Don't Label Twice: Quantity Beats Quality when Comparing Binary Classifiers on a Budget

Florian E. Dorner [1 2 3] **Moritz Hardt** [1 2]

Abstract

We study how to best spend a budget of noisy labels to compare the accuracy of two binary classifiers. It's common practice to collect and aggregate multiple noisy labels for a given data point into a less noisy label via a majority vote. We prove a theorem that runs counter to conventional wisdom. If the goal is to identify the better of two classifiers, we show it's best to spend the budget on collecting a single label for more samples. Our result follows from a non-trivial application of Cramér's theorem, a staple in the theory of large deviations. We discuss the implications of our work for the design of machine learning benchmarks, where they overturn some time-honored recommendations. In addition, our results provide sample size bounds superior to what follows from Hoeffding's bound.

1. Introduction

Data annotators are the "AI revolution's unsung heros," Gray & Suri (2019) argued. The labor of human annotators has powered a growing industry of machine learning datasets and benchmarks since the 1980s (Hardt & Recht, 2022). Human labels are a precious, yet unreliable resource. Errors easily creep into data labor at scale. The designer of a benchmark has to cope with the reality of conflicting labels for the same data point.

Many benchmarks follow a common strategy. Each data point in a sample gets noisy labels from multiple human annotators. The candidate labels then determine a single label via an aggregation function, such as a majority vote in the case of binary labels. For a sample of size n and a choice of m labels per data point, the cost of this design scales as mn. Although ubiquitous, we prove that this strategy is wasteful for creating the test set.

[1]Max Planck Institute for Intelligent Systems, Tübingen [2]Tübingen AI Center [3]ETH Zürich. Correspondence to: Florian E. Dorner <florian.dorner@tuebingen.mpg.de>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

When the goal is to compare the population accuracy of binary classifiers, it is better to sample mn data points and collect a single noisy label for each. The basis of our main result is a simple mathematical model that captures the essential question. Data points are independent and identically distributed. For each data point, we can request an odd number $m \geq 1$ of binary labels, drawn independently from a distribution that picks the correct label with some probability strictly greater than chance. We then aggregate the m labels into a single label using a majority vote. Fix two classifiers, one better than the other in terms of population accuracy by some positive margin. We have a budget k to spend on labels. Given an annotator number m, we can create a labeled sample of size $n = k/m$. We pick the classifier with the higher empirical accuracy on this sample. How should we pick an annotator number m so as to maximize the probability of picking the better classifier? Our main theorem provides the answer.

Theorem 1 (Informal)**.** *For a sufficiently large sample budget k, the probability of identifying the better of two binary classifiers is maximized at $m = 1$ labels per data point.*

As a rough intuition, the gains in label accuracy from aggregation are outweighed by the loss in sample size and information about classifier disagreements. Formally, our result follows from a non-trivial—and rather lengthy—reduction to Cramér's theorem. Cramér's theorem is fundamental to the theory of large deviations. It provides precise control over tail probabilities, based on the Legendre transform of the logarithmic moment generating function. The theorem is asymptotic with respect to n, complicating the application to our problem. However, standard concentration inequalities, such as Hoeffding's bound, are insufficient for our purposes as they only provide upper bounds.

Our theorem extends to the case where label errors are correlated with classifier errors, possibly even in a data dependent way. It only fails in the unusual cases where label noise systematically aligns in favor of the worse classifier, the effect of aggregation on label quality systematically aligns in favor of the better classifier, or the cost of unlabelled data is large. While we prove our main theorem for sufficiently large sample sizes, we conjecture that the statement holds for all $n \geq 1$. We have verified the conjecture

numerically in a vast parameter sweep spanning more than four billion values. The numerical tool we created for verifying the conjecture also serves as an effective way to calculate tight sample size requirements for given parameters and is available at https://labelnoise.is.tuebingen.mpg.de.

Our result applies to the case of many classifiers via the union bound. Here, it gives an answer to the question how many classifiers we can reliably rank in a machine learning benchmark. This question is commonly answered in theory by combining the union bound with Hoeffding's inequality. We demonstrate that our bound permits exponentially more comparisons than the standard argument for the same sample budget. Figure 1 illustrates the improvement.

There is a common belief that benchmark designers should invest in cleaning noisy labels through aggregation. Our result suggests a surprising departure. For the purpose of comparing and ranking binary classifiers, quantity beats quality. A single label per data point is optimal.

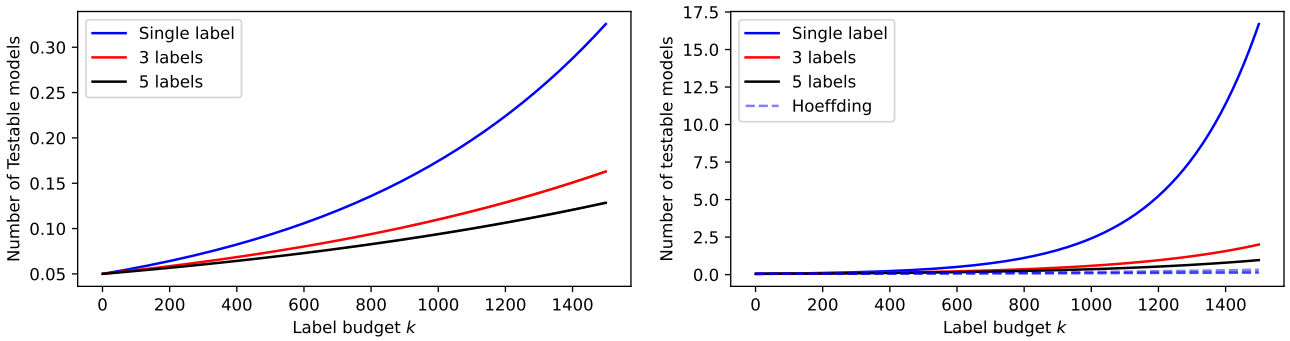
1.1. Related Work

Label aggregation in dataset creation. Human-provided labels are at the heart of modern machine learning, both in industry (Gray & Suri, 2019) and academic benchmarking. Many important datasets have been labeled by humans, with "gold standard" labels produced by aggregating multiple annotators' labels: In image recognition, labels for CIFAR-10 (Krizhevsky et al., 2009) were verified by the work's authors after being initially labeled by others, while labels in ImageNet (Russakovsky et al., 2015) are aggregated from multiple crowdworker annotations. Similarly, the target label for medical datasets is often established by a majority vote over expert annotators like sonographers (Tanno et al., 2019) or radiologists (Nguyen et al., 2022). In natural language processing, classic benchmarks like MSRP (Dolan & Brockett, 2005), SST (Socher et al., 2013), SICK (Marelli et al., 2014) and MNLI (Bowman et al., 2015) all base labels on a per-instance majority vote after collecting multiple labels for each instance. More recently, label aggregation has been used to define test labels in Kaggle's Jigsaw Unintended Bias in Toxicity Classification challenge (Jigsaw, 2019) and for evaluating the safety of LLama2 (Touvron et al., 2023). Similarly, OpenAssistant (Köpf et al., 2023) aggregates users' rankings for the same list of model outputs into a "consensus opinion". Recht et al. (2019) suggest to "employ a separate labeling process for the test set that relies on more costly expert annotations." In line with this, it is common to collect a larger amount of labels per instance for *testing* than for training (Williams et al., 2017; Dorner et al., 2022; Nguyen et al., 2022) to increase label quality.

The impact of label aggregation on learning. While label aggregation is a common practice in dataset and benchmark creation, its impacts on training and evaluation are not fully understood: On the theoretical side, Crammer et al. (2005) provide performance bounds that depend on the quality and size of training data and can be used to heuristically choose between data sources. Wei et al. (2023) analyze whether duplicate labels for the same data point should be aggregated or treated independently for empirical risk minimization and find the latter to perform better if disagreement is common. Empirically Sheng et al. (2008) show that for certain decision tree learners, a large number of noisy labels per instance x beats single labels for more data points when labels are very noisy. On the other hand, Chen et al. (2021) provide empirical evidence that for realistic label noise, the opposite is true for finetuning modern language models. In line with that, Lin et al. (2014) show that the benefits of relabeling can depend both on the problem domain and hyperparameters of the learning algorithm. In contrast to these works, our work focuses on comparing already learnt classifiers, not classifier training.

Annotator disagreement as a feature. Aroyo & Welty (2013) argue that due to the lack of objective ground truth for many tasks, taking annotator disagreement into account is essential. The authors suggest to to use non-binary labels that encompass disagreement. Ramponi & Leonardelli (2022) and Sandri et al. (2023) use predicting annotator disagreement as an auxillary task for detecting offensive language, while Cheplygina & Pluim (2018) show that annotator disagreement itself can be an informative feature in medical image analysis. Meanwhile, Tanno et al. (2019) and Davani et al. (2022) suggest to predict individual annotators' responses. This approach, combined with focusing on annotators relevant for a given contexts, is also used to mitigate majoritarian biases caused by aggregation (Gordon et al., 2022; Fleisig et al., 2023). As these approaches require annotator-level labels, it is often recommended for dataset creators to release these rather than already aggregated labels (Prabhakaran et al., 2021; Denton et al., 2021). Our work is orthogonal: We focus on cases where the target label is agreed upon to be given by a (fictitious) majority vote over the whole crowdworker population. In this setting, we demonstrate that collecting and aggregating multiple labels per data point is *statistically* suboptimal in terms of identifying the better of two classifiers.

The theory of benchmarking. Benchmarking plays an important role in machine learning, but is rarely studied. An exception is work on *adaptive overfitting*: For the test error to estimate the population risk without bias, models have to be trained without knowledge about the test set, which is rarely true for real benchmarks. To see whether this causes problems in practice, Recht et al. (2019) recre-



(a) Hoeffding-based guarantees for $\delta = 0.05$

(b) Cramér-based guarantees for $\delta = 0.05$

Figure 1. Number of testable classifiers according to the Hoeffding (a) and Cramér-based (b) upper bounds on the error probability and a union bound (see Section 3.3) for accuracies $p = q = 0.75$, margin $\epsilon = 0.1$ and error tolerance $\delta = 0.05$. Note the different y axes.

ated the ImageNet test set based on the original procedure. They find that classifier accuracy on the new test set is lower, but strongly correlates with the original accuracy such that model rankings are remarkably stable. Mania & Sra (2020) theoretically explain these observations based on correlations between classifiers. Lastly, Blum & Hardt (2015) show that the impacts of adaptive overfitting can be reduced by only revealing a classifier's test accuracy if it is substantially better than the previous best.

2. Formal Setup

Let $\mathcal{D}$ be a distribution of data points x with binary correct labels $y_{True}(x) \in \{0, 1\}$. For a binary classifier c, we define the population risk as the expected frequency of classification errors

$$\mathcal{R}(c) := \mathop{\mathbb{E}}_{x \sim \mathcal{D}}[\mathbb{I}(y_{True}(x) \neq c(x))],$$

where $\mathbb{I}$ denotes the indicator function. We consider two arbitrary classifiers c_b and c_w (where b stands for "better" and w for "worse"), such that

$$1 - p = \mathcal{R}(c_w) > \mathcal{R}(c_b) = 1 - p - \epsilon$$

for accuracy $p \in [0.5, 1]$ and margin $\epsilon \in (0, 1-p]$. We want to use a limited labeling budget k to create a test set T on which test accuracy is likely to be higher for the better classifier, without using any information about the two specific classifiers at hand. We assume that test sets are created using the following sampling procedure: Independently (with replacement) sample a dataset D of n data points $x \sim \mathcal{D}$. Then, for each $x \in D$, sample $m = \frac{k}{n}$ labelers l from a population of crowdworkers $\mathcal{D}_{crowd}$, again independently and with replacement, and have each of them provide a label $y_l(x)$ for x. For a given data point x, we then set the test label $y_{Test}(x)$ equal to the majority of the labels $y_l(x)$. The main question tackled in this work is then, how to allocate the label budget k between n and m in order to have the best chance of correctly identifying the better classifier c_b using the constructed test set. We will particularly focus on comparing the case of $m = 1$ to $m > 1$, as we find strong evidence that $m = 1$ is optimal in most cases.

For a fixed data point x, we set $q(x) \in (0.5, 1]$ to the probability that a crowdworker label $y_l(x)$ is correct, marginalized over l, i.e. $q(x) := \mathbb{P}_l(y_l(x) = y_{True}(x))$. Similarly, q denotes the same probability marginalized over both x and l: $q := \mathbb{P}_{x,l}(y_l(x) = y_{True}(x))$. We note, that in this setup, the case of collecting m labels $y_l(x)$ for a given x with correctness probability $q(x)$ yields the same distribution of labels as collecting a single label with correctness probability $q'(x) = M_m(q(x))$, where

$$M_m(q) := \mathbb{P}(\text{Majority of } m \text{ independent voters correct})$$

under the assumption that each voter is correct with probability q. To compare the two classifiers c_b and c_w on our test set, we define the gap indicator G

$$G := \begin{cases} 1: & c_b(x) = y_{Test(x)} \neq c_w(x) \\ -1: & c_w(x) = y_{Test(x)} \neq c_b(x) \\ 0: & c_w(x) = c_b(x) \end{cases},$$

where x and y_{Test} are sampled as described above. The gap indicator G describes the unnormalized accuracy gap between the classifiers c_b and c_w on the test set, as we can express

$$\frac{1}{n} \sum_{i=1}^{n} G_i = \mathrm{Acc}_{Test}(c_b) - \mathrm{Acc}_{Test}(c_w)$$

for independent copies G_i of G, where $\mathrm{Acc}_{Test}(c) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(y_{Test}(x) = c(x))$. In particular $\sum_{i=1}^{n} G_i$ is positive if and only if our test set correctly identifies the better classifier c_b.

3. Parameterizing the Gap Indicator

We begin by considering the fully independent case in which the error events $c_w(x) \neq y_{True}(x)$, $c_b(x) \neq y_{True}(x)$ are independent of each other and the label accuracy $q(x)$. We will treat the gap indicator G as a function of q and assume homogeneous label errors over x, i.e. $q(x) = q$. This assumption allows us to use the equivalence of a single labeler with accuracy $M_m(q(x))$ and m labelers with accuracies $q(x)$ each, to compare $G(q)$ and $G(M_m(q))$ rather than explicitly parameterizing G by m. We note that this assumption yields the best case for the $m-$label approach: If the label accuracy $q(x)$ depends strongly on x, majority voting might not actually yield noticeable benefits in terms of label accuracy. As an extreme example, if $q(x)$ only takes on values in $\{0, 1\}$, majority voting has no benefits at all. Formally, Jensen's inequality and the well-known concavity of the majority vote in $M_m(z)$ in z for $z \in (0.5, 1]$ (Boland et al., 1989) imply $\mathbb{E}_x[M_m(q(x))] \leq M_m(\mathbb{E}_x[q(x)]) = M_m(q)$. This means that $M_m(q)$ can only overestimate label quality for the m-label case. The following proposition provides a precise parametric characterization for G with $m = 1$ in that case, and is proven in Appendix B.

Proposition 1. *Assuming mututally independent classifier and labeler errors, G can be written as follows:*

$$G(q, p, \epsilon) = \begin{cases} 1 & \text{w.p. } q\epsilon + (1-p-\epsilon)p \\ -1 & \text{w.p. } (1-q)\epsilon + (1-p-\epsilon)p \\ 0 & \text{else } p(p+\epsilon) + (1-p-\epsilon)(1-p) \end{cases},$$

for label accuracy q, classifier accuracy p and margin ϵ.

The expectation of the random variable $G(q, p, \epsilon)$ thus equals $(2q-1)\epsilon$, which is positive as $q > 0.5$.

We are now interested in whether using an $m-$majority vote of the noisy crowdworker labels provides more information about which of the two classifiers is better than using m times as many data points with a single label each. More precisely, we would like to find out whether the better classifier is more likely to win with a single label and more data points, or with aggregated labels. In technical terms, we thus want to know whether

$$\mathbb{P}\left(\sum_{i=0}^{mn} G_i(q, p, \epsilon) > 0\right) > \mathbb{P}\left(\sum_{i=0}^{n} G_i(M_m(q), p, \epsilon) > 0\right)$$

for independent copies G_i of G. For small n and m we can calculate the exact probabilities of identifying the better classifier c_b using test sets with different levels of label accuracy q. According to a large scale grid search over the possible values of p, q and ϵ that evaluated nearly five billion configurations, detailed in Appendix A, using $m = 1$ labels is consistently the best approach. Figure 2 demonstrates this, showing the exact probabilities for fixed accuracy $p = 0.8$, margin $\epsilon = 0.01$ and varying values of the label accuracy q and label budget k.

3.1. Hoeffding Bounds

Hoeffding's inequality yields the following lemma proven in Appendix C that allows us to lower bound the probability that c_b beats c_w on a test set:

Lemma 1. *For independent copies X_i of any random variable X with $\mathbb{E}[X] > 0$ and values in $[-1, 1]$, we can bound*

$$\mathbb{P}\left(\sum_{i=0}^{n} X_i \leq 0\right) \leq e^{\frac{-n\,\mathbb{E}[X]^2}{2}} =: B(X, n).$$

We will use this lemma to gain some initial intuition about the quality of test sets constructed with $m = 1$, compared to $m > 1$ labelers per data point x. Specifically, we get a higher lower bound for $\mathbb{P}(\sum_{i=0}^{mn} G_i(q, p, \epsilon) > 0)$ than for $\mathbb{P}(\sum_{i=0}^{n} G_i(M_m(q), p, \epsilon) > 0)$, whenever

$$nm\,\mathbb{E}[G_i(q, p, \epsilon)]^2 > n\,\mathbb{E}[G_i(M_m(q), p, \epsilon)]^2. \quad (1)$$

Informally, equation (1) states that the gains in terms of squared expectation from aggregating multiple labels do not outweigh the simple factor m achieved by labeling multiple data points. It is equivalent to

$$\sqrt{m} > \frac{\mathbb{E}[G_i(M_m(q), p, \epsilon)]}{\mathbb{E}[G_i(q, p, \epsilon)]} = \frac{2M_m(q) - 1}{2q - 1}. \quad (2)$$

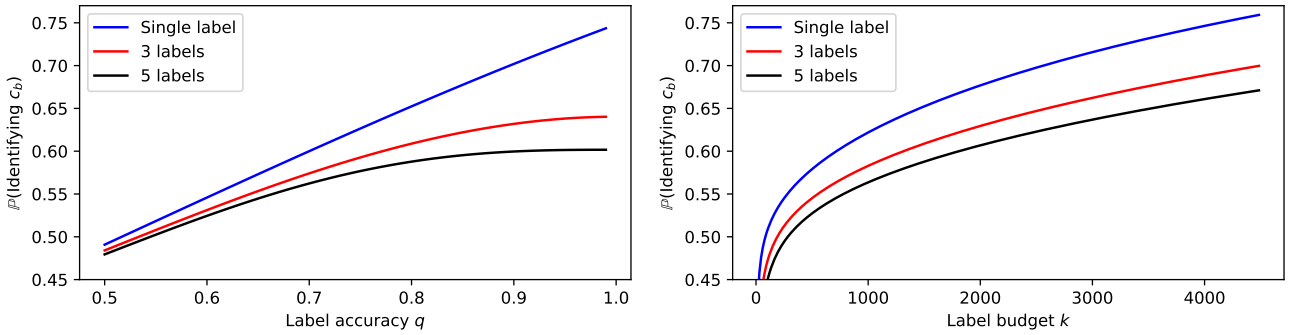
For $m = 3$, this becomes

$$\sqrt{3} > \frac{6q^2 - 4q^3 - 1}{2q - 1} = 1 + 2(1-q)(q),$$

with the right side maximized at $q = 0.5$, with a value of $1.5 < 1.73 \approx \sqrt{3}$, such that (2) is true for all $q \in (0.5, 1]$. In Appendix C, we prove that $B(G(q, p, \epsilon), mn) < B(G(M_m(q), p, \epsilon), n)$ holds for any $m > 1$.

3.2. Correlated Classifiers

The previous Sections assumed both classifiers and the labels to be independent, which is unlikely in practice, as certain examples might be more difficult than others. In this Section, we relax this assumption by modelling the worse classifier c_w to be correct with probability $p_w \in [0.5, 1]$. Then, the better classifier c_b is correct with probability $p_b^0 \in [0.5, 1]$ conditional on c_w being incorrect on a given datapoint x and $p_b^1 \in [0.5, 1]$ conditional on c_w being correct. The assumption that c_b has lower risk than c_w implies $(1-p_w)p_b^0 + p_w p_b^1 > p_w$ or equivalently $(1-p_w)p_b^0 + p_w(p_b^1 - 1) > 0$. We also model correlations between the two classifiers and the labels by denoting $q_b \in (0.5, 1]$ as the probability that the label is correct,



(a) Exact probabilities for $p = 0.8$, $\epsilon = 0.01$, $k = 1500$

(b) Exact probabilities for $p = 0.8$, $\epsilon = 0.01$, $q = 0.8$

Figure 2. Probability of identifying c_b for accuracy $p = 0.8$, margin $\epsilon = 0.01$, budget $k = 1500$ (a), label accuracy $q = 0.8$ (b).

conditional on the event E_b that c_b is correct and c_w is incorrect, and $q_w \in (0.5, 1]$ as the probability that the label is correct in the case that c_b is incorrect and c_w is correct, termed E_w.[1]

Proposition 2. *Assuming correlated classifiers and labels with the above parameterization, we have:*

$$G(q,p) = \begin{cases} 1 & \text{w.p. } q_b(1-p_w)p_b^0 \\ & +(1-q_w)p_w\big(1-p_b^1\big) \\ -1 & \text{w.p. } (1-q_b)(1-p_w)p_b^0 \\ & +q_w p_w\big(1-p_b^1\big) \\ 0 & \text{else} \end{cases}.$$

Then, the expectation of the gap indicator G equals

$$(2q_b-1)(1-p_w)p_b^0 - (2q_w-1)p_w\big(1-p_b^1\big),$$

which is larger than zero if and only if

$$(2q_b-1) > (2q_w-1)\frac{p_w\big(1-p_b^1\big)}{(1-p_w)p_b^0}.$$

The factor $\frac{p_w\left(1-p_b^1\right)}{(1-p_w)p_b^0}$ is smaller than one, as long as $(1-p_w)p_b^0 - p_w\big(1-p_b^1\big) > 0$, which is true as c_b has lower risk than c_w. This means that G is guaranteed to have positive expectation, whenever $q_b \geq q_w$. This essentially ensures that c_w is not overfit to the label noise more than c_b. We assume this to be true:

Assumption 1. *No biased label accuracy:* $q_b \geq q_w$.

An alternative interpretation of assumption 1 is, that the examples in E_w for which the better classifier is incorrect are be more "difficult" than the ones in E_b. For example, consider c_b correct on all but for the top 10% most difficult examples, and c_w random. Then E_w is a subset of the top 10% most difficult examples, such that annotators make more errors and q_w is low. At the same time, E_b only contains examples in the bottom 90%, such that q_b is large.

If assumption 1 does not hold, for example because c_w was trained on parts of the test set, the expectation of G can become negative, such that $\mathbb{P}(\sum_i^n G_i > 0)$ converges to zero. In these cases, narrowing the gap between q_b and q_w by aggregating labels ($M_m(q_w) \approx M_m(q_b)$ for large m) can have large benefits by causing the expectation to become positive, thus flipping the limit of $\mathbb{P}(\sum_i^n G_i > 0)$ from zero to one.

For the $m-$label case, we again focus on (conditionally) homogeneous label errors over x, i.e. $q(x) = q_b$ when $x \in E_b$ and $q(x) = q_w$ when $x \in E_w$, so that we can replace $G(q_b, q_w)$ by $G(M_m(q_b), M_m(q_w))$ rather than explicitly parameterizing G by m. Note that in this case, homogeneity in the label accuracy $q(x)$ is not necessarily the best case for $m > 1$ any more: Heterogeneity lowering the label accuracy of the majority vote can be beneficial as long it is restricted to E_w, where c_b is incorrect. We assume that heterogeneity does not disproportionately harm label accuracy when the better classifier is incorrect:

Assumption 2. *No biased heterogeneity:*

$$\frac{(1-p_w)p_b^0}{p_w(1-p_b^1)}\Big(M_m(q_b) - \mathop{\mathbb{E}}_x[M_m(q(x))|E_b]\Big) \geq M_m(q_w) - \mathop{\mathbb{E}}_x[M_m(q(x))|E_w]. \quad (3)$$

The $\frac{(1-p_w)p_b^0}{p_w\left(1-p_b^1\right)}$ factor is larger than one as c_b is more accurate than c_w. Because $q_b > q_w$ and $M_3(x)$ is more concave for larger $x > 0.5$, this means that for $m = 3$ assumption 2 is expected to hold whenever there are similar levels of heterogeneity conditional on the events E_b and

[1] As data points x for which both agree do not influence the gap indicator, label accuracy can be arbitrary for such points, i.e. x that are neither in E_b nor in E_w.

E_w. For simplicity of notation, we will sometimes use p as a shorthand for p_w, p_b^0, p_b^1, q as a shorthand for q_b, q_w and $M_m(q)$ as a shorthand for $M_m(q_b), M_m(q_w)$, again obtaining $B(G(q,p),nm) > B(G(M_m(q),p),n)$ for any $m > 1$ under assumption 1, as proven in Appendix C.

3.3. Application to Benchmarking

The different bounds on the error probabilities for a single vs m labels are straightforward to extend to benchmarking, where we compare multiple classifiers: Formally, we consider a classifier c_b with risk $\mathcal{R}(c_b) = 1 - p - \epsilon$ that is better than k other classifiers $c_i, i \leq k$ with (larger) risk $\mathcal{R}(c_i) \geq 1-p$. A test set is a good benchmark, if c_b has the highest test accuracy with high probability. We can bound the probability that the benchmark fails to identify the best classifier c_b using a standard union bound argument:

$$\mathbb{P}\left(\mathrm{Acc}_{Test}(c_b) \leq \max_{i\leq k} \mathrm{Acc}_{Test}(c_i)\right) \leq \sum_{i\leq k} \mathbb{P}(\mathrm{Acc}_{Test}(c_b) \leq \mathrm{Acc}_{Test}(c_i)).$$

Now if $\mathbb{P}(\mathrm{Acc}_{Test}(c_b) \leq \mathrm{Acc}_{Test}(c_i)) \leq e^{-dn\epsilon^2}$ for some $d > 0$ and all $i \leq k$ as suggested by the Hoeffding bounds from the last Section, we get

$$\delta := \mathbb{P}\left(\mathrm{Acc}_{Test}(c_b) \leq \max_{i\leq k} \mathrm{Acc}_{Test}(c_i)\right) \leq ke^{-dn\epsilon^2}.$$

If we want to bound the probability of not identifying the best classifier c_b to a fixed $\delta > 0$, we can thus test at most $k = e^{dn\epsilon^2}\delta$ different classifiers. Correspondingly under the assumptions from before, moving from an $e^{-d_1 n\epsilon^2}$ to an $e^{-d_2 n\epsilon^2}$ bound for $d_2 > d_1$ by not collecting multiple labels per data point allows us to benchmark $e^{(d_2-d_1)n\epsilon^2}$ times as many classifiers while guaranteeing a given bound on the error probability δ.

This exponential improvement is illustrated in Figure 1, which also illustrates the lack of tightness of Hoeffding bounds in our setting, when compared to the bounds provided by Cramér's Theorem discussed in the next Section: For a label budget of $k = 1500$, Cramér guarantees the testability of more than 17 models in the single label case, while Hoeffding is too loose to provide a guarantee for two models at error tolerance $\delta = 0.05$.

4. Proof of the Main Theorem

The results proven above text are suggestive, but do not prove that a single label is optimal. This is because we compare lower bounds that could have systematically different levels of tightness for the single label compared to the $m-$label case. As a large test set not correctly identifying the better classifier is a tail event, we use tools from the theory on large deviations, more specifically Cramér's Theorem to provide a proof for sufficiently large values of n.

Cramér's Theorem. *(Adapted from (Klenke, 2013)) Let X_i be iid real random variables for $i \in \mathbb{N}$ such that*

$$\Lambda(t) := \log \mathbb{E}[e^{tX_1}] < \infty$$

for all $t \in \mathbb{R}$. Define the Legendre transform

$$\Lambda^*(x) := \sup_t (tx - \Lambda(t)).$$

Then for all $z \in \mathbb{R}$ such that $z > \mathbb{E}[X_1]$, we have

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(S_n = \sum_{i=0}^{n} X_i \geq zn\right) = -\Lambda^*(z),$$

where the limit is an upper bound for all n.

We apply the theorem to the random variables $X = -G(M_m(q),p,\epsilon)$ and $X' = -\sum_{i=0}^{m} G_i(q,p,\epsilon)$ at $z = 0$, which is possible, as both $-X$ and $-X'$ have positive expectation, such that $z = 0 > \mathbb{E}[X]$. This yields limits

$$\lim_{n\to\infty} \frac{1}{n} \log\left(\mathbb{P}\left(\sum_i^n G_i(M_m(q),p,\epsilon) \leq 0\right)\right) = -\Lambda_X^*(0)$$

and

$$\lim_{n\to\infty} \frac{1}{n} \log\left(\mathbb{P}\left(\sum_i^{mn} G_i(q,p,\epsilon) \leq 0\right)\right) = -\Lambda_{X'}^*(0)$$

respectively. Because $\mathbb{P}(X \leq 0) = 1 - \mathbb{P}(X > 0)$ for any random variable X, we can conclude that

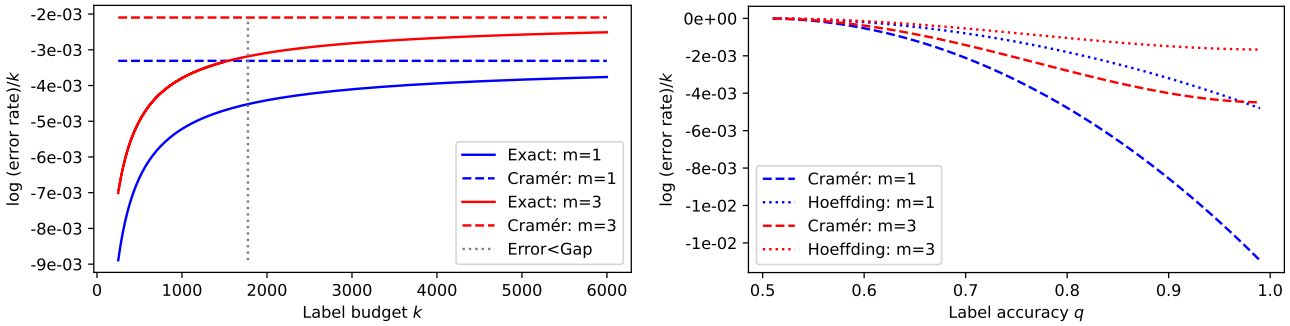
$$\mathbb{P}\left(\sum_i^{mn} G_i(q,p,\epsilon) > 0\right) > \mathbb{P}\left(\sum_i^{n} G_i(M_m(q),p,\epsilon) > 0\right)$$

will be true for sufficiently large n as long as $-\Lambda_X^*(0) > -\Lambda_{X'}^*(0)$. Figure 3 a) illustrates the convergence implied by Cramér's theorem for a fixed set of parameters. As the Cramér rates are upper bounds, we can conclude that the single label approach is better, as soon as the absolute gap between the Cramér rates exceeds the maximum of the approximation errors (here around $k = 1800$). Meanwhile, Figure 3 b) shows the tightness of Cramér's bound compared to Hoeffding's bound. While both are very close when labels are random ($q = 0.5$), Cramér's bound becomes a lot smaller when labels are accurate.

To prove $-\Lambda_X^*(0) > -\Lambda_{X'}^*(0)$, we make use of the simple ternary structure of G and the following lemma characterising $-\Lambda_X^*(0)$ for ternary random variables:

Lemma 2. *For X ternary with $\mathbb{P}(X = 1) = x$, $\mathbb{P}(X = -1) = y$, and $\mathbb{P}(X = 0) = z$,*

$$-\Lambda_X^*(0) = \log\left(2\sqrt{xy} + z\right).$$



(a) Error rates for $q = 0.75$, $p = 0.7$, $\epsilon = 0.1$

(b) Error bounds for $k = 2000$, $p = 0.7$, $\epsilon = 0.1$

Figure 3. a): Convergence of normalized log error rates to the values implied by Cramér's Theorem for label accuracy $q = 0.75$, classifier accuracy $p = 0.7$, margin $\epsilon = 0.1$ and $m \in \{1, 3\}$. b): Upper bounds on normalized log error rate for Cramér's bound compared to Hoeffding's bound.

For sums of independent copies X_i of ternary variables $S_n = \sum_{i=0}^{n} X_i$,

$$-\Lambda^*_{S_n}(0) = n \log\left(2\sqrt{xy} + z\right).$$

Lemma 2 can then be used to prove our main theorem:

Theorem 2. *For G as defined as in Section 3.2, $m > 1$ and $q_b, q_w, p_w, p_b^0, p_b^1$ fixed such that assumption 1 holds, there exist an $N \in \mathbb{N}$ such that for $n > N$*

$$\mathbb{P}\left(\sum_i^n G_i(M_m(q), p) > 0\right) < \mathbb{P}\left(\sum_i^{mn} G_i(q, p) > 0\right).$$

Under assumption 2, this implies that the single label strategy outperforms the $m-$label strategy for these n.

In other words, under the assumptions from Section 3.2 on the joint distribution of the labels and classifiers and sufficiently large label budgets mn, it is always better to collect a single label for mn data points rather than m labels for n data points, when it comes to classifier comparison. We note that the assumptions in Section 3.2 generalize those from Section 2 by taking into account correlations between classifiers and labels, such that Theorem 2 also holds for the independent case discussed in Section 2.

We begin by sketching the proof on a high level: First, we observe that $-m\Lambda^*_X(0) = -\Lambda^*_{X'}(0)$ whenever $q = q_b = q_w = 0.5$, as $M_m(0.5) = 0.5$. As both of these terms have to be negative, this establishes $-\Lambda^*_X(0) > -m\Lambda^*_X(0) = -\Lambda^*_{X'}(0)$. We then show that in the setting of Section 2, the derivative of $-\Lambda^*_X(0) + \Lambda^*_{X'}(0)$ with respect to q is always positive, establishing the independent case. Then, we extended the same argument to the case of correlated classifiers, before decoupling q_b and q_w for the fully correlated case from Section 3.2, setting $q_b = q_w + \delta$ for $\delta \geq 0$ based on assumption 1. Noting that by the previous proof, the theorem is correct for $\delta = 0$, we again establish consistently positive derivatives of $-m\Lambda^*_X(0) = -\Lambda^*_{X'}(0)$, this time with respect to δ. Finally, after establishing $-\Lambda^*_X(0) > -\Lambda^*_{X'}(0)$, and thus the first half of the theorem statement, we use assumption 2 to reduce the case of heterogeneous label accuracies $q(x)$ to the homogeneous case via stochastic dominance.

We continue with additional details for the independent case: Setting

$$d := \epsilon(1 - p - \epsilon)p + ((1 - p - \epsilon)p)^2,$$
$$f^*(q) := M_m(q)(1 - M_m(q))\epsilon^2 + d,$$
$$g^*(q) := q(1 - q)\epsilon^2 + d,$$
$$c := 1 - \epsilon - 2p(1 - p - \epsilon),$$

it is possible to rewrite

$$-\Lambda^*_X(0) + \Lambda^*_{X'}(0)$$
$$= \log\left(2\sqrt{f^*(q)} + c\right) - m\log\left(2\sqrt{g^*(q)} + c\right).$$

such that

$$\frac{d}{dq}(-\Lambda^*_X(0) + \Lambda^*_{X'}(0))$$
$$= \frac{f^{*'}(q)}{\left(2f^*(q) + c\sqrt{f^*(q)}\right)} - m\frac{g^{*'}(q)}{\left(2g^*(q) + c\sqrt{g^*(q)}\right)}.$$

This is positive whenever

$$f^{*'}(q) \geq mg^{*'}(q)\frac{\left(2f^*(q) + c\sqrt{f^*(q)}\right)}{\left(2g^*(q) + c\sqrt{g^*(q)}\right)}. \tag{4}$$

To establish inequality (4), we calculate

$$g^{*'}(x) = \epsilon^2(1 - 2q)$$

and

$$f^{*'}(x) = \epsilon^2(1-2M_m(q))m\binom{m-1}{\frac{m-1}{2}}q^{\frac{m-1}{2}}(1-q)^{\frac{m-1}{2}},$$

where the last equation uses the well known equality of

$$M_{2n+1}(q) = (2n+1)\binom{2n}{n}\int_0^q x^n(1-x)^n dx$$

(Boland et al., 1989). Using various algebraic manipulations to get rid of additive constants, this allows us to reduce (4) to

$$\frac{2M_m(q)-1}{2q-1}\binom{m-1}{\frac{m-1}{2}}q^{\frac{m+1}{2}}(1-q)^{\frac{m+1}{2}} \leq M_m(q)(1-M_m(q)). \quad (5)$$

Using an additive recursion for $M_m(q)$, we then show that both sides of the inequality approach zero for $q \to 1$. We conclude with a series of further algebraic manipulations to establish that the right hand side has a smaller derivative than the left hand side, thus growing faster as q is decreased starting from $q = 1$.

The proof of the general case again centers around equation (4), now interpreted as a function of δ. However, $f^*(\delta)$ and $g^*(\delta)$ become substantially more complicated, as they now involve both $M_m(q_w)$ and $M_m(q_w + \delta)$ terms that need to be treated separately. The proof again makes heavy use of the additive recursion for $M_m(q)$, as well as algebraic manipulations that simplify inequalities of fractions by allowing us to ignore certain terms, eventually reducing equation (4) to equation (5) again. For the sake of brevity, we defer further details to Appendix D.

5. Conclusion

Our results suggest that while collecting multiple labels per instance can be useful for better understanding disagreement about a classification task, collecting a single label per instance is optimal for comparing binary classifiers' accuracy in terms of the annotators' majority label. Thus, while we agree with Aroyo & Welty (2015) that "one [label] is enough" is a myth when it comes to a fine-grained understanding of annotator labels, we find that one label is all you need for simple benchmarking, where a model's performance is *for better or worse* reduced to its test accuracy.

In order to better understand ambiguities in their task definition and how annotators' identity influences their labels (Denton et al., 2021), we still encourage practitioners to initially collect multiple annotations for a small sample of instances when designing a new benchmark based on crowdsourced labels. This understanding can then be used to adjust the task instructions and annotator pool such that the expected annotator label for each instance reflects the intended task as well as possible, and in particular such that $q > 0.5$. Achieving that might require a data-dependent annotator pool, preferentially assigning annotators to instances for which they possess relevant expertise.

Once the task description and annotator pool are fixed and it comes to evaluation at scale, we generally recommend practitioners to build their test set using a large number of instances with a single label each, according to their budget. The only exceptions are if a) estimating the precise risk $\mathcal{R}(c)$ of a classifier c is more important than ranking classifiers, b) the cost of unlabeled data is not negligible, or c) there is good reason to believe that one of our assumptions is violated, i.e. label errors are more common when the better classifier c_b is correct or there is substantially more heterogeneity in $q(x)$ when the worse classifier c_w is correct. In the latter case, using single labels can still often be preferable, and we provide a calculator for the exact probabilities at https://labelnoise.is.tuebingen.mpg.de.

While we do not study the effects of aggregation for datasets that have already been constructed using multiple labels per instance, we would like to reiterate Denton et al. (2021)'s recommendation to "Consider what valuable information might be lost through such aggregation". If such a dataset is, privacy permitting, released with all annotators' labels, users have to choice whether and how to aggregate labels (Prabhakaran et al., 2021). If only majority labels are released, it is impossible for others to obtain information about annotator disagreement, or even simply use a different aggregation method more suited to their needs.

Our work opens up multiple theoretical problems. First, while we consistently observe the single label approach outperform $m > 1$ in our experiments, our main theorem is asymptotic. We conjecture, that this is always true:

Conjecture 1. *For G defined as in Section 3.2 with $m > 1$,*

$$\mathbb{P}\left(\sum_i^n G_i(M_m(q_b), M_m(q_w), p_w, p_b) > 0\right) < \mathbb{P}\left(\sum_i^{mn} G_i(q_b, q_w, p_w, p_b) > 0\right)$$

for all $n > 0$ as long as assumption 1 holds. Under assumption 2, this implies that the single label strategy outperforms the $m-$label strategy.

Proving this conjecture likely requires different methods than employed in the current paper, as Cramér's theorem is not particularly tight for small n.

Second, as our proofs involve a series of non-tight inequalities, assumptions 1 and 2 could likely be further relaxed at the cost of additional complexity.

Third, while binary classification is at the heart of many contemporary human-labeled tasks, most notably reward

modelling for Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022), multiclass classification remains an important task. Extending our results to that setting is a challenging open problem. Solving this likely requires precise modelling of class-conditional error probabilities and results might depend on details of the aggregation procedure: For example, when there are more than three labelers, plurality and absolute majority can diverge, and it is conceivable that plurality voting could extract sufficient additional signal to make collecting multiple labels competitive in some scenarios. Similarly, smarter *adaptive* labeling strategies, like first collecting two labels and only collecting a third in case of a tie, could make collecting multiple labels more competitive in the binary case, but these strategies are harder to implement and analyse.

Acknowledgements

We would like to thank mathoverflow user Kostya_I for pointing out the connection of our problem to large deviation theory. We would also like to thank Rediet Abebe, Amin Charusaie, André Cruz, Mila Gorecki, Vivian Nastl, Olawale Salaudeen, Ana-Andreea Stoica, Sven Wang, and Jiduan Wu for helpful discussions and feedback on draft versions of this work. Florian Dorner is grateful for financial support from the Max Planck ETH Center for Learning Systems (CLS).

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here. That said, we would like to reiterate, that using single labels is only optimal for classifier comparison, once it has been established that each annotator label is more likely to be correct than not. If that is not the case, benchmark results can easily become misleading and in the worst case anti-correlated with actual task performance. As discussed in Section 5, collecting multiple labels for single data points during benchmark conceptualization and analyzing different annotators' disagreements can play an important role for better understanding the task definition, improving annotator instructions, and ensuring a sufficiently diverse and representative annotator pool.

References

Aroyo, L. and Welty, C. Crowd truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. *WebSci2013. ACM*, 2013(2013), 2013.

Aroyo, L. and Welty, C. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1): 15–24, 2015.

Blum, A. and Hardt, M. The ladder: A reliable leaderboard for machine learning competitions. In *International Conference on Machine Learning*, pp. 1006–1014. PMLR, 2015.

Boland, P. J., Proschan, F., and Tong, Y. L. Modelling dependence in simple and indirect majority systems. *Journal of Applied Probability*, 26(1):81–88, 1989.

Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*, 2015.

Chen, D., Yu, Z., and Bowman, S. R. Clean or annotate: How to spend a limited data collection budget. *arXiv preprint arXiv:2110.08355*, 2021.

Cheplygina, V. and Pluim, J. P. Crowd disagreement about medical images is informative. In *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis: 7th Joint International Workshop, CVII-STENT 2018 and Third International Workshop, LABELS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3*, pp. 105–111. Springer, 2018.

Crammer, K., Kearns, M., and Wortman, J. Learning from data of variable quality. *Advances in Neural Information Processing Systems*, 18, 2005.

Davani, A. M., Díaz, M., and Prabhakaran, V. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110, 2022.

Denton, E., Díaz, M., Kivlichan, I., Prabhakaran, V., and Rosen, R. Whose ground truth? accounting for individual and collective identities underlying dataset annotation. *arXiv preprint arXiv:2112.04554*, 2021.

Dolan, B. and Brockett, C. Automatically constructing a corpus of sentential paraphrases. In *Third International Workshop on Paraphrasing (IWP2005)*, 2005.

Dorner, F. E., Peychev, M., Konstantinov, N., Goel, N., Ash, E., and Vechev, M. Human-guided fair classification for natural language processing. *arXiv preprint arXiv:2212.10154*, 2022.

Fleisig, E., Amstutz, A., Atalla, C., Blodgett, S. L., Daumé III, H., Olteanu, A., Sheng, E., Vann, D., and Wallach, H. Fair-prism: Evaluating fairness-related harms in text generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics*, 2023.

Gordon, M. L., Lam, M. S., Park, J. S., Patel, K., Hancock, J., Hashimoto, T., and Bernstein, M. S. Jury learning: Integrating dissenting voices into machine learning models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–19, 2022.

Gray, M. L. and Suri, S. *Ghost work: How to stop Silicon Valley from building a new global underclass*. Eamon Dolan Books, 2019.

Hardt, M. and Recht, B. *Patterns, predictions, and actions: Foundations of machine learning*. Princeton University Press, 2022.

Jigsaw. Jigsaw unintended bias in toxicity classification. https://www.kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification/data, 2019. Accessed: 2024-01-24.

Klenke, A. *Probability theory: a comprehensive course*. Springer Science & Business Media, 2013.

Köpf, A., Kilcher, Y., von Rütte, D., Anagnostidis, S., Tam, Z.-R., Stevens, K., Barhoum, A., Duc, N. M., Stanley, O., Nagyfi, R., et al. Openassistant conversations–democratizing large language model alignment. *arXiv preprint arXiv:2304.07327*, 2023.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images, 2009.

Lin, C., Weld, D., et al. To re (label), or not to re (label). In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 2, pp. 151–158, 2014.

Mania, H. and Sra, S. Why do classifier accuracies show linear trends under distribution shift? *arXiv preprint arXiv:2012.15483*, 2020.

Marelli, M., Bentivogli, L., Baroni, M., Bernardi, R., Menini, S., and Zamparelli, R. Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pp. 1–8, 2014.

Nguyen, H. Q., Lam, K., Le, L. T., Pham, H. H., Tran, D. Q., Nguyen, D. B., Le, D. D., Pham, C. M., Tong, H. T., Dinh, D. H., et al. Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations. *Scientific Data*, 9(1):429, 2022.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Prabhakaran, V., Davani, A. M., and Diaz, M. On releasing annotator-level labels and information in datasets. *arXiv preprint arXiv:2110.05699*, 2021.

Ramponi, A. and Leonardelli, E. Dh-fbk at semeval-2022 task 4: leveraging annotators' disagreement and multiple data views for patronizing language detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pp. 324–334, 2022.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pp. 5389–5400. PMLR, 2019.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115: 211–252, 2015.

Sandri, M., Leonardelli, E., Tonelli, S., and Ježek, E. Why don't you do it right? analysing annotators' disagreement in subjective tasks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2420–2433, 2023.

Sheng, V. S., Provost, F., and Ipeirotis, P. G. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 614–622, 2008.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., and Potts, C. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1631–1642, 2013.

Stanica, P. Good lower and upper bounds on binomial coefficients. *Journal of Inequalities in Pure and Applied Mathematics*, 2(3):30, 2001.

Tanno, R., Saeedi, A., Sankaranarayanan, S., Alexander, D. C., and Silberman, N. Learning from noisy labels by regularized estimation of annotator confusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11244–11253, 2019.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Wei, J., Zhu, Z., Luo, T., Amid, E., Kumar, A., and Liu, Y. To aggregate or not? learning with separate noisy labels. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2523–2535, 2023.

Williams, A., Nangia, N., and Bowman, S. R. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*, 2017.

A. Numerical Evidence

We conducted a large scale parameter sweep for

$$n \in [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 100, 101, 1000, 1001],$$

$$m \in [3, 11]$$

and

$$q_b, q_w, p_w, p_b^0, p_b^1 \in S^5,$$

where S is a set of 50 evenly spaced points $s \in [0.5, 1]$ with a resolution of 0.01. For all of the almost five billon grid points that fulfilled $(1 - p_w)p_b^0 + p_w p_b^1 > p_w$, we both explicitly calculated

$$\mathbb{P}\left(\sum_{i=0}^{mn} G_i(q, p, \epsilon) > 0\right)$$

and

$$\mathbb{P}\left(\sum_{i=0}^{n} G_i(M_m(q), p, \epsilon) > 0\right)$$

(using iterated convolutions of the base variable G, sped up via exponentiation by squaring) and additionally approximated the probabilities based on sampling each of the sums 100 times. Under the assumptions from section 3.2, the exact calculations consistently yielded

$$\mathbb{P}\left(\sum_{i=0}^{mn} G_i(q, p, \epsilon) > 0\right) \geq \mathbb{P}\left(\sum_{i=0}^{n} G_i(M_m(q), p, \epsilon) > 0\right),$$

with the only exceptions happening when both probabilities are extremely close to 1 (maximal distance of the order $1e-12$). These exceptions do not provide meaningful evidence against our conjecture, as they are most likely caused by numerical instability (notably, they often coincide with calculated probabilities that exceed one). In particular, there were no parameters for which both the exact probabilities and the sampled probabilities were better for the $m-$label case, even though this happened for the sampled probabilities alone in 1.6% of the cases (as to be expected from the relatively small sample size of 100). As an additional sanity check, the sampled probabilities generally approximated the exact probabilities well, with the average distance over all parameters being on the order of $1e-7$, and the average MSE of the order 0.01 for both the single and the $m-$label case.

Notably, the single label approach still performed better in two thirds of the parameter configurations with $q_w > q_b$, with this number slowly decreasing for larger values of q_w. This suggests that our (already not particularly restrictive) assumptions could be relaxed substantially further.

B. Parameterizations of the Gap Indicator

Proposition 1. *Assuming mututally independent classifier and labeler errors, G can be written as follows:*

$$G(q, p, \epsilon) = \begin{cases} 1 & \textit{w.p. } q\epsilon + (1-p-\epsilon)p \\ -1 & \textit{w.p. } (1-q)\epsilon + (1-p-\epsilon)p \\ 0 & \textit{else } p(p+\epsilon) + (1-p-\epsilon)(1-p) \end{cases},$$

for label accuracy q, classifier accuracy p and margin ϵ.

Proof. The better classifier c_b wins for a given x (i.e. $G = 1$) if $c_b(x)$ and the label $y_{Test}(x)$ are correct, while $c_w(x)$ is not, or if both $c_b(x)$ and the label $y_{Test}(x)$ are incorrect, while $c_w(x)$ is correct. The former happens with probability $((p+\epsilon)(1-p)q)$, and the latter with probability $((p)(1-p-\epsilon)(1-q))$. Summing up yields

$$\begin{aligned} \mathbb{P}(G=1) =& ((p+\epsilon)(1-p)q) + (p)(1-p-\epsilon)(1-q) \\ &= qp - qp^2 + q\epsilon - qp\epsilon + (1-q)\big(p - p^2 - p\epsilon\big) \\ &= qp - qp^2 + q\epsilon - qp\epsilon + p - p^2 - p\epsilon - qp + qp^2 + qp\epsilon \\ &= q\epsilon + p - p^2 - p\epsilon = q\epsilon + p(1-p-\epsilon). \end{aligned}$$

For the worse classifier c_w to win ($G = -1$), we get the opposite cases conditional on the label, with respective probabilities of $((p+\epsilon)(1-p)(1-q))$ and $((p)(1-p-\epsilon)q$. These sum up as follows:

$$\begin{aligned}\mathbb{P}(G=-1) &= ((p+\epsilon)(1-p)(1-q)) + (p)(1-p-\epsilon)q \\ &= \big(p+\epsilon-p^2-p\epsilon\big)(1-q) + qp - qp^2 - qp\epsilon \\ &= p+\epsilon-p^2-p\epsilon-qp-q\epsilon+qp^2+qp\epsilon+qp-qp^2-qp\epsilon \\ &= p+\epsilon-p^2-p\epsilon-q\epsilon \\ &= (1-p-\epsilon)p + (1-q)\epsilon.\end{aligned}$$

Adding up both probabilities yields

$$\begin{aligned}\mathbb{P}(G\neq 0) &= 2p(1-p-\epsilon)+\epsilon \\ &= 2p-2p^2-2p\epsilon+\epsilon \\ &= 1-p(p+\epsilon)+2p-p^2-p\epsilon+\epsilon-1 \\ &= 1-p(p+\epsilon)-(1-p-\epsilon)(1-p),\end{aligned}$$

which makes sense as the gap indicator $G(p,q,\epsilon)$ is zero whenever both classifiers produce the same answer, independent of the label. □

Proposition 2. *Assuming correlated classifiers and labels with the above parameterization, we have:*

$$G(q,p) = \begin{cases} 1 & \textit{w.p. } q_b(1-p_w)p_b^0 \\ & +(1-q_w)p_w\big(1-p_b^1\big) \\ -1 & \textit{w.p. } (1-q_b)(1-p_w)p_b^0 \\ & +q_w p_w\big(1-p_b^1\big) \\ 0 & \textit{else} \end{cases}.$$

Proof. The better classifier c_b "wins" on a given datapoint, whenever it and the label are correct, while the worse classifier is not, or if the label and the better classifier are incorrect, while the worse classifier is correct. The former happens with probability $q_b(1-p_w)p_b^0$ and the latter with probability $(1-q_w)p_w\big(1-p_b^1\big)$. The case of the worse classifier winning is symmetric, with q_i and $1-q_i$ reversed. This yields

$$G\big(q_b,q_w,p_w,p_b^0,p_b^1\big) = \begin{cases} 1 & \text{w.p. } q_b(1-p_w)p_b^0 + (1-q_w)p_w\big(1-p_b^1\big) \\ -1 & \text{w.p. } (1-q_b)(1-p_w)p_b^0 + q_w p_w\big(1-p_b^1\big) \\ 0 & \text{else} \end{cases}.$$

□

C. Details on Hoeffding Bounds

We first establish, that it is sufficient to focus on the case of m uneven, as going from m uneven to $m+1$ even reduces the number of data points we can label n, while *reducing* $M_m(q)$ due to additional ties, rather than increasing it:

Lemma C.1. *For even $k>1$, we have that*

$$M_k(q) < M_{k-1}(q).$$

Proof. For even k, a majority can only be obtained, if there is already a majority for the first $k-1$ votes. In that case, the majority is always retained, unless the margin was exactly one, and the new vote goes against the majority. For our case, this means that

$$M_k(q) = M_{k-1}(q) - \binom{k-1}{\frac{k}{2}} q^{\frac{k}{2}}(1-q)^{\frac{k}{2}} < M_{k-1}(q).$$

□

We proceed by proving lemma 1:

Lemma 1. *For independent copies X_i of any random variable X with $\mathbb{E}[X] > 0$ and values in $[-1, 1]$, we can bound*

$$\mathbb{P}\left(\sum_{i=0}^{n} X_i \leq 0\right) \leq e^{\frac{-n\,\mathbb{E}[X]^2}{2}} =: B(X, n).$$

Proof. For independent copies X_i of any random variable X with values in $[-1, 1]$, we have

$$\sum_{i=0}^{n} X_i \leq 0 \iff \sum_{i=0}^{n}(X_i - \mathbb{E}[X]) \leq -n\,\mathbb{E}[X] \iff \sum_{i=0}^{n}(-X_i - \mathbb{E}[-X]) \geq n\,\mathbb{E}[X].$$

If $\mathbb{E}[X] > 0$, we can then apply Hoeffding's inequality to $-X$ to obtain

$$\mathbb{P}\left(\sum_{i=0}^{n} X_i \leq 0\right) = \mathbb{P}\left(\sum_{i=0}^{n}(-X_i - \mathbb{E}[-X]) \geq n\,\mathbb{E}[X]\right) \leq e^{\frac{-2n^2\,\mathbb{E}[X]^2}{4n}} = e^{\frac{-n\,\mathbb{E}[X]^2}{2}}.$$

□

With this, we focus on

Proposition C.1. *For any uneven $m > 1$, equation (2) is true, i.e.*

$$\sqrt{m} > \frac{2M_m(q) - 1}{2q - 1}.$$

Correspondingly, for any $m > 1$

$$B(G(q, p, \epsilon), nm) < B(G(M_m(q), p, \epsilon), n),$$

where B is the Hoeffding lower bound on the success probability.

Proof. To prove proposition C.1, we need the following lemma:

Lemma C.2. *Setting $\sigma(m, q) = \sum_{k\ uneven}^{m-2} \binom{k}{\lceil \frac{k}{2} \rceil} q^{\lceil \frac{k}{2} \rceil}(1-q)^{\lceil \frac{k}{2} \rceil}$ for uneven m, we have that*

$$M_m(q) = q + (2q - 1)\sigma(m, q).$$

Proof. Let $b_n(q, k)$ be the probability of k successes in a binomial with n trials and with probability of success p for a single trial. Then:

$$\begin{aligned}
M_m(q) &= M_{m-2}(q) + q^2 b_{m-2}\left(q, \lfloor \frac{m-2}{2} \rfloor\right) - (1-q)^2 b_{m-2}\left(q, \lceil \frac{m-2}{2} \rceil\right) \\
&= M_{m-2}(q) + q^2 \frac{1-q}{q} b_{m-2}\left(q, \lceil \frac{m-2}{2} \rceil\right) - (1-q)^2 b_{m-2}\left(q, \lceil \frac{m-2}{2} \rceil\right) \\
&= M_{m-2}(q) + \left(q - q^2 - 1 + 2q - q^2\right) b_{m-2}\left(q, \lceil \frac{m-2}{2} \rceil\right) \\
&= M_{m-2}(q) + \left(3q - 2q^2 - 1\right) b_{m-2}\left(q, \lceil \frac{m-2}{2} \rceil\right) \\
&= M_{m-2}(q) + (1-q)(2q-1) b_{m-2}\left(q, \lceil \frac{m-2}{2} \rceil\right) \\
&= M_{m-2}(q) + (1-q)(2q-1)\binom{m-2}{\lceil \frac{m-2}{2} \rceil} q^{\lceil \frac{m-2}{2} \rceil}(1-q)^{\lceil \frac{m-2}{2} \rceil - 1} \\
&= M_{m-2}(q) + (2q-1)\binom{m-2}{\lceil \frac{m-2}{2} \rceil} q^{\lceil \frac{m-2}{2} \rceil}(1-q)^{\lceil \frac{m-2}{2} \rceil}.
\end{aligned}$$

The first equation captures the fact that a majority of m trials consists of all events that have a majority for the first $m-2$ trials (first term), except for those with a margin of one that simultaneously have two misses in the last two trials (third term), in addition to all events that miss a majority in the first $m-2$ trials by a margin of one, but have two successes in the last two trials (second term). The statement of the Lemma then follows by unrolling the additive recursion. □

With Lemma C.2, (2) can be rewritten as

$$\sqrt{m} > \frac{2M_m(q)-1}{2q-1} = \frac{2q+2(2q-1)\sigma(m,q)-1}{2q-1} = 1+2\sigma(m,q). \tag{6}$$

We can control the right term using another Lemma:

Lemma C.3.

$$1+2\sigma(m,q) \leq 1+\frac{1}{\sqrt{\pi}}(2\sqrt{\frac{m-1}{2}}-1)$$

Proof. We use an upper bound version of Stirling's approximation based on Theorem 2.6 in (Stanica, 2001):

$$\binom{m-1}{\frac{m-1}{2}} < \frac{4^{\frac{m-1}{2}}}{\sqrt{\pi\frac{m-1}{2}}},$$

the fact that $q(1-q)$ is maximized at $q=0.5$ and the monontonicity of $\frac{1}{\sqrt{k}}$ to estimate

$$\begin{aligned}
2\sigma(m,q) &= 2\sum_{k \text{ uneven}}^{m-2}\binom{k}{\lceil\frac{k}{2}\rceil}q^{\lceil\frac{k}{2}\rceil}(1-q)^{\lceil\frac{k}{2}\rceil} && = 2\sum_{k \text{ uneven}}^{m-2}\binom{k}{\frac{k+1}{2}}q^{\frac{k+1}{2}}(1-q)^{\frac{k+1}{2}} \\
&= 2\sum_{k \text{ uneven}}^{m-2}\frac{\frac{k+1}{2}}{k+1}\binom{k+1}{\frac{k+1}{2}}q^{\frac{k+1}{2}}(1-q)^{\frac{k+1}{2}} && = \sum_{k \text{ uneven}}^{m-2}\binom{k+1}{\frac{k+1}{2}}q^{\frac{k+1}{2}}(1-q)^{\frac{k+1}{2}} \\
&\leq \sum_{k \text{ uneven}}^{m-2}\frac{4^{\frac{k+1}{2}}}{\sqrt{\pi\frac{k+1}{2}}}q^{\frac{k+1}{2}}(1-q)^{\frac{k+1}{2}} && \leq \sum_{k \text{ uneven}}^{m-2}\frac{1}{\sqrt{\pi\frac{k+1}{2}}} \\
&= \frac{1}{\sqrt{\pi}}\sum_{k>0 \text{ even}}^{m-1}\frac{1}{\sqrt{\frac{k}{2}}} && = \frac{1}{\sqrt{\pi}}\sum_{k=1}^{\frac{m-1}{2}}\frac{1}{\sqrt{k}} \\
&= \frac{1}{\sqrt{\pi}}\left(1+\sum_{k=2}^{\frac{m-1}{2}}\frac{1}{\sqrt{k}}\right) && \leq \frac{1}{\sqrt{\pi}}\left(1+\int_{k=1}^{\frac{m-1}{2}}\frac{1}{\sqrt{k}}\right) \\
&= \frac{1}{\sqrt{\pi}}\left(1+2\sqrt{\frac{m-1}{2}}-2\right) && = \frac{1}{\sqrt{\pi}}\left(2\sqrt{\frac{m-1}{2}}-1\right) \quad .
\end{aligned}$$

□

With this, (2) reduces to

$$1-\frac{1}{\sqrt{\pi}}+\sqrt{\frac{2}{\pi}(m-1)} = 1+\frac{1}{\sqrt{\pi}}\left(2\sqrt{\frac{m-1}{2}}-1\right) < \sqrt{m}.$$

At $m=3$, this becomes

$$1.56 \approx 1+\frac{1}{\sqrt{\pi}} < \sqrt{3} \approx 1.73.$$

On the other hand the derivative of the gap with respect to m,

$$\frac{d}{dm}\left(\sqrt{m}-1+\frac{1}{\sqrt{\pi}}-\sqrt{\frac{2}{\pi}(m-1)}\right)=\frac{1}{2\sqrt{m}}-\frac{1}{\sqrt{2\pi}\sqrt{m-1}}$$

is positive whenever

$$\frac{1}{2\sqrt{m}}>\frac{1}{\sqrt{2\pi}\sqrt{m-1}}$$

or

$$1.25\approx\frac{\sqrt{2\pi}}{2}>\sqrt{\frac{m}{m-1}}=\sqrt{1+\frac{1}{m-1}}.$$

For $m\geq 3$, the right side is clearly at most $\sqrt{1+\frac{1}{3-1}}\approx 1.22$, such that the derivative is positive for all $m>3$ and (2) holds for $m\geq 3$. □

Next, we focus on the general case with correlated classifiers and labels:

Proposition C.2. *When classifiers and labels are correlated, as long as $q_b\geq q_w$ and $(1-p_w)p_b^0+p_wp_b^1>p_w$,*

$$B(G(q,p),nm)<B(G(M_m(q),p),n)$$

holds for any $m>1$, where B is the Hoeffding lower bound on the success probability.

Proof. In this setting, Equation (1) becomes

$$\sqrt{m}>\frac{(2M_m(q_b)-1)(1-p_w)p_b^0-(2M_m(q_w)-1)p_w\big(1-p_b^1\big)}{(2q_b-1)(1-p_w)p_b^0-(2q_w-1)p_w(1-p_b^1)}. \tag{7}$$

To prove this, we need the following Lemma:

Lemma C.4. *Let A,B,C,D,c_1,c_2 be positive constants such that $Ac_1-Bc_2>0$ and $Cc_1-Dc_2>0$. Then $\frac{Ac_1-Bc_2}{Cc_1-Dc_2}\leq\frac{A}{C}$ is true if and only if $CB\geq DA$.*

Proof.

$$\begin{aligned}\frac{Ac_1-Bc_2}{Cc_1-Dc_2}\leq\frac{A}{C}&\iff Ac_1-Bc_2\leq\frac{A(Cc_1-Dc_2)}{C}\\&\iff C(Ac_1-Bc_2)\leq A(Cc_1-Dc_2)\\&\iff CAc_1-CBc_2\leq CAc_1-DAc_2\\&\iff -CBc_2\leq -DAc_2\\&\iff CB\geq DA\end{aligned}$$

□

With this, we set

$$\begin{aligned}A&=2M_m(q_b)-1,\\B&=2M_m(q_w)-1,\\C&=2q_b-1,\\D&=2q_w-1,\end{aligned}$$

and

$$\begin{aligned}c_1&=(1-p_w)p_b^0,\\c_2&=p_w\big(1-p_b^1\big).\end{aligned}$$

Then, $CB \geq DA$ is equivalent to

$$(2q_b - 1)(2M_m(q_w) - 1) \geq (2q_w - 1)(2M_m(q_b) - 1),$$

i.e.

$$\frac{2M_m(q_w) - 1}{2q_w - 1} \geq \frac{2M_m(q_b) - 1}{2q_b - 1},$$

which is equivalent to

$$1 + 2\sigma(m, q_w) \geq 1 + 2\sigma(m, q_b),$$

and holds for $q_b \geq q_w$ as $\sigma(m, x)$ is clearly monotonically decreasing in x. Lemma C.4 combined with Equation (6) thus allows us to upper bound

$$\frac{(2M_m(q_b) - 1)(1 - p_w)p_b^0 - (2M_m(q_w) - 1)p_w(1 - p_b^1)}{(2q_b - 1)(1 - p_w)p_b^0 - (2q_w - 1)p_w(1 - p_b^1)} \leq \frac{2M_m(q_b) - 1}{2q_b - 1} \leq \sqrt{m},$$

proving the proposition. □

D. Proving Theorem 2

Theorem 2. *For G as defined as in Section 3.2, $m > 1$ and $q_b, q_w, p_w, p_b^0, p_b^1$ fixed such that assumption 1 holds, there exist an $N \in \mathbb{N}$ such that for $n > N$*

$$\mathbb{P}\left(\sum_i^n G_i(M_m(q), p) > 0\right) < \mathbb{P}\left(\sum_i^{mn} G_i(q, p) > 0\right).$$

Under assumption 2, this implies that the single label strategy outperforms the $m-$label strategy for these n.

The proof of theorem 2 is based on Cramér's Theorem:

Cramér's Theorem. *(Adapted from (Klenke, 2013)) Let X_i be iid real random variables for $i \in \mathbb{N}$ such that*

$$\Lambda(t) := \log \mathbb{E}[e^{tX_1}] < \infty$$

for all $t \in \mathbb{R}$. Define the Legendre transform

$$\Lambda^*(x) := \sup_t (tx - \Lambda(t)).$$

Then for all $z \in \mathbb{R}$ such that $z > \mathbb{E}[X_1]$, we have

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(S_n = \sum_{i=0}^n X_i \geq zn\right) = -\Lambda^*(z),$$

where the limit is an upper bound for all n.

This means that $\mathbb{P}(S_n = \sum_{i=0}^n X_i \geq zn)$ is eventually roughly of the order $e^{-n\Lambda^*(z)}$. Furthermore, a glance at the prove of Cramér's Theorem, reveals that this exponential is actually an upper bound for the error probability independent of n in our case of $z = 0$. We want to eventually apply the theorem to $X = -G(M_m(q), p, \epsilon)$ and $X' = -\sum_{i=0}^m G_i(q, p, \epsilon)$ respectively. Because these random variables have negative expectation, the theorem can be applied to $z = 0 > \mathbb{E}[X]$, yielding limits for

$$\frac{1}{n} \log \mathbb{P}(S_n \geq 0) := \frac{1}{n} \log\left(\mathbb{P}\left(\sum_i^n G_i(M_m(q), p, \epsilon) \leq 0\right)\right) = \frac{1}{n} \log\left(1 - \mathbb{P}\left(\sum_i^n G_i(M_m(q), p, \epsilon) > 0\right)\right)$$

and

$$\frac{1}{n} \log \mathbb{P}(S'_n \geq 0) := \frac{1}{n} \log\left(1 - \mathbb{P}\left(\sum_i^{mn} G_i(q, p, \epsilon) > 0\right)\right).$$

If we can prove that

$$-\Lambda_X^*(0) > -\Lambda_{X'}^*(0), \tag{8}$$

it follows that there is an $N \in \mathbb{N}$ such that for $n > N$ we have

$$\frac{1}{n}\log\left(1-\mathbb{P}\left(\sum_i^n G_i(M_m(q),p,\epsilon)>0\right)\right) > \frac{1}{n}\log\left(1-\mathbb{P}\left(\sum_i^{mn} G_i(q,p,\epsilon)>0\right)\right)$$

and thus by monotonicity

$$\mathbb{P}\left(\sum_i^n G_i(M_m(q),p,\epsilon)>0\right) < \mathbb{P}\left(\sum_i^{mn} G_i(q,p,\epsilon)>0\right).$$

We first consider a general ternary X with negative expectation:

$$X = \begin{cases} 1 & \text{w.p. } x \\ -1 & \text{w.p. } y \\ 0 & \text{w.p. } z \end{cases}$$

for $y > x$.

Lemma 2. *For X ternary with $\mathbb{P}(X=1)=x$, $\mathbb{P}(X=-1)=y$, and $\mathbb{P}(X=0)=z$,*

$$-\Lambda_X^*(0) = \log\left(2\sqrt{xy}+z\right).$$

For sums of independent copies X_i of ternary variables $S_n = \sum_{i=0}^n X_i$,

$$-\Lambda_{S_n}^*(0) = n\log\left(2\sqrt{xy}+z\right).$$

Proof. We have that

$$-\Lambda_X^*(0) = -\left(\sup_t 0\cdot t - \Lambda_X(t)\right) = \inf_t \Lambda_X(t)$$

Here,

$$\Lambda_X(t) = \log\mathbb{E}[e^{tX}] = \log\big(xe^t + ye^{-t} + z\big).$$

Differentiating yields

$$\frac{d}{dt}\Lambda_X(t) = \frac{xe^t - ye^{-t}}{xe^t + ye^{-t} + z}.$$

The numerator is positive for large positive t and negative for large negative t, with a unique zero at $xe^t = ye^{-t}$, i.e. $\frac{y}{x} = e^{2t}$ or $t = 0.5\log\frac{y}{x}$, such that $\Lambda(t)$ is minimized at this t. This means that

$$\begin{aligned}
\inf_t \Lambda_X(t) &= \log\Big(xe^{0.5\log\frac{y}{x}} + ye^{-0.5\log\frac{y}{x}} + z\Big) \\
&= \log\left(x\sqrt{e^{\log\frac{y}{x}}} + y\frac{1}{\sqrt{e^{\log\frac{y}{x}}}} + z\right) \\
&= \log\left(x\sqrt{\frac{y}{x}} + y\sqrt{\frac{x}{y}} + z\right) \\
&= \log(2\sqrt{xy}+z).
\end{aligned}$$

Now for S_n, we get

$$\Lambda_{S_n}(t) = \log\mathbb{E}[e^{tSn}] = \log\prod_{i=0}^n \mathbb{E}[e^{tX_i}] = n\log\big(xe^t + ye^{-t} + z\big).$$

The optimization is not affected by multiplying by n, so we get

$$\inf_t \Lambda_{S_n}(t) = n\log(2\sqrt{xy}+z)$$

□

D.1. Independent Classifiers

We first focus on the independent case with

$$G(q,p,\epsilon) = \begin{cases} 1 & \text{w.p. } q\epsilon + (1-p-\epsilon)p \\ -1 & \text{w.p. } (1-q)\epsilon + (1-p-\epsilon)p \\ 0 & \text{else } p(p+\epsilon) + (1-p-\epsilon)(1-p) \end{cases},$$

where $q_b = q_w = q$, $p_w = p$ and $p_b^0 = p_b^1 = p + \epsilon$ as defined in section 2 and apply Lemma 2 to

$$X = -G(M_m(q), p, \epsilon)$$

and

$$X' = -\sum_{i=0}^{m} G_i(q,p,\epsilon)$$

to obtain

$$-\Lambda_X^*(0) = \log\left(2\sqrt{M_m(q)(1-M_m(q))\epsilon^2 + \epsilon(1-p-\epsilon)p + ((1-p-\epsilon)p)^2} + 1 - \epsilon - 2p(1-p-\epsilon)\right).$$

and

$$-\Lambda_{X'}^*(0) = m\log\left(2\sqrt{q(1-q)\epsilon^2 + \epsilon(1-p-\epsilon)p + ((1-p-\epsilon)p)^2} + 1 - \epsilon - 2p(1-p-\epsilon)\right).$$

To get some intuition, we fix $p = 0.5$, such that

$$-\Lambda_X^*(0) = \log\left(2\sqrt{\left(M_m(q)(1-M_m(q)) - \frac{1}{4}\right)\epsilon^2 + \frac{1}{16}} + \frac{1}{2}\right),$$

which for the aggregated case yields an asymptotic error rate of

$$e^{-\Lambda_X^*(0)n} = \left(2\sqrt{\left(M_m(q)(1-M_m(q)) - \frac{1}{4}\right)\epsilon^2 + \frac{1}{16}} + \frac{1}{2}\right)^n$$

The error rate has a second order taylor expansion around $\epsilon = 0$ of

$$e^{-n\Lambda_X^*(0)} \approx 1 + (4(M_m(q)(1-M_m(q)) - 1))n\epsilon^2.$$

For $q = M_m(q) = 1$, we thus get

$$e^{-n\Lambda_X^*(0)} \approx 1 - n\epsilon^2,$$

which is consistent with the statistical intuition that $n \gg \frac{1}{\epsilon^2}$ samples are needed to detect a coin with a bias of order ϵ. Meanwhile as q goes to 0.5, $M_m(q)(1-M_m(q))$ approaches 4 and the amount of required samples explodes.

Back to general $p \geq 0.5$, we note that by the AM-GM inequality, $2\sqrt{xy} + z \leq x + y + z = 1$ for any x, y, z that describe a ternary random variable as above with equality only if $x = y$, which cannot happen for G under our assumptions because of its positive expectation. This means that the logarithms in the Λ^* are always strictly negative. In particular, at $q = 0.5$ and $q = 1$, $M_m(q) = q$ such that the terms in the logarithm are equal and we get $-\Lambda_X^*(0) \geq -m\Lambda_X^*(0) = -\Lambda_{X'}^*(0)$. In general, we have

$$\begin{aligned} &-\Lambda_X^*(0) + \Lambda_{X'}^*(0) \\ &= \log\left(2\sqrt{M_m(q)(1-M_m(q))\epsilon^2 + \epsilon(1-p-\epsilon)p + ((1-p-\epsilon)p)^2} + 1 - \epsilon - 2p(1-p-\epsilon)\right) \\ &- m\log\left(2\sqrt{q(1-q)\epsilon^2 + \epsilon(1-p-\epsilon)p + ((1-p-\epsilon)p)^2} + 1 - \epsilon - 2p(1-p-\epsilon)\right). \end{aligned} \tag{9}$$

Because (8) holds for $q = 0.5$ independent of ϵ and p as both terms are the same except for the factor m in that case, it is sufficient to show that (9) always has a positive derivative in q. To show this, we set

$$f^*(q) = M_m(q)(1 - M_m(q))\epsilon^2 + \epsilon(1 - p - \epsilon)p + ((1 - p - \epsilon)p)^2,$$
$$g^*(q) = q(1 - q)\epsilon^2 + \epsilon(1 - p - \epsilon)p + ((1 - p - \epsilon)p)^2$$

and

$$c = 1 - \epsilon - 2p(1 - p - \epsilon),$$

such that

$$-\Lambda^*_X(0) + \Lambda^*_{X'}(0) = \log\Big(2\sqrt{f^*(q)} + c\Big) - m\log\Big(2\sqrt{g^*(q)} + c\Big). \tag{10}$$

Differentiating yields

$$\begin{aligned}
\frac{d}{dq}(-\Lambda^*_X(0) + \Lambda^*_{X'}(0)) &= \frac{d}{dq}\log\Big(2\sqrt{f^*(q)} + c\Big) - m\frac{d}{dq}\log\Big(2\sqrt{g^*(q)} + c\Big) \\
&= \frac{\frac{d}{dq}2\sqrt{f^*(q)}}{\Big(2\sqrt{f^*(q)} + c\Big)} - m\frac{\frac{d}{dq}2\sqrt{g^*(q)}}{\Big(2\sqrt{g^*(q)} + c\Big)} \\
&= \frac{\frac{f^{*'}(q)}{\sqrt{f^*(q)}}}{\Big(2\sqrt{f^*(q)} + c\Big)} - m\frac{\frac{g^{*'}(q)}{\sqrt{g^*(q)}}}{\Big(2\sqrt{g^*(q)} + c\Big)} \\
&= \frac{f^{*'}(q)}{\sqrt{f^*(q)}\Big(2\sqrt{f^*(q)} + c\Big)} \\
&\quad - m\frac{g^{*'}(q)}{\sqrt{g^*(q)}\Big(2\sqrt{g^*(q)} + c\Big)}.
\end{aligned}$$

Correspondingly using (9), (8) reduces to

$$f^{*'}(q) \geq mg^{*'}(q)\frac{\sqrt{f^*(q)}\Big(2\sqrt{f^*(q)} + c\Big)}{\sqrt{g^*(q)}\Big(2\sqrt{g^*(q)} + c\Big)} \tag{11}$$

We can calculate

$$g^{*'}(x) = \frac{d}{dq}q(1 - q)\epsilon^2 = \epsilon^2(1 - 2q)$$

and

$$\begin{aligned}
f^{*'}(x) &= \epsilon^2\frac{d}{dq}M_m(q)(1 - M_m(q)) \\
&= \epsilon^2(1 - 2M_m(q))\frac{d}{dq}M_m(q) \\
&= \epsilon^2(1 - 2M_m(q))m\binom{m-1}{\frac{m-1}{2}}q^{\frac{m-1}{2}}(1 - q)^{\frac{m-1}{2}},
\end{aligned}$$

where the last equation uses the equality of

$$M_{2n+1}(q) = (2n + 1)\binom{2n}{n}\int_0^q x^n(1 - x)^n dx$$

(Boland et al., 1989). Correspondingly, (11) holds if and only if

$$(1 - 2M_m(q))m\binom{m-1}{\frac{m-1}{2}}q^{\frac{m-1}{2}}(1 - q)^{\frac{m-1}{2}} \geq m(1 - 2q)\frac{\sqrt{f^*(q)}\Big(2\sqrt{f^*(q)} + c\Big)}{\sqrt{g^*(q)}\Big(2\sqrt{g^*(q)} + c\Big)}.$$

For $0.5 < q < 1$, this is equivalent to

$$\frac{2M_m(q)-1}{2q-1}\binom{m-1}{\frac{m-1}{2}}q^{\frac{m-1}{2}}(1-q)^{\frac{m-1}{2}} \leq \frac{\sqrt{f^*(q)}\left(2\sqrt{f^*(q)}+c\right)}{\sqrt{g^*(q)}\left(2\sqrt{g^*(q)}+c\right)}. \tag{12}$$

Lemma D.1. *Let $0 < x < y$ and $c > 0$. Then, $\frac{2x+c\sqrt{x}}{2y+c\sqrt{y}} \geq \frac{x}{y}$*

Proof.

$$\begin{aligned}
\frac{2x+c\sqrt{x}}{2y+c\sqrt{y}} \geq \frac{x}{y} &\iff \left(2x+\sqrt{x}c\right)y \geq (2y+\sqrt{y}c)x \\
&\iff 2xy + c\sqrt{x}y \geq 2xy + c\sqrt{y}x \\
&\iff \sqrt{x}y \geq \sqrt{y}x \\
&\iff \frac{y}{\sqrt{y}} \geq \frac{x}{\sqrt{x}} \\
&\iff \sqrt{y} \geq \sqrt{x} \\
&\iff y \geq x
\end{aligned}$$

□

As $f^*(q)$ and $g^*(q)$ can be written as $M_m(q)(1-M_m(q))\epsilon^2 + d$ and $q(1-q)\epsilon^2 + d$ respectively for $d = \epsilon(1-p-\epsilon)p + ((1-p-\epsilon)p)^2$, and because $k(x) = x(1-x)$ is monontonously falling in x while $M_m(x)$ grows in m, $g^*(x) \geq f^*(x)$, and Lemma D.1 implies that it is sufficient to show

$$\frac{2M_m(q)-1}{2q-1}\binom{m-1}{\frac{m-1}{2}}q^{\frac{m-1}{2}}(1-q)^{\frac{m-1}{2}} \leq \frac{f^*(q)}{g^*(q)}. \tag{13}$$

Lemma D.2. *Let $0 < x < y$ and $d > 0$. Then, $\frac{x+d}{y+d} \geq \frac{x}{y}$*

Proof.

$$\begin{aligned}
\frac{x+d}{y+d} \geq \frac{x}{y} &\iff y(x+d) \geq x(y+d) \\
&\iff xy + yd \geq yx + xd \\
&\iff y \geq x
\end{aligned}$$

□

Lemma D.2 implies that (13) can be reduced to

$$\frac{2M_m(q)-1}{2q-1}\binom{m-1}{\frac{m-1}{2}}q^{\frac{m-1}{2}}(1-q)^{\frac{m-1}{2}} \leq \frac{\epsilon^2 M_m(q)(1-M_m(q))}{\epsilon^2 q(1-q)}. \tag{14}$$

Lemma C.2 allows to rewrite (14) as

$$\begin{aligned}
&(1+2\sigma(m,q))\binom{m-1}{\frac{m-1}{2}}q^{\frac{m-1}{2}}(1-q)^{\frac{m-1}{2}} \\
&= \frac{2q+2(2q-1)\sigma(m,q)-1}{2q-1}\binom{m-1}{\frac{m-1}{2}}q^{\frac{m-1}{2}}(1-q)^{\frac{m-1}{2}} \\
&\leq \frac{M_m(q)(1-M_m(q))}{q(1-q)}
\end{aligned}$$

or equivalently

$$(1+2\sigma(m,q))\binom{m-1}{\frac{m-1}{2}}q^{\frac{m+1}{2}}(1-q)^{\frac{m+1}{2}} \leq M_m(q)(1-M_m(q)). \tag{15}$$

We note that both sides approach zero from above as $q \to 1$, such that (15) holds for $q = 1$. It is thus sufficient to show, that the right side grows faster than the left side when decreasing q, i.e.

$$\frac{d}{dq}\left((1+2\sigma(m,q))\binom{m-1}{\frac{m-1}{2}}q^{\frac{m+1}{2}}(1-q)^{\frac{m+1}{2}}\right) \geq \frac{d}{dq}(M_m(q)(1-M_m(q))). \tag{16}$$

We have

$$\frac{d}{dq}(M_m(q)(1-M_m(q))) = (1-2M_m(q))m\binom{m-1}{\frac{m-1}{2}}q^{\frac{m-1}{2}}(1-q)^{\frac{m-1}{2}}$$

Meanwhile,

$$\begin{aligned}
&\frac{d}{dq}(1+2\sigma(m,q))\binom{m-1}{\frac{m-1}{2}}q^{\frac{m+1}{2}}(1-q)^{\frac{m+1}{2}} \\
&= (1+2\sigma(m,q))\frac{m+1}{2}(1-2q)\binom{m-1}{\frac{m-1}{2}}q^{\frac{m-1}{2}}(1-q)^{\frac{m-1}{2}} + \binom{m-1}{\frac{m-1}{2}}q^{\frac{m-1}{2}}(1-q)^{\frac{m-1}{2}}q(1-q)2\frac{d}{dq}\sigma(m,q) \\
&= \binom{m-1}{\frac{m-1}{2}}q^{\frac{m-1}{2}}(1-q)^{\frac{m-1}{2}}\left((1+2\sigma(m,q))\frac{m+1}{2}(1-2q) + q(1-q)2\frac{d}{dq}\sigma(m,q)\right).
\end{aligned}$$

Because $\binom{m-1}{\frac{m-1}{2}}q^{\frac{m-1}{2}}(1-q)^{\frac{m-1}{2}} > 0$ for $q < 1$, (16) or

$$\frac{d}{dq}(1+2\sigma(m,q))\binom{m-1}{\frac{m-1}{2}}q^{\frac{m+1}{2}}(1-q)^{\frac{m+1}{2}} \geq \frac{d}{dq}(M_m(q)(1-M_m(q)))$$

holds whenever

$$(1+2\sigma(m,q))\frac{m+1}{2}(1-2q) + q(1-q)2\frac{d}{dq}\sigma(m,q) \geq m(1-2M_m(q)). \tag{17}$$

Dividing by the (negative) $1-2M_m(q)$ term yields

$$(1+2\sigma(m,q))\frac{m+1}{2}\frac{(1-2q)}{1-2M_m(q)} + \frac{q(1-q)}{1-2M_m(q)}2\frac{d}{dq}\sigma(m,q) \leq m.$$

which is equivalent to

$$\frac{m+1}{2} + \frac{q(1-q)}{1-2M_m(q)}2\frac{d}{dq}\sigma(m,q) \leq m$$

as $\frac{(1-2q)}{1-2M_m(q)} = \frac{1}{1+2\sigma(m,q)}$. Rewriting yields

$$\begin{aligned}
\frac{m-1}{2} &= m - \frac{m+1}{2} \qquad (18)\\
&\geq -\frac{q(1-q)}{2M_m(q)-1}2\frac{d}{dq}\sigma(m,q) \\
&= -2\frac{q(1-q)}{2M_m(q)-1}(1-2q)\sum_{k \text{ uneven}}^{m-2}\frac{k+1}{2}\binom{k}{\frac{k+1}{2}}q^{\frac{k-1}{2}}(1-q)^{\frac{k-1}{2}} \\
&= 2\frac{2q-1}{2M_m(q)-1}\sum_{k \text{ uneven}}^{m-2}\frac{k+1}{2}\binom{k}{\frac{k+1}{2}}q^{\frac{k+1}{2}}(1-q)^{\frac{k+1}{2}} \\
&= 2\frac{1}{1+2\sigma(m,q)}\sum_{k \text{ uneven}}^{m-2}\frac{k+1}{2}\binom{k}{\frac{k+1}{2}}q^{\frac{k+1}{2}}(1-q)^{\frac{k+1}{2}} \qquad (19)
\end{aligned}$$

We can upper bound

$$\sum_{k \text{ uneven}}^{m-2} \frac{k+1}{2}\binom{k}{\frac{k+1}{2}} q^{\frac{k+1}{2}}(1-q)^{\frac{k+1}{2}} \leq \frac{m-2+1}{2} \sum_{k \text{ uneven}}^{m-2} \binom{k}{\frac{k+1}{2}} q^{\frac{k+1}{2}}(1-q)^{\frac{k+1}{2}}$$
$$= \frac{m-1}{2}\sigma(m,q)$$

such that (18) reduces to

$$\frac{m-1}{2} \geq \frac{m-1}{2}\frac{2\sigma(m,q)}{1+2\sigma(m,q)}, \tag{20}$$

which is clearly true, as $\frac{x}{1+x} < \frac{x}{x} = 1$ for all $x > 0$.

D.2. Correlated Classifiers

We now analyze the case of correlated classifiers discussed in 3.2, at first keeping $q = q_b = q_w$ fixed to be equal. As a reminder, we now have

$$G(q,p,\epsilon) = \begin{cases} 1 & \text{w.p. } q(1-p_w)p_b^0 + (1-q)p_w\big(1-p_b^1\big) \\ -1 & \text{w.p. } (1-q)(1-p_w)p_b^0 + qp_w\big(1-p_b^1\big) \\ 0 & \text{else} \end{cases}.$$

with expectation

$$(2q-1)(1-p_w)p_b^0 - (2q-1)p_w\big(1-p_b^1\big)$$
$$= (2q-1)\big((1-p_w)p_b^0 + p_w\big(p_b^1-1\big)\big) > 0.$$

We also note that

$$\begin{aligned}\mathbb{P}(G(q,p,\epsilon)=0) &= 1-\mathbb{P}(G(q,p,\epsilon)=1)-\mathbb{P}(G(q,p,\epsilon)=-1)\\ &= 1-q(1-p_w)p_b^0-(1-q)p_w\big(1-p_b^1\big)\\ &\quad -(1-q)(1-p_w)p_b^0-qp_w\big(1-p_b^1\big)\\ &= 1-(1-p_w)p_b^0-p_w\big(1-p_b^1\big) =: c_0\end{aligned}$$

is constant in q. Repeating the argument from above, we now obtain

$$\begin{aligned}&\Lambda_X'^*(0)\\ &= m\log\left(2\sqrt{\mathbb{P}(G(q,p,\epsilon)=1)\,\mathbb{P}(G(q,p,\epsilon)=-1)}+c_0\right)\\ &= m\log\Bigg(2\Bigg(\big(q(1-p_w)p_b^0(1-q)(1-p_w)p_b^0+q(1-p_w)p_b^0qp_w\big(1-p_b^1\big)\big)\\ &+(1-q)p_w\big(1-p_b^1\big)(1-q)(1-p_w)p_b^0+(1-q)p_w\big(1-p_b^1\big)qp_w\big(1-p_b^1\big)\Bigg)^{\frac{1}{2}}+c_0\Bigg)\\ &= m\log\Big(2\sqrt{q(1-q)c_1+qqc_2+(1-q)(1-q)c_3+(1-q)qc_4}+c_0\Big),\end{aligned}$$

where the c_i are constants that do not depend on q. We also note, that $p_w\big(1-p_b^1\big)(1-p_w)p_b^0 = c_2 = c_3$.

We now consider

$$\begin{aligned}f^*(q) &= (c_1+c_4)M_m(q)(1-M_m(q)) + c_2\Big(M_m(q)^2+(1-M_m(q))^2\Big)\\ &= (c_1+c_4)M_m(q)(1-M_m(q)) + c_2\Big(M_m(q)^2+1-2M_m(q)+M_m(q)^2\Big)\\ &= (c_1+c_4)M_m(q)(1-M_m(q)) + c_2\Big(2M_m(q)^2-2M_m(q)\Big)+c_2\\ &= (c_1+c_4)M_m(q)(1-M_m(q)) - 2c_2(M_m(q)(1-M_m(q)))+c_2\\ &= (c_1+c_4-2c_2)M_m(q)(1-M_m(q))+c_2\end{aligned}$$

and

$$g^*(q) = (c_1 + c_4 - 2c_2)q(1-q) + c_2,$$

such that

$$-\Lambda_X^*(0) + \Lambda_{X'}^*(0) = \log\Big(2\sqrt{f^*(q)} + c_0\Big) - m\log\Big(2\sqrt{g^*(q)} + c_0\Big),$$

where c_0 does not depend on q. This is exactly (10) with c_0 replacing c. A brief glance reveals that

$$\frac{(1-p_w)^2\big(p_b^0\big)^2 + \big(1-p_b^1\big)^2(p_w)^2}{2} \geq \big((1-p_w)p_b^0 p_w\big(1-p_b^1\big)\big)$$

by the AM-GM inequality, such that

$$c_1 + c_4 - 2c_2 > 0.$$

This means that $f^*(q)$ and $g^*(q)$ are exactly of the form $d_1 M_m(q)(1-M_m(q)) + d_2$ and $d_1 q(1-q) + d_2$ for constants $d_1 = c_1 + c_4 - 2c_2 > 0$ and $d_2 = c_2 > 0$. As it did not rely on the specific values for these constants beyond their positivity, the reasoning from the last section (where $d_1 = \epsilon^2$ and $d_2 = \epsilon(1-p-\epsilon)p + ((1-p-\epsilon)p)^2$) can be repeated one to one, proving our main result for correlated classifiers,

D.3. Correlated Classifiers and Labels

As in 3.2, we now consider

$$G(q,p,\epsilon) = \begin{cases} 1 & \text{w.p. } q_b(1-p_w)p_b^0 + (1-q_w)p_w\big(1-p_b^1\big) \\ -1 & \text{w.p. } (1-q_b)(1-p_w)p_b^0 + q_w p_w\big(1-p_b^1\big) \\ 0 & \text{else} \end{cases}.$$

with expectation

$$(2q_b - 1)(1-p_w)p_b^0 - (2q_w - 1)p_w\big(1-p_b^1\big) > 0.$$

We note that

$$\begin{aligned}\mathbb{P}(G(q,p,\epsilon)=0) &= 1 - \mathbb{P}(G(q,p,\epsilon)=1) - \mathbb{P}(G(q,p,\epsilon)=-1) \\ &= 1 - q_b(1-p_w)p_b^0 - (1-q_w)p_w\big(1-p_b^1\big) \\ &\quad - (1-q_b)(1-p_w)p_b^0 - q_w p_w\big(1-p_b^1\big) \\ &= 1 - (1-p_w)p_b^0 - p_w\big(1-p_b^1\big)\end{aligned}$$

still does not depend on either of the q_i, nor their difference. By assumption 1, we can reparameterise $q_b = q_w + \delta = q + \delta$ for $\delta \geq 0$ and we know by the previous calculations that (8) holds for $\delta = 0$. We now obtain

$$-\Lambda_X'^*(0) = m\log\Bigg(2\Big((q+\delta)(1-q-\delta)c_1 + (q+\delta)qc_2 + (1-q)(1-q-\delta)c_3 + (1-q)qc_4\Big)^{\frac{1}{2}} + c_0\Bigg),$$

where the constants c_i are as before and neither depend on q nor δ. We set

$$\begin{aligned}f^*(\delta) &= c_1 M_m(q+\delta)(1-M_m(q+\delta)) + c_2\Big(M_m(q)M_m(q+\delta) + (1-M_m(q))(1-M_m(q+\delta))\Big) \\ &\quad + c_4(1-M_m(q))M_m(q)\end{aligned}$$

and

$$g^*(\delta) = c_1(q+\delta)(1-q-\delta) + c_2(q(q+\delta) + (1-q)(1-q-\delta)) + c_4(1-q)q,$$

such that

$$-\Lambda_X^*(0) + \Lambda_{X'}^*(0) = \log\Big(2\sqrt{f^*(\delta)} + c_0\Big) - m\log\Big(2\sqrt{g^*(\delta)} + c_0\Big),$$

and we again have to show (11), i.e.

$$f^{*'}(\delta) \geq m g^{*'}(\delta) \frac{\sqrt{f^*(\delta)}\left(2\sqrt{f^*(\delta)} + c_0\right)}{\sqrt{g^*(\delta)}\left(2\sqrt{g^*(\delta)} + c_0\right)}$$

as we already know $-\Lambda_X^*(0) + \Lambda_{X'}^*(0)$ to be positive for $\delta = 0$. This time,

$$g^{*'}(\delta) = c_1(1 - 2(q+\delta)) - c_2(1-2q)$$

and

$$\begin{aligned}
f^{*'}(\delta) &= c_1(1 - 2M_m(q+\delta))m\binom{m-1}{\frac{m-1}{2}}(q+\delta)^{\frac{m-1}{2}}(1-q-\delta)^{\frac{m-1}{2}} \\
&+ c_2 M_m(q) m\binom{m-1}{\frac{m-1}{2}}(q+\delta)^{\frac{m-1}{2}}(1-q+\delta)^{\frac{m-1}{2}} \\
&- c_2(1 - M_m(q))m\binom{m-1}{\frac{m-1}{2}}(q+\delta)^{\frac{m-1}{2}}(1-q+\delta)^{\frac{m-1}{2}} \\
&= (c_1(1 - 2M_m(q+\delta)) - c_2(1-2M_m(q)))m\binom{m-1}{\frac{m-1}{2}}(q+\delta)^{\frac{m-1}{2}}(1-q-\delta)^{\frac{m-1}{2}}.
\end{aligned}$$

We note that

$$c_1 - c_2 = \left((1-p_w)p_b^0\right)^2 - (1-p_w)p_b^0 p_w\left(1-p_b^1\right),$$

which is positive if

$$(1-p_w)p_b^0 - p_w\left(1-p_b^1\right) > 0,$$

i.e.

$$(1-p_w)p_b^0 + p_w\left(p_b^1 - 1\right) > 0,$$

which we assumed to be true. Correspondingly, $c_1 > c_2$ and because $1-2(q+\delta)$ and $1-2M_m(q+\delta)$ are monotonously falling in δ, both $f^{*'}$ and $g^{*'}$ are negative. As such, (11) reduces to

$$\begin{aligned}
\frac{f^{*'}}{m g^{*'}} &= \binom{m-1}{\frac{m-1}{2}}(q+\delta)^{\frac{m-1}{2}}(1-q-\delta)^{\frac{m-1}{2}} \frac{(c_1(1 - 2M_m(q+\delta)) - c_2(1-2M_m(q)))}{c_1(1-2(q+\delta)) - c_2(1-2q)} \\
&\leq \frac{\sqrt{f^*(\delta)}\left(2\sqrt{f^*(\delta)} + c_0\right)}{\sqrt{g^*(\delta)}\left(2\sqrt{g^*(\delta)} + c_0\right)}.
\end{aligned}$$

To get a better handle on this inequality, we need the following lemma:

Lemma D.3. *Let c_1, c_2 be positive and A, B, C, D be negative constants such that $c_1 A - c_2 B < 0$ and $c_1 C - c_2 D < 0$. Then $\frac{c_1 A - c_2 B}{c_1 C - c_2 D} \leq \frac{A}{C}$ is true if and only if $CB \geq DA$.*

Proof.

$$\begin{aligned}
\frac{c_1 A - c_2 B}{c_1 C - c_2 D} \leq \frac{A}{C} &\iff c_1 A - c_2 B \geq \frac{A(c_1 C - c_2 D)}{C} \\
&\iff C(c_1 A - c_2 B) \leq A(c_1 C - c_2 D) \\
&\iff c_1 CA - c_2 CB \leq c_1 CA - c_2 DA \\
&\iff -c_2 CB \leq -c_2 DA \\
&\iff CB \geq DA
\end{aligned}$$

□

We set

$$A = (1 - 2M_m(q + \delta)),$$
$$B = (1 - 2M_m(q)),$$
$$C = (1 - 2(q + \delta)),$$
$$D = (1 - 2q),$$

such that $CB \geq DA$ is equivalent to

$$(1 - 2(q + \delta))(1 - 2M_m(q)) \geq (1 - 2q)(1 - 2M_m(q + \delta)),$$

or

$$(1 - 2M_m(q)) \leq (1 - 2q)\frac{(1 - 2M_m(q + \delta))}{(1 - 2(q + \delta))},$$

i.e.

$$\frac{(1 - 2M_m(q))}{(1 - 2q)} \geq \frac{(1 - 2M_m(q + \delta))}{(1 - 2(q + \delta))},$$

which is equivalent to

$$1 + 2\sigma(m, q) \geq 1 + 2\sigma(m, q + \delta),$$

which holds as $\sigma(m, x)$ is clearly monotonically decreasing in x for $x > 0.5$.

Lemma D.3 allows us to upper bound

$$\begin{aligned}\frac{(c_1(1 - 2M_m(q + \delta)) - c_2(1 - 2M_m(q)))}{c_1(1 - 2(q + \delta)) - c_2(1 - 2q)} &\leq \frac{(1 - 2M_m(q + \delta))}{1 - 2(q + \delta)} \\ &= 1 + 2\sigma(m, q + \delta).\end{aligned}$$

Correspondingly, (11) reduces to

$$\begin{aligned}&\binom{m-1}{\frac{m-1}{2}}(q + \delta)^{\frac{m-1}{2}}(1 - q - \delta)^{\frac{m-1}{2}}(1 + 2\sigma(m, q + \delta)) \\ &\leq \frac{\sqrt{f^*(\delta)}\Big(2\sqrt{f^*(\delta)} + c_0\Big)}{\sqrt{g^*(\delta)}\Big(2\sqrt{g^*(\delta)} + c_0\Big)}.\end{aligned} \tag{21}$$

To control this, we need another lemma:

Lemma D.4. *Let $c, f_1, f_2, g_1, g_2 > 0$; $f_1 \leq g_1$ and $f_2 \geq g_2$. Then*

$$\frac{f_1}{g_1} \leq \frac{2(f_1 + f_2) + c\sqrt{f_1 + f_2}}{2(g_1 + g_2) + c\sqrt{g_1 + g_2}}$$

Proof. We first note that

$$(f_1 - g_1)f_1g_1 \leq g_1^2 f_2 - f_1^2 g_2,$$

as the left side is always negative because $f_1 \leq g_1$, while the right side is always positive as $g_1 \geq f_1$ and $f_2 \geq g_2$. With this, we calculate

$$\begin{aligned}&(f_1 - g_1)f_1g_1 \leq g_1^2 f_2 - f_1^2 g_2 \\ &\iff f_1^2 g_1 + f_1^2 g_2 \leq g_1^2 f_1 + g_1^2 f_2 \\ &\iff f_1^2(g_1 + g_2) \leq g_1^2(f_1 + f_2) \\ &\iff f_1\sqrt{g_1 + g_2} \leq g_1\sqrt{f_1 + f_2}.\end{aligned}$$

With this,

$$\begin{aligned}
&\frac{f_1}{g_1} \leq \frac{2(f_1+f_2)+c\sqrt{f_1+f_2}}{2(g_1+g_2)+c\sqrt{g_1+g_2}} \\
&\iff f_1\big(2(g_1+g_2)+c\sqrt{g_1+g_2}\big) \leq g_1\Big(2(f_1+f_2)+c\sqrt{f_1+f_2}\Big) \\
&\iff f_1\big(2g_2+c\sqrt{g_1+g_2}\big) \leq g_1\Big(2f_2+c\sqrt{f_1+f_2}\Big).
\end{aligned}$$

The inequality now holds for the second terms on each side by our previous calculations, and for the first terms on each side as $f_1 \leq g_1$ and $g_2 \leq f_2$. □

We set

$$f_1 = c_1 M_m(q+\delta)(1-M_m(q+\delta)) + c_4(1-M_m(q))M_m(q),$$
$$g_1 = c_1(q+\delta)(1-q-\delta) + c_4(1-q)q,$$

as well as

$$f_2 = c_2(M_m(q)M_m(q+\delta) + (1-M_m(q))(1-M_m(q+\delta))),$$

and

$$g_2 = c_2(q(q+\delta) + (1-q)(1-q-\delta)),$$

such that

$$f_1 + f_2 = f^*(\delta)$$

and

$$g_1 + g_2 = g^*(\delta).$$

If we can prove the preconditions for D.4, (21) will reduce to

$$\binom{m-1}{\frac{m-1}{2}}(q+\delta)^{\frac{m-1}{2}}(1-q-\delta)^{\frac{m-1}{2}}(1+2\sigma(m,q+\delta)) \leq \frac{f_1}{g_1}. \tag{22}$$

$f_1 \leq g_1$ is easy to see, based on the increasingness of $M_m(q)$ in m, and the decreasingness of $x(1-x)$ in x for $x > 0.5$. We can thus focus on showing $g_2 \leq f_2$, i.e.

$$\begin{aligned}
&(q(q+\delta)+(1-q)(1-q-\delta)) \\
&\leq (M_m(q)M_m(q+\delta)+(1-M_m(q))(1-M_m(q+\delta))).
\end{aligned} \tag{23}$$

At $\delta = 1-q$, (23) becomes

$$q \leq M_m(q),$$

which is clearly true. At $\delta = 0$, we get

$$q^2 + (1-q)^2 \leq M_m(q)^2 + (1-M_m(q))^2.$$

We note that

$$x^2 + (1-x)^2 = 1 + 2\big(x^2 - x\big)$$

has the derivative $4x-2$, which is positive for $x > 0.5$. Correspondingly, the $M_m(q)$ term is larger than the q term. Having shown that (23) holds at both extreme values for δ, it is sufficient for Lemma D.4 to hold to show that the second derivative of

$$(q(q+\delta)+(1-q)(1-q-\delta)) - (M_m(q)M_m(q+\delta)+(1-M_m(q))(1-M_m(q+\delta)))$$

with respect to δ is positive, such that the function is convex. As the left term is linear in δ, this derivative equals

$$-M_m(q)\frac{d^2}{d^2\delta}M_m(q+\delta) + (1-M_m(q))\frac{d^2}{d^2\delta}M_m(q+\delta),$$

which equals

$$(1 - 2M_m(q))\frac{d^2}{d^2\delta}M_m(q+\delta)$$

and thus has the opposite sign of $\frac{d^2}{d^2\delta}M_m(q+\delta)$, which is negative due to the well-known concavity of the majority vote in $M_m(x)$ in x for $x > 0.5$ (Boland et al., 1989).

To prove (22), we need one last lemma:

Lemma D.5. *Let $A, B, C, D > 0$ and $AD \leq BC$. Then, $\frac{A}{C} \leq \frac{A+B}{C+D}$*

Proof.

$$\frac{A}{C} \leq \frac{A+B}{C+D} \iff AC + AD \leq AC + BC \iff AD \leq BC$$

□

We set

$$\begin{aligned} A &= c_1 M_m(q+\delta)(1 - M_m(q+\delta)), \\ B &= c_4(1 - M_m(q))M_m(q), \\ C &= c_1(q+\delta)(1-q-\delta), \\ D &= c_4(1-q)q, \end{aligned}$$

such that

$$f_1 = A + B$$

and

$$g_1 = C + D.$$

If we can show that $AD \leq BC$, (22) would reduce to

$$\begin{aligned} &\binom{m-1}{\frac{m-1}{2}}(q+\delta)^{\frac{m-1}{2}}(1-q-\delta)^{\frac{m-1}{2}}(1 + 2\sigma(m, q+\delta)) \\ &\leq \frac{M_m(q+\delta)(1 - M_m(q+\delta))}{(q+\delta)(1-q-\delta)}, \end{aligned} \tag{24}$$

which is equivalent to (14) and true by the calculations in section D.1. $AD \leq BC$ is equivalent to

$$M_m(q+\delta)(1 - M_m(q+\delta))(1-q)q \leq (1 - M_m(q))M_m(q)(q+\delta)(1-q-\delta).$$

This is again clearly true for $\delta = 0$ where both sides are equal, such that it is sufficient to show that

$$\frac{M_m(q+\delta)(1 - M_m(q+\delta))(1-q)q}{(1 - M_m(q))M_m(q)(q+\delta)(1-q-\delta)}$$

or

$$\frac{(1-q)q}{(1 - M_m(q))M_m(q)} \frac{M_m(q+\delta)(1 - M_m(q+\delta))}{(q+\delta)(1-q-\delta)}$$

is maximized at $\delta = 0$. As the first term does not depend on δ, we only need to analyze the second term. Reparameterizing $x = q + \delta$, it is thus sufficient to show that

$$\frac{M_m(x)(1 - M_m(x))}{x(1-x)}$$

decreases monotonously in x. We take derivatives with respect to x, obtaining

$$\frac{(1 - 2M_m(x))m\binom{m-1}{\frac{m-1}{2}}x^{\frac{m+1}{2}}(1-x)^{\frac{m+1}{2}} - M_m(x)(1 - M_m(x))(1-2x)}{x^2(1-x)^2}.$$

This is negative, whenever

$$(1-2M_m(x))m\binom{m-1}{\frac{m-1}{2}}x^{\frac{m+1}{2}}(1-x)^{\frac{m+1}{2}} \leq M_m(x)(1-M_m(x))(1-2x)$$

or equivalently

$$\frac{1-2M_m(x)}{1-2x}m\binom{m-1}{\frac{m-1}{2}}x^{\frac{m+1}{2}}(1-x)^{\frac{m+1}{2}} \geq M_m(x)(1-M_m(x)),$$

i.e.

$$(1+2\sigma(m,x))m\binom{m-1}{\frac{m-1}{2}}x^{\frac{m+1}{2}}(1-x)^{\frac{m+1}{2}} \geq M_m(x)(1-M_m(x)).$$

As both sides tend to zero for $x \to 1$, it is sufficient to show that the right term increases more slowly as x decreases, i.e.

$$\frac{d}{dq}\left((1+2\sigma(m,x))m\binom{m-1}{\frac{m-1}{2}}x^{\frac{m+1}{2}}(1-x)^{\frac{m+1}{2}}\right) \leq \frac{d}{dq}(M_m(x)(1-M_m(x))). \tag{25}$$

Note, that this equation is the reverse of (16), but with an additional factor of m on the left side. Repeating the calculations from Section D.1, (25) reduces to

$$m\left(\frac{m+1}{2}+\frac{q(1-q)}{1-2M_m(q)}2\frac{d}{dq}\sigma(m,q)\right) \geq m$$

or

$$\frac{m+1}{2}+\frac{q(1-q)}{1-2M_m(q)}2\frac{d}{dq}\sigma(m,q) \geq 1.$$

The $\frac{q(1-q)}{1-2M_m(q)}2\frac{d}{dq}\sigma(m,q)$ term is positive, as both the first and the second factor are clearly negative, such that the equation holds, finishing our proof of

$$\mathbb{P}\left(\sum_i^n G_i(M_m(q),p)>0\right) < \mathbb{P}\left(\sum_i^{mn} G_i(q,p)>0\right).$$

It remains to show that for fixed $q_b \geq q_w$ and $m > 1$ uneven, G in the heterogeneous case stochastically dominates G for the homogeneous whenever assumption 2 holds. This would imply that the sum of G_i follows the same dominance relation, such that the probability of correctly identifying c_b is larger for the $m-$label case assuming homogeneity rather than explicitly modelling heterogeneity. We note that $\mathbb{P}(G(q,p)=0)$ does not depend on q, such that it is sufficient to show that $\mathbb{P}(G(q,p)=1)$ is larger in the homogeneous case. We rewrite

$$\begin{aligned}
&\frac{(1-p_w)p_b^0}{p_w(1-p_b^1)}\Big(M_m(q_b)-\mathop{\mathbb{E}}_x[M_m(q(x))|E_b]\Big) \geq M_m(q_w)-\mathop{\mathbb{E}}_x[M_m(q(x))|E_w]\\
\iff &(1-p_w)p_b^0\Big(M_m(q_b)-\mathop{\mathbb{E}}_x[M_m(q(x))|E_b]\Big)\\
&\geq p_w\big(1-p_b^1\big)\Big(M_m(q_w)-\mathop{\mathbb{E}}_x[M_m(q(x))|E_w]\Big)\\
\iff &(1-p_w)p_b^0M_m(q_b)-p_w\big(1-p_b^1\big)M_m(q_w)\\
&\geq (1-p_w)p_b^0\mathop{\mathbb{E}}_x[M_m(q(x))|E_b]-p_w\big(1-p_b^1\big)\mathop{\mathbb{E}}_x[M_m(q(x))|E_w]\\
\iff &(1-p_w)p_b^0M_m(q_b)-p_w\big(1-p_b^1\big)(1-M_m(q_w))\\
&\geq (1-p_w)p_b^0\mathop{\mathbb{E}}_x[M_m(q(x))|E_b]-p_w\big(1-p_b^1\big)\Big(1-\mathop{\mathbb{E}}_x[M_m(q(x))|E_w]\Big)\\
\iff &\mathbb{P}(G(M_m(q_b),M_m(q_w),p)=1)\\
&\geq \mathbb{P}\Big(G\Big(\mathop{\mathbb{E}}_x[M_m(q(x))|E_b],\mathop{\mathbb{E}}_x[M_m(q(x))|E_w],p\Big)=1\Big),
\end{aligned}$$

showing that the heterogeneous case is dominated by the homogeneous case under assumption 2.