infty$ and have a minimum (maximum) in $z_i^{t-1} = 0$ for any $w_{j^t} \neq 0$. Hence, if a task i is preferred to i' under one policy, it is also preferred under the other. □

3.5. Policy comparison

Now that we have established bounds on the performance of the three policies `UNI`, `US` and `IG`, we can draw a meaningful comparison between them. On the one hand, all of them exhibit an asymptotic tradeoff between the number of labels R and the final accuracy in the form $\mathbb{P}(error) \leq \exp(-cR)$. On the other hand, the specific constant factor c varies from policy to policy. In order to see this more clearly, let us derive c for the three policies.

First, consider the probability of an error for the `UNI` policy in Equation 16. This can be rewritten in exponential form as follows:

$$\mathbb{E}_{\boldsymbol{p}}\big\{\mathbb{P}(\hat{y}_i \neq y_i)\big\} \leq \frac{1}{2}\exp\left[R\log\left(\mathbb{E}_{f_p}\left\{2\sqrt{p_j(1-p_j)}\right\}\right)\right] \tag{41}$$

which yields the following value for the constant c_{uni}:

$$c_{uni} = -\log\left(\mathbb{E}_{f_p}\left\{2\sqrt{p_j(1-p_j)}\right\}\right) \tag{42}$$

Similarly, we can define a shared constant c_{ada} for the two adaptive policies `US` and `IG` thanks to the results in Theorems 4 and 5:

$$c_{ada} = \mathbb{E}_{f_s}\{s_{ij}\} = \mathbb{E}_{f_p}\left\{(2p_j - 1)\log\left(\frac{p_j}{1-p_j}\right)\right\} \tag{43}$$

In general, a larger value of c means that the policy is asymptotically more efficient in using additional labels to improve the classification accuracy. In this respect, we can show that the policies `US` and `IG` are always superior to the `UNI` policy.

Theorem 7. *For any distribution f_p, we have $c_{ada} \geq 2c_{uni}$*

Proof. By Jensen's inequality we have:

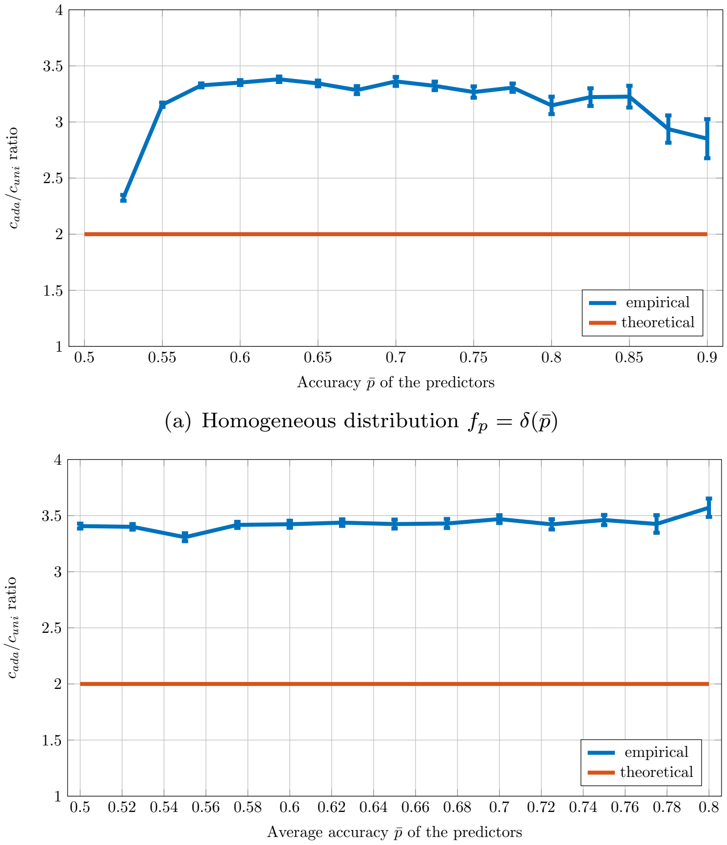
$$c_{uni} \leq \mathbb{E}_{p_j}\left\{-\frac{1}{2}\log\left(4p_j(1-p_j)\right)\right\} \tag{44}$$

Comparing the argument of Equations 43 and 44, we can notice that:

$$\forall p \in (0,1), \qquad (2p-1)\log\left(\frac{p}{1-p}\right) \geq -\log\left(4p(1-p)\right) \tag{45}$$

Then, the result in the theorem follows from the monotonicity property of the expected value. □

At the same time, the result in Theorem 7 does not explain how much the c_{ada}/c_{uni} ratio is larger than its theoretical baseline of 2. To study this, we run synthetic experiments with different populations of predictors f_p, and estimate the value of the c_{ada}/c_{uni} ratio empirically. In Figure 3(a) we report the results with $f_p = \delta(\bar{p})$, i.e. all the predictors have the same accuracy $p_j = \bar{p}$, for



(a) Homogeneous distribution $f_p = \delta(\bar{p})$

(b) Mixed distribution $f_p = \text{Uniform}(\bar{p} - 0.2, \bar{p} + 0.2)$

Figure 3: Comparison between the empirical c_{ada}/c_{uni} ratio and its theoretical baseline.

different values of $\bar{p}$. In Figure 3(b) we report the results with $f_p = \text{Uniform}(\bar{p} - 0.2, \bar{p} + 0.2)$, i.e. the predictors are uniformly distributed in an interval of width 0.4 around $\bar{p}$. This choice of parameters covers both homogeneous and mixed populations of predictors, and varying degrees of average accuracy.

As the figures show, the empirical value of the c_{ada}/c_{uni} ratio is around 3 in most cases. A notable exception is the homogeneous population $f_p = \delta(\bar{p})$ with $\bar{p}$ close to 0.5. In this scenario, the output of the predictors is almost random, and the advantage of the `US` and `IG` policies is reduced. In the extreme case of $p_j = 0.5$ for all predictors, the aggregator can only output random guesses no matter the collection policy used. In such circumstances we can expect the value of the c_{ada}/c_{uni} ratio to be equal to 1.

4. Estimated predictors' accuracy

In this section, we remove the assumption that the predictors' accuracy $\boldsymbol{p}$ is known, and study the case where we only have access to some noisy estimates $\hat{\boldsymbol{p}}$. On the one hand, these estimates can depend on the past performance of the predictors: for example, how many times a physician has correctly assessed their patients' health, how many times a patrol team has reported the correct status of a location back to base, or how many times a crowdworker has identified the correct class of a task in the past. On the other hand, these estimates can also depend on side information about the predictors: for instance, the medical background of each physician, the type of equipment available to each police patrol, or the reputation score each crowdworker has accumulated in the past interactions with the crowdsourcing platform. Specific assumptions about the nature of these estimates are introduced alongside our theoretical discussion when needed.

The contributions we present in this section are as follows. First, we propose novel upper and lower bounds on the three collection policies `UNI`, `US` and `IG`. In so doing, we provide closed formulas for the error-resources tradeoff when the predictors' accuracy $\hat{\boldsymbol{p}}$ is computed by Bayesian estimation (see Section 2.1). This has been identified as an important open problem by Berend and Kontorovich [2]. Second, we establish that the adaptive policies `US` and `IG` are equivalent in this scenario, and quantify their performance advantage over the non-adaptive `UNI` policy. For an easier comparison with the corresponding results in the known accuracy case, the structure of this section closely matches that of Section 3.

4.1. Random walk on the estimated log-odds

The presence of the estimates $\hat{\boldsymbol{p}}$ in the weighted majority aggregation rule introduces a number of changes in how the random walk on the (estimated) log-odds $\hat{z}_i^t$ evolves. Now, every step assumes the form $\hat{s}_{ij} \equiv x_{ij}\hat{w}_j$, where x_{ij} is still the observed label from predictor j on task i, but $\hat{w}_j = \log(\hat{p}_j/(1-\hat{p}_j))$ is its estimated weight. Coincidentally, the probability density function of these steps can be written as follows:

$$\begin{aligned} f_{\hat{s}}(\hat{s}_{ij}) &= \mathbb{P}(\hat{w}_j = \hat{s}_{ij}, x_{ij} = 1) + \mathbb{P}(\hat{w}_j = -\hat{s}_{ij}, x_{ij} = -1) \\ &= \int_0^1 \Big[p_j \mathbb{P}(\hat{w}_j = \hat{s}_{ij}|p_j) + (1-p_j)\mathbb{P}(\hat{w}_j = -\hat{s}_{ij}|p_j)\Big]\mathbb{P}(p_j)dp_j \end{aligned} \tag{46}$$

where $\mathbb{P}(\hat{w}_j|p_j)$ is the conditional probability of estimating the weight $\hat{w}_j$ given the real accuracy p_j. Note that in the case where $\hat{w}_j$ is computed by Bayesian estimation (see Section 2.1), such conditional probability becomes easy to compute as it depends solely on the number of correct answers c_j of each worker.

4.2. Performance of the `UNI` *policy*

As per the known accuracy case (see Equation 11), we have a direct way to estimate the probability of an error under the `UNI` policy:

$$\mathbb{P}(\hat{y}_i \neq y_i) = \int_{-\infty}^{0-} F_{\hat{s}}^{R}(\hat{z}_i) d\hat{z}_i + \frac{1}{2} F_{\hat{s}}^{R}(0) \tag{47}$$

where $F_{\hat{s}}^{R} = \mathop{\ast}_{k=1}^{R} f_{\hat{s}}$ is the probability density function of the estimated steps $\hat{s}_{ij}$ convolved with itself R times. In a similar way to the known accuracy case, this equation has no closed form in general, which makes it difficult to compare it with the corresponding results for the other policies. In this regard, we can derive more useful results by bounding the probability of an error from above:

Theorem 8. *Given a subset of predictors with accuracies $p_1, \ldots, p_N$ and estimates $\hat{p}_1, \ldots, \hat{p}_N$, the probability of an error is upper bounded by:*

$$\mathbb{P}(\hat{y}_i \neq y_i) \leq \prod_{j:x_{ij}\neq 0} \left(\frac{p_j}{\hat{p}_j} + \frac{1-p_j}{1-\hat{p}_j} \right) \sqrt{\hat{p}_j(1-\hat{p}_j)} \tag{48}$$

Proof. Define the estimated halved log-odds on task i as $\hat{h}_i \equiv \sum_{j:x_{ij}\neq 0} x_{ij} \frac{1}{2} \hat{w}_j$. Then, the probability of an error given $\boldsymbol{p}$ and $\hat{\boldsymbol{p}}$ is:

$$\mathbb{P}(\hat{y}_i \neq y_i) = \sum_{\boldsymbol{x}_i} \left(\frac{1}{2} \mathbb{I}\{\hat{h}_i = 0\} + \mathbb{I}\{\hat{h}_i < 0\} \right) \mathbb{P}(\boldsymbol{x}_i) \tag{49}$$

Now, we apply a procedure similar to the proof of Theorem 1, and rewrite the probability of $\boldsymbol{x}_i$ as follows:

$$\begin{aligned} \mathbb{P}(\boldsymbol{x}_i) &= \prod_{j:x_{ij}\neq 0} \frac{\mathbb{P}(x_{ij})}{g(x_{ij})} \frac{g(x_{ij}) \exp(-x_{ij} \frac{1}{2} \hat{w}_j)}{\sqrt{\hat{p}_j(1-\hat{p}_j)}} \frac{\sqrt{\hat{p}_j(1-\hat{p}_j)}}{\exp(-x_{ij} \frac{1}{2} \hat{w}_j)} \\ &= \exp(\hat{h}_i) \prod_{j:x_{ij}\neq 0} \frac{\mathbb{P}(x_{ij})}{g(x_{ij})} \sqrt{\hat{p}_j(1-\hat{p}_j)} \end{aligned} \tag{50}$$

where $g(+1) \equiv \hat{p}_j$ and $g(-1) \equiv 1 - \hat{p}_j$. Substituting Equation 50 into Equation 49, we can derive the following bound:

$$\mathbb{P}(\hat{y}_i \neq y_i^*) \leq \sum_{\boldsymbol{x}_i} 1 \prod_{j:x_{ij}\neq 0} \frac{\mathbb{P}(x_{ij})}{g(x_{ij})} \sqrt{\hat{p}_j(1-\hat{p}_j)} \tag{51}$$

which yields the result in the theorem after some trivial algebraic manipulations. □

Compared to the corresponding result with known accuracies $\boldsymbol{p}$ (see Equation 12), we have the additional term $\left(\frac{p_j}{\hat{p}_j} + \frac{1-p_j}{1-\hat{p}_j} \right)$ which accounts for the errors in the estimates $\hat{\boldsymbol{p}}$. When these estimates converge to the real values $\boldsymbol{p}$, the bound in Equation 12 is restored (up to a factor of 0.5). Note also that Theorem 8 is exponentially tighter than the corresponding result in Gao et al. [7] (see Theorem 4.1 and its proof therein), since our result does not depend on the estimates $\hat{\boldsymbol{p}}$ being arbitrarily close to the real values $\boldsymbol{p}$.

Equation 48 shows clearly the dependence between the number of predictors R and the probability of an error. When we take the expectation over the probability of observing a predictor with real accuracy p_j and estimate $\hat{p}_j$, we can form a prediction on the performance of the collection process. In the case of Bayesian estimation, this expectation has the closed form presented below. Note that the theorem that follows solves the open problem first identified by Berend and Kontorovich [2].

Theorem 9. *For a set of R predictors with real accuracy $p_j \sim Beta(\alpha, \beta)$ and estimated accuracy on Q trials $\hat{p}_j = \frac{\alpha+c_j}{\alpha+\beta+Q}$, we have*

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{p},\hat{\boldsymbol{p}}}\{\mathbb{P}(\hat{y}_i \neq y_i)\}\\ &\qquad \leq \left[2\sum_{c_j=0}^{Q}\binom{Q}{c_j}\frac{B(c_j+\alpha, Q-c_j+\beta)}{B(\alpha,\beta)}\left(\frac{\sqrt{(c_j+\alpha)(Q-c_j+\beta)}}{Q+\alpha+\beta}\right)\right]^R\end{aligned} \tag{52}$$

Proof. Taking the expectation of Equation 48 on p_j and $\hat{p}_j$ yields the following:

$$\mathbb{E}_{\boldsymbol{p},\hat{\boldsymbol{p}}}\{\mathbb{P}(\hat{y}_i \neq y_i)\} \leq \prod_{j=1}^{R}\mathbb{E}_{f_p,f_{\hat{p}}}\left\{p_j\sqrt{\frac{1-\hat{p}_j}{\hat{p}_j}}+(1-p_j)\sqrt{\frac{\hat{p}_j}{1-\hat{p}_j}}\right\} \tag{53}$$

Now, thanks to the assumption of a beta-distributed p_j and Bayesian estimation for $\hat{p}_j$, we can compute the expectation of Equation 53 in closed form. For the two addends therein we have respectively:

$$\mathbb{E}_{f_p,f_{\hat{p}}}\left\{p_j\sqrt{\frac{1-\hat{p}_j}{\hat{p}_j}}\right\} = \sum_{c_j=0}^{Q}\binom{Q}{c_j}\frac{\mathrm{B}(c_j+\alpha+1, Q-c_j+\beta)}{\mathrm{B}(\alpha,\beta)}\sqrt{\frac{Q-c_j+\beta}{c_j+\alpha}} \tag{54}$$

$$\mathbb{E}_{f_p,f_{\hat{p}}}\left\{(1-p_j)\sqrt{\frac{\hat{p}_j}{1-\hat{p}_j}}\right\} = \sum_{c_j=0}^{Q}\binom{Q}{c_j}\frac{\mathrm{B}(c_j+\alpha, Q-c_j+\beta+1)}{\mathrm{B}(\alpha,\beta)}\sqrt{\frac{c_j+\alpha}{Q-c_j+\beta}} \tag{55}$$

Finally, by considering that $B(d+1,e) = B(d,e)\frac{d}{d+e}$ and $B(d,e+1) = B(d,e)\frac{e}{d+e}$, we can merge Equations 54 and 55 into the result in the theorem. □

On a different note, we can also bound the probability of an error under the `UNI` policy from below:

Theorem 10. *Given a subset of predictors with accuracies $p_1, \dots, p_N$ and estimates $\hat{p}_1, \dots, \hat{p}_N$, the probability of an error is lower bounded by:*

$$\mathbb{P}(\hat{y}_i \neq y_i) \geq 0.36 \Bigg[\prod_{j:x_{ij} \neq 0} \min\left(\frac{p_j}{\hat{p}_j}, \frac{1-p_j}{1-\hat{p}_j} \right) 2\sqrt{\hat{p}_j(1-\hat{p}_j)} \Bigg] \exp\left(-\frac{1}{2} \sqrt{\sum_{j=1}^{R} \hat{w}_j^2} \right) \tag{56}$$

Proof. Combining Equations 49 and 50, we can bound the probability of an error as follows:

$$\begin{aligned} \mathbb{P}(\hat{y}_i \neq y_i) \geq \sum_{\boldsymbol{x}_i} & \left(\frac{1}{2} \mathbb{I}\{\hat{h}_i = 0\} + \mathbb{I}\{\hat{h}_i < 0\} \right) \\ & \exp(\hat{h}_i) \prod_{j:x_{ij} \neq 0} \min_{x_{ij}} \left\{ \frac{\mathbb{P}(x_{ij})}{g(x_{ij})} \right\} \sqrt{\hat{p}_j(1-\hat{p}_j)} \end{aligned} \tag{57}$$

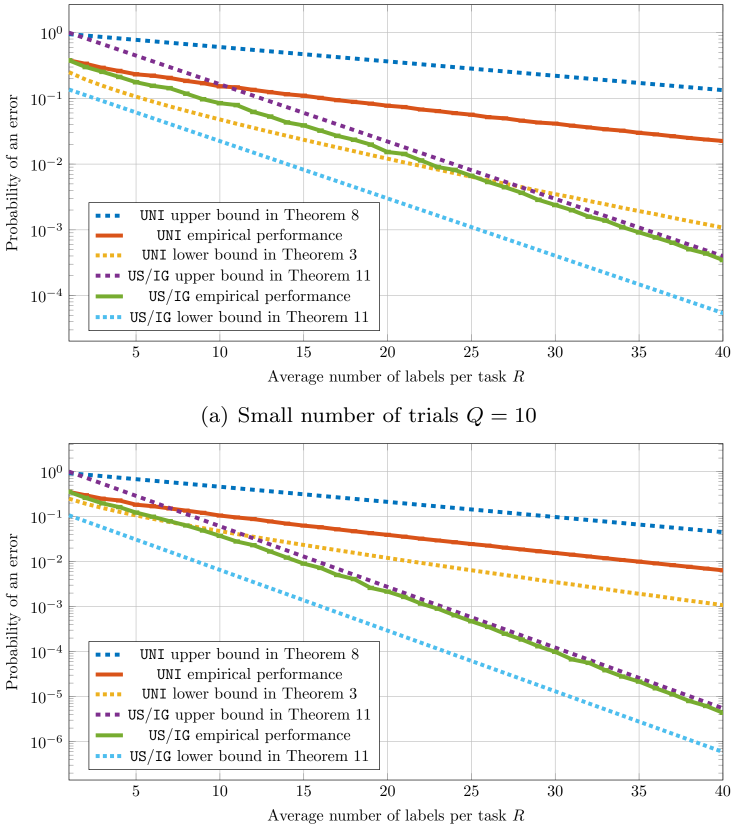
Then, we observe that the term in the product does not depend on $\boldsymbol{x}_i$ and thus we can apply the same procedure we use to prove Theorem 3. □

As is the case for Theorem 8, when the estimates $\hat{\boldsymbol{p}}$ converge to the real values $\boldsymbol{p}$, the bound in Theorem 10 converges to the corresponding result with known accuracies (see Theorem 3). At the same time, we can see that for noisy estimates of $\boldsymbol{p}$ the upper and lower bound in Theorems 8 and 10 do not match asymptotically. This is because the minimisation term in Equation 56 can make the value of the lower bound arbitrarily small, especially for large values of R. When this is the case, Theorem 3 can be used as an alternative lower bound, since knowing the real accuracies $\boldsymbol{p}$ is a best-case scenario in this context.

An example of this is presented in Figure 4. There, we plot the empirical performance of the `UNI` policy with the respective upper bound in Theorem 8 and the best-case lower bound in Theorem 3. As in the known accuracy case (see Figure 1(a)), the empirical performance was evaluated on synthetic data with predictors distributed according to $f_p = \text{Beta}(4, 3)$. Accordingly, the estimates $\hat{\boldsymbol{p}}$ were computed by Bayesian estimation with matching prior. Notice how the performance of the policy improves as the number of trials increases from $Q = 10$ (Figure 4(a)) to $Q = 100$ (Figure 4(b)). In the limit for $Q \to \infty$, the estimates $\hat{\boldsymbol{p}}$ converge to $\boldsymbol{p}$ and we obtain the results of the known accuracy case (Figure 1(a)).

4.3. Performance of the `US` policy

From a high-level perspective, our analysis of the behaviour of the `US` policy in the known accuracy case (see Section 3.3) still holds, with the only difference that here the distribution of steps is $f_{\hat{s}}$ instead of f_s. However, we need an additional condition on the nature of the estimates $\hat{\boldsymbol{p}}$ in order to translate our past results to this new setup. More specifically, we need the estimates $\hat{\boldsymbol{p}}$ to be unbiased so that the relationship $\mathbb{P}(\hat{y}_i = y_i) = \sigma(|\hat{z}_i|)$ still holds. This ensures that we can always use the estimated log-odds as an indicator of how



(a) Small number of trials $Q = 10$

(b) Large number of trials $Q = 100$

Figure 4: Comparison between the theoretical bounds and the empirical performance of the `UNI` and `US` policies under the Bayesian estimation aggregator, with $p_j \sim \text{Beta}(4, 3)$.

uncertain each task is. The Bayesian estimator presented in Section 2.1 is a notable example of such an unbiased estimator.

With these assumptions, we can prove the following two bounds:

Theorem 11. *The probability of a classification error under the* `US` *policy is upper bounded by:*

$$\mathbb{P}(\hat{y}_i \neq y_i) \leq \exp\big(-\mathbb{E}_{f_{\hat{s}}}\{\hat{s}_{ij}\}(R-1)\big) \tag{58}$$

and lower bounded by:

$$\mathbb{P}(\hat{y}_i \neq y_i) \geq \frac{1}{2}\exp\big(-\mathbb{E}_{f_{\hat{s}}}\{\hat{s}_{ij}\}R - \mathbb{E}_{f_{\hat{s}}}\{|\hat{s}_{ij}|\} - 0.56\big) \tag{59}$$

Proof. The proof follows the same steps of the proofs of Theorems 4 and 5. First, we set an arbitrary threshold $\hat{z}_B$. Then, we compute the number of steps

r required to reach it (recall that now the steps are distributed according to $f_{\hat{s}}$ instead of f_s). This yields the following upper and lower bounds:

$$\hat{z}_B > \mathbb{E}_{f_{\hat{s}}}\{\hat{s}_{ij}\}(R-1) \tag{60}$$

$$\hat{z}_B \leq R\mathbb{E}_{f_{\hat{s}}}\{\hat{s}_{ij}\} + \mathbb{E}_{f_{\hat{s}}}\{|\hat{s}_{ij}|\} + 0.56 \tag{61}$$

Thanks to the property that $\mathbb{P}(\hat{y}_i = y_i) = \sigma(|\hat{z}_i|)$, we can use Equations 60 and 61 to get the result in the theorem (after the same algebraic manipulations used for Equations 28 and 33). □

In the case of Bayesian estimation, we assume that $p_j \sim \text{Beta}(4,3)$ and that each predictor has been tested on Q independent trials. As a result, we can compute the value of $\mathbb{E}_{f_{\hat{s}}}\{\hat{s}_{ij}\}$ in closed form as follows:

$$\begin{aligned}\mathbb{E}_{f_{\hat{s}}}\{\hat{s}_{ij}\} &= \sum_{c_j=0}^{Q} \mathbb{P}(c_j|f_p)w_j(c_j)\Big[\mathbb{P}(x_{ij}=1|c_j) - \mathbb{P}(x_{ij}=-1|c_j)\Big] \\ &= \sum_{c_j=0}^{Q} \binom{Q}{c_j}\frac{\mathrm{B}(c_j+\alpha, Q-c_j+\beta)}{\mathrm{B}(\alpha,\beta)} \\ &\qquad\qquad \log\left(\frac{c_j+\alpha}{Q-c_j+\beta}\right)\frac{(c_j+\alpha)-(Q-c_j+\beta)}{Q+\alpha+\beta}\end{aligned} \tag{62}$$

Similarly, the expected value of $\mathbb{E}_{f_{\hat{s}}}\{|\hat{s}_{ij}|\}$ is as follows:

$$\mathbb{E}_{f_{\hat{s}}}\{|\hat{s}_{ij}|\} = \sum_{c_j=0}^{Q} \binom{Q}{c_j}\frac{\mathrm{B}(c_j+\alpha, Q-c_j+\beta)}{\mathrm{B}(\alpha,\beta)}\left|\log\left(\frac{c_j+\alpha}{Q-c_j+\beta}\right)\right| \tag{63}$$

Note that the expected values in Equations 62 and 63 converge to those of the real distribution f_s as the number of samples Q increases. In other terms, the bounds in Theorem 11 converge to those in Theorems 4 and 5 as the Bayesian estimator is allowed more data to form its estimates $\hat{\boldsymbol{p}}$.

An example of this is presented in Figure 4. There, we compare the performance of the US policy with the theoretical bounds derived in Theorem 11. The empirical performance was evaluated on synthetic data with predictors distributed according to $f_p = \text{Beta}(4,3)$, as in the known accuracy case (see Figure 1(b)), and the estimates $\hat{\boldsymbol{p}}$ were computed by Bayesian estimation with matching prior. Notice how the performance of the policy improves as the number of trials increases from $Q = 10$ (Figure 4(a)) to $Q = 100$ (Figure 4(b)). In the limit for $Q \to \infty$, the estimates $\hat{\boldsymbol{p}}$ converge to $\boldsymbol{p}$ and we obtain the results of the known accuracy case (Figure 1(b)).

4.4. Performance of the IG *policy*

Under the assumption in Section 4.3, we can prove that the IG policy is equivalent to the US policy, as we did in the known accuracies case (see Section 3.4):

Theorem 12. *Given the current set of labels X^{t-1} and a worker with estimated weight $\hat{w}_j \neq 0$, the two policies* `US` *and* `IG` *select the same task, except the in case of a tie.*

Proof. The proof follows the same outline as the proof of the corresponding result with known accuracy (see Theorem 6). When computing the expected information gain, it suffices to note that in this scenario we have the following identities:

$$\mathbb{P}(y_i = +1|X^{t-1}) = \sigma(\hat{z}_i^{t-1}) \tag{64}$$

and similarly:

$$\mathbb{P}(x_{ij}|X^{t-1}) = \sigma(\hat{z}_i^{t-1})\sigma(x_{ij^t}\hat{w}_{j^t}) + \sigma(-\hat{z}_i^{t-1})\sigma(-x_{ij^t}\hat{w}_{j^t}) \tag{65}$$

where $\sigma(\hat{w}_{j^t})$ is the probability that the new label x_{ij^t} is correct, given that the estimates $\hat{\boldsymbol{p}}$ are unbiased. □

4.5. Policy comparison

The results presented in Sections 4.2, 4.3 and 4.4 closely resemble those derived in the known accuracy case (see Section 3). In this section we investigate how much the introduction of the estimates $\hat{\boldsymbol{p}}$ has an impact on the relative efficiency of the policies. Since the two adaptive policies `US` and `IG` are equivalent, we are interested in quantifying their advantage over the non-adaptive `UNI` policy. As before, we achieve this goal by considering the value of the constants c in the exponential tradeoff $\mathbb{P}(error) \leq \exp(-cR)$ that all these policies exhibit. For the `UNI` policy, we can derive it from the upper bound of Theorem 8 taken in expectation (see Equation 53):

$$c_{uni} = -\log\left(\mathbb{E}_{f_p, f_{\hat{p}}}\left\{p_j\sqrt{\frac{1-\hat{p}_j}{\hat{p}_j}} + (1-p_j)\sqrt{\frac{\hat{p}_j}{1-\hat{p}_j}}\right\}\right) \tag{66}$$

Similarly, for the two adaptive policies `US` and `IG` we can use the result in Theorem 11 and write the constant $\mathbb{E}_{f_{\hat{s}}}\{s_{ij}\}$ in its extended form:

$$c_{ada} = \mathbb{E}_{f_p, f_{\hat{p}}}\left\{(2p_j - 1)\log\left(\frac{\hat{p}_j}{1-\hat{p}_j}\right)\right\} \tag{67}$$

Taken together, we can use the constants in Equations 66 and 67 to quantify the advantage of the adaptive `US` and `IG` policies over the non-adaptive `UNI` policy. This turns out to be the same as the corresponding result in the known accuracy case (see Theorem 7):

Theorem 13. *For any distribution f_p and any estimates $\hat{\boldsymbol{p}}$ such that $\mathbb{P}(\hat{y}_i = y_i) = \sigma(|\hat{z}_i|)$, we have $c_{ada} \geq 2c_{uni}$.*

Proof. By applying Jensen's inequality twice to Equation 66, we have:

$$\begin{aligned} c_{uni} &\leq \mathbb{E}_{f_p, f_{\hat{p}}} \left\{ -\log \left(p\sqrt{\frac{1-\hat{p}}{\hat{p}}} + (1-p)\sqrt{\frac{\hat{p}}{1-\hat{p}}} \right) \right\} \\ &\leq \mathbb{E}_{f_p, f_{\hat{p}}} \left\{ (2p-1)\log \left(\sqrt{\frac{\hat{p}}{1-\hat{p}}} \right) \right\} \end{aligned} \tag{68}$$

Comparing the argument of the expected value in Equations 67 and 68, we obtain the result in the theorem. □

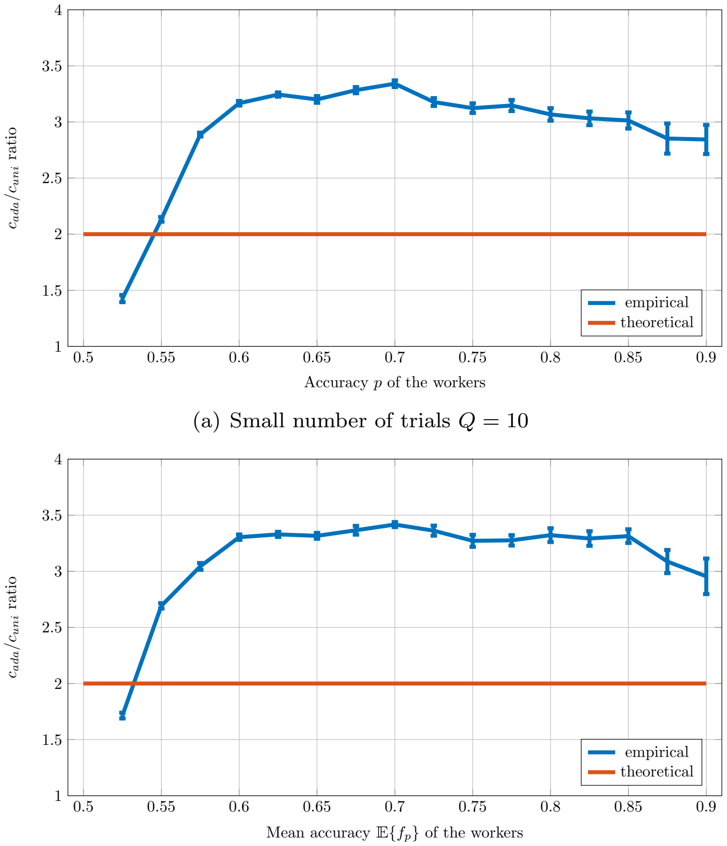
At the same time, the result in Theorem 13 is based on the assumption that $\mathbb{P}(\hat{y}_i = y_i) = \sigma(|\hat{z}_i|)$, which is needed to derive Equation 67. In some instances, this can be impossible to guarantee, for example if we are using Bayesian estimation but the prior distribution on the predictors' accuracy f_p is unknown. In this case, a typical solution is to run the Bayesian estimation with an uninformative prior, i.e. setting $\alpha = \beta = 1$. To study this, we run synthetic experiments with different populations of predictors f_p, and estimate the value of the c_{ada}/c_{uni} ratio empirically. To make the comparison easier, we use the same settings of the corresponding experiments for the known accuracy case (see Figure 3). In this regard, the results with homogeneous predictors are presented in Figure 5, whereas those with mixed predictors are presented in Figure 6.

As the figures show, the empirical value of the c_{ada}/c_{uni} ratio is around 3 for most distributions f_p. As in the known accuracy case, we have lower values for homogeneous distributions $f_p = \delta(\bar{p})$ with $\bar{p}$ close to 0.5. Again, this phenomenon is due to the amount of randomness in the predictors' output. This reduces the ability of the aggregator to discriminate between the two classes $y_i = \pm 1$, thus making the impact of the data collection policy less important. Finally, notice how the number of trials Q does not have a considerable effect on the relative performance of the policies. In the limit for $Q \to \infty$ the estimates $\hat{\boldsymbol{p}}$ converge to $\boldsymbol{p}$ and we obtain the same c_{ada}/c_{uni} ratio as the one presented in Figure 3.

5. Conclusions

In this paper, we analysed the efficiency of data collection policies in a setting with multiple Naïve Bayes classifiers and shared predictors. This setting arises when we have a set of classification tasks to solve, and only access to a limited pool of independent predictors to gather information from. As the existing literature shows, the policy we use to assign the predictors to the classification tasks has a considerable impact on the accuracy of the final predictions. However, the existing works are mainly empirical and fail to explain the underlying theoretical reasons behind this phenomenon.

This is particularly unfortunate from a system designer's perspective, as the lack of strong theoretical guarantees makes every design choice more challenging.



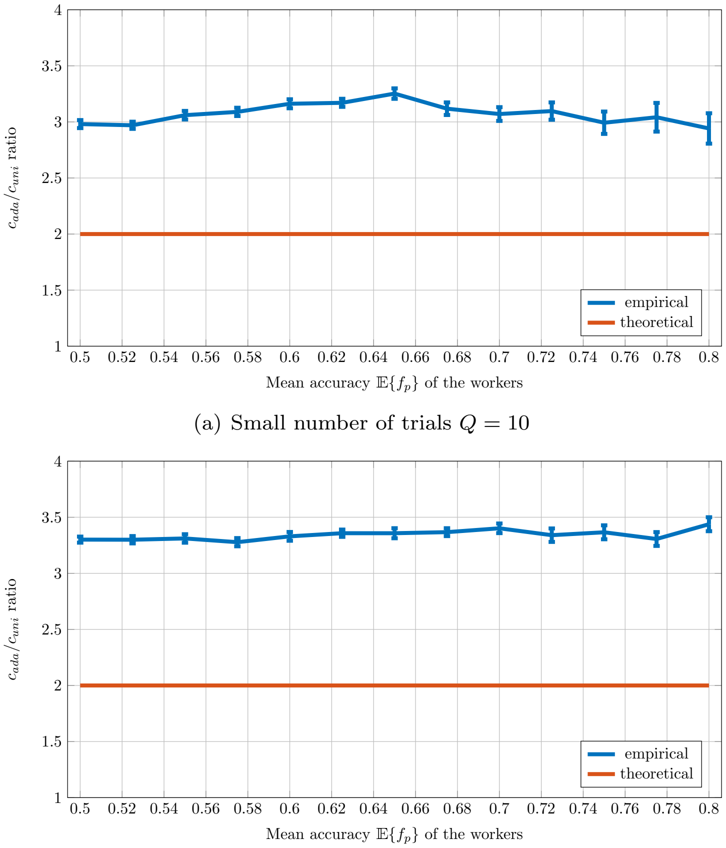
(a) Small number of trials $Q = 10$

(b) Large number of trials $Q = 100$

Figure 5: Comparison between the empirical c_{ada}/c_{uni} and its theoretical baseline for a homogeneous distribution $f_p = \delta(\bar{p})$.

In fact, not only is it difficult to predict the correct amount of resources needed to attain the target accuracy, but also it is not clear which collection policy performs the best under the given circumstances. Furthermore, in contexts like crowdsourcing, the pool of predictors can change rapidly over time. This is due to workers constantly dropping in and out of the crowdsourcing platform. As a result, it may be difficult to test and compare the merits of different methods in the real world. On the contrary, having strong theoretical guarantees can help the pratictioner make more informed decisions on the amount of resources to allocate and which data collection policy to use.

In order to address these issues, we analysed three of the most popular collection policies: the non-adaptive uniform allocation policy, and the adaptive policies of uncertainty sampling and information gain maximisation. By representing them as random walks in the log-odds domain, we were able to explain



(a) Small number of trials $Q = 10$

(b) Large number of trials $Q = 100$

Figure 6: Comparison between the empirical c_{ada}/c_{uni} and its theoretical baseline for a mixed distribution $f_p = \text{Uniform}(\bar{p} - 0.2, \bar{p} + 0.2)$.

their behaviour and derive new upper and lower bounds on their accuracy. With this view, we derived the following three main results. First, we proved that the two adaptive policies of uncertainty sampling and information gain maximisation are equivalent. Second, we proved that these two adaptive policies are superior to the non-adaptive uniform allocation policy and measured the performance gap between them for the first time. Third, we repeated our analysis under two different scenarios: when we have perfect knowledge of the accuracy of each single predictor, and when we only have access to some estimates of it. By comparing the outcomes of our analysis of the two cases, we showed that the performance gap between the policies remains constant and does not depend on our knowledge of the predictors' accuracy.

These results confirm the empirical intuition that adaptive policies like uncertainty sampling and information gain maximisation have an intrinsic advantage over non-adaptive ones like uniform allocation. Consequently, the former should be always preferred if the circumstances allow it. Furthermore, our bounds on the error-resources tradeoff of the three collection policies we analysed in this paper can be used to predict the amount of resources, i.e. number of individual predictors, needed to reach a given target accuracy. Similarly, if the amount of resources is fixed in advance, our bounds can be used to predict the resulting classification accuracy.

In addition to this, our work also constitutes a general contribution in the field of ensemble classifiers [3]. This is when several machine learning subsystems are trained separately and their outputs aggregated to form a final prediction. In this regard, they can be seen as a group of predictors solving a single classification task. When we can assume independence between them, and we have at least an estimate of their individual accuracy, the results in this paper apply. More specifically, the bounds in Theorems 1, 3, 8 and 10 can be used to compute the number of machine learning subsystems needed to attain the desired aggregated accuracy.

On a different note, we believe that the techniques showed here can be extended to many other cases. A first direction is that of non-binary classifiers. In this regard, the work of Gao et al. [7] can provide a guide on how to translate similar results from the binary to the non-binary case. A second direction is that of probabilistic inference aggregators. In this case, the predictors' accuracy is estimated from the set of labels X^t itself. This has the effect of coupling the decisions of the data collection policy on which task to label with the estimation of $\hat{\boldsymbol{p}}$. Moreover, the weights $\hat{\boldsymbol{w}}$ change at every timestep, making the evolution of the random walk more complex to model. A third direction is that of tasks with different levels of difficulty. Here, the main challenge lies in computing the drift of the random walk as it varies from one task to another. The work of Khetan and Oh [10] provides a different approach to this case, but their results are only asymptotical. Instead, we believe that our approach can lead to tight finite-sample bounds on the error-resources tradeoff of the collection policies.

Acknowledgments

This work is an extension of a paper accepted at the 2018 IJCAI conference in Stockholm [16]. This research was funded by the UK Research Council project ORCHID, grant EP/I011587/1. The authors acknowledge the use of the IRIDIS High Performance Computing Facility, and associated support services at the University of Southampton.

References

[1] D. W. Barowy, C. Curtsinger, E. D. Berger, and A. McGregor. AutoMan: A Platform for Integrating Human-based and Digital Computation. In *Proceedings of the ACM International Conference on Object Oriented Programming Systems Languages and Applications*, pages 639–654, 2012.

[2] D. Berend and A. Kontorovich. A Finite Sample Analysis of the Naive Bayes Classifier. *Journal of Machine Learning Research*, 16(1):1519–1545, 2015.

[3] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, 2006.

[4] T. Bonald and R. Combes. A Minimax Optimal Algorithm for Crowdsourcing. In *Proceedings of the Thirtieth International Conference on Neural Information Processing Systems*, pages 4355–4363. 2017.

[5] X. Chen, Q. Lin, and D. Zhou. Optimistic Knowledge Gradient Policy for Optimal Budget Allocation in Crowdsourcing. In *Proceedings of the Thirtieth International Conference on Machine Learning*, volume 28, pages 64–72, 2013.

[6] A. P. Dawid and A. M. Skene. Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):20–28, 1979.

[7] C. Gao, Y. Lu, and D. Zhou. Exact Exponent in Optimal Rates for Crowdsourcing. In *Proceedings of the Thirty-Third International Conference on Machine Learning*, pages 603–611, 2016.

[8] C.-J. Ho, S. Jabbari, and J. W. Vaughan. Adaptive Task Assignment for Crowdsourced Classification. In *Proceedings of the Thirtieth International Conference on Machine Learning*, pages 534–542, 2013.

[9] D. R. Karger, S. Oh, and D. Shah. Budget-Optimal Task Allocation for Reliable Crowdsourcing Systems. *Operations Research*, 62(1):1–24, 2014.

[10] A. Khetan and S. Oh. Achieving Budget-Optimality with Adaptive Schemes in Crowdsourcing. In *Proceedings of the Twenty-Ninth International Conference on Neural Information Processing Systems*, pages 4844–4852, 2016.

[11] S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[12] D. D. Lewis and W. A. Gale. A Sequential Algorithm for Training Text Classifiers. In *Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3–12, 1994.

[13] N. Littlestone and M. K. Warmuth. The Weighted Majority Algorithm. *Information and Computation*, 108(2):212–261, 1994.

[14] Q. Liu, J. Peng, and A. Ihler. Variational Inference for Crowdsourcing. In *Proceedings of the Twenty-Fifth International Conference on Neural Information Processing Systems*, pages 692–700, 2012.

[15] D. J. C. MacKay. Information-Based Objective Functions for Active Data Selection. *Neural Computation*, 4(4):590–604, 1992.

[16] E. Manino, L. Tran-Thanh, and N. R. Jennings. On the Efficiency of Data Collection for Crowdsourced Classification. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pages 1568–1575, 2018.

[17] Marquis de Condorcet. *Essai sur l'Application de l'Analyse à la Probabilité des Décisions Rendus à la Pluralité des Voix.* Imprimerie Royale, Paris, 1785.

[18] S. Nitzan and J. Paroush. Optimal Decision Rules in Uncertain Dichotomous Choice Situations. *International Economic Review*, 23(2):289–297, 1982.

[19] D. Oleson, A. Sorokin, G. Laughlin, V. Hester, J. Le, and L. Biewald. Programmatic Gold: Targeted and Scalable Quality Assurance in Crowdsourcing. In *Proceedings of the 11th AAAI Conference on Human Computation*, pages 43–48, 2011.

[20] S. D. Ramchurn, F. Wu, W. Jiang, J. E. Fischer, S. Reece, S. Roberts, T. Rodden, C. Greenhalgh, and N. R. Jennings. Human–agent collaboration for disaster response. In *Proceedings of the 15th International Conference on Autonomous Agents and Multiagent Systems*, pages 82–111, 2016.

[21] E. Simpson and S. Roberts. Bayesian Methods for Intelligent Task Assignment in Crowdsourcing Systems. In *Scalable Decision Making: Uncertainty, Imperfection, Deliberation*, pages 1–32. Springer, 2014.

[22] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng. Cheap and Fast – but is It Good?: Evaluating Non-expert Annotations for Natural Language Tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 254–263, 2008.

[23] A. Wald. On Cumulative Sums of Random Variables. *The Annals of Mathematical Statistics*, 15(3):283–296, 1944.

[24] P. Welinder and P. Perona. Online Crowdsourcing: Rating Annotators and Obtaining Cost-Effective Labels. In *Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition – Workshops*, pages 25–32, 2010.