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

Florian E. Dorner[1,2] and Moritz Hardt[1]
[1]Max Planck Institute for Intelligent Systems, Tübingen and Tübingen AI Center
[2]ETH Zürich

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. We discuss the implications of our work for the design of machine learning benchmarks, where they overturn some time-honored recommendations.

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.

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. This is particularly relevant for selecting fine-tuned binary classifiers for deployment in settings where labeled data is sparse, but also has implications for efficiently benchmarking reward models for RLHF Ouyang et al. (2022) at scale, as these models act as binary classifiers on pairs of outputs.

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.*

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

Corresponding author: florian.dorner@tuebingen.mpg.de

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 our main theorem is asymptotic, we conjecture that the statement holds for all $n \geq 1$. We have verified this numerically in a vast parameter sweep spanning more than four billion values. The numerical tool we created also serves as an effective way to calculate tight sample size requirements and is available at https://labelnoise.pythonanywhere.com.

There is a common belief that benchmark designers should clean 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 (Gray & Suri, 2019). "Gold standard" labels for datasets like ImageNet (Russakovsky et al., 2015) are routinely produced by aggregating multiple annotators' labels. In natural language processing, classic benchmarks like SST (Socher et al., 2013) and MNLI (Bowman et al., 2015) also base labels on a per-instance majority vote after collecting multiple labels for each instance. More recently, label aggregation has been used for evaluating the safety of LLama2 (Touvron et al., 2023), while OpenAssistant (Köpf et al., 2023) aggregates users' rankings 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 more labels per instance for *testing* than for training (Williams et al., 2017; Dorner et al., 2022).

Annotator disagreement as a feature. Aroyo & Welty (2013) argue that due to the lack of objective ground truth, taking annotator disagreement into account is essential. The authors suggest to to use non-binary labels that encompass disagreement. Similarly, predicting individual annotators' responses is a common approach (Tanno et al., 2019; Davani et al., 2022; Gordon et al., 2022; Fleisig et al., 2023). To enable this, 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 assume that the target label is agreed upon to be given by a 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.

2 Formal results

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 frequency of classification errors

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

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

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

for accuracy $p_w \in [0.5, 1]$ and margin $\epsilon \in (0, 1 - p_w]$. 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 as follows: 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.

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. $q(x) \coloneqq \mathbb{P}_l(y_l(x) = y_{True}(x))$. Similarly, q denotes the same probability marginalized over both x and l: $q \coloneqq \mathbb{P}_{x,l}(y_l(x) = y_{True}(x))$. 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 $M_m(q) \coloneqq \mathbb{P}(\text{Majority of } m \text{ independent voters correct})$ 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

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

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

We model 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 $(1 - p_w)p_b^0 + p_w\big(p_b^1 - 1\big) > 0$. 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, 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.

We abbreviate q_w, q_b as q, p_w, p_b^0, p_b^1 as p and $M_m(q_b), M_m(q_w)$ as $M_m(q)$ and treat the treat the gap indicator G as a function of q and p. We focus on the case of homogeneous label errors over x, i.e. $q(x) = q_k$ for k determined by $x \in E_k$. This allows us to use the equivalence of a single labeler with accuracy $M_m(q_k(x))$ and m labelers with accuracies $q_k(x)$ each, to compare $G(q)$ and $G(M_m(q))$ rather than explicitly parameterizing G by m.

We are 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 want to know whether for independent copies G_i of G

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

For small n and m this can be checked by numerically calculating the exact probabilities. It is consistently true 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. Figure 1 shows these exact probabilities for fixed accuracies $p = p_w = 0.8$, $p_b^0 = p_b^1 = p + \epsilon = 0.81$ and varying values of the label accuracy q and label budget k. As can be seen, the single label approach consistently outperforms $m > 1$ labels.

If labels are substantially more accurate in the event E_w, 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. We assume that this does not happen:

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

Note that the homogeneity in label accuracy $q(x)$ we assumed is not necessarily the best case for $m > 1$: 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:

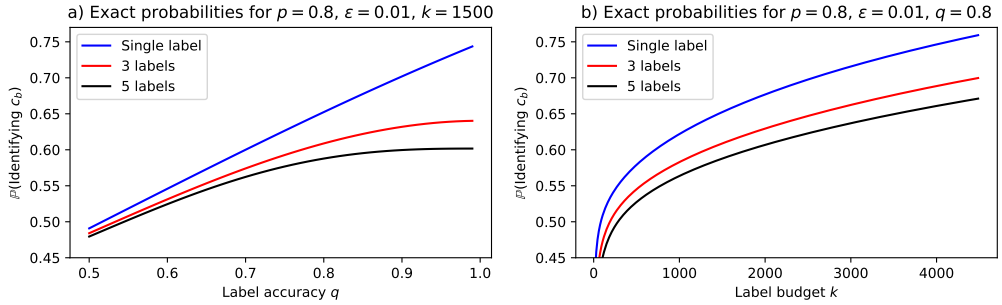

Figure 1: Probability of identifying c_b for accuracy $p = 0.8$, margin $\epsilon = 0.01$, budget $k = 1500$ (a), label accuracy $q = 0.8$ (b).

Assumption 2. *No biased heterogeneity:*

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

The $\frac{(1-p_w)p_b^0}{p_w(1-p_b^1)}$ 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 when heterogeneity is similar conditional on the events E_b and E_w.

Theorem 2. *For G as defined above, $m > 1$ 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 assumptions 1 and 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. The theorem makes heavy use of Cramér's theorem from large deviation theory (Klenke, 2013) and is proven in Appendix C.

3 Discussion

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. 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. When designing a new benchmark, we still encourage practitioners to initially collect multiple annotations for a few instances, in order to better understand ambiguities in their task definition and how annotators' identity influences their labels Denton et al. (2021). This 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 well. Once the instructions and annotator pool are fixed and it comes to evaluation at scale, we recommend to build the test set using a large number of instances with a single label each, according to budget.

While we do not study the effects of aggregation for datasets that include multiple labels per instance, we 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 can choose whether and how to aggregate labels Prabhakaran et al. (2021). Otherwise, they cannot obtain information about annotator disagreement, or simply use another aggregation method more suited to their needs.

4 Limitations

Our recommendations might not hold 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.pythonanywhere.com.

On the theoretical side, our main theorem is asymptotic. Based on our numerical results, we conjecture that it is always true, but are unable to proof this at the current stage:

Conjecture 1. *For G defined as in Section 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.

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.

Similarly, smarter *adaptive* labeling strategies that are out of the scope of this work, 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.

5 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).

References

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

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

Philip J Boland, Frank Proschan, and Yung Liang Tong. Modelling dependence in simple and indirect majority systems. *Journal of Applied Probability*, 26(1):81–88, 1989.

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

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

Emily Denton, Mark Díaz, Ian Kivlichan, Vinodkumar Prabhakaran, and Rachel Rosen. Whose ground truth? accounting for individual and collective identities underlying dataset annotation. *arXiv preprint arXiv:2112.04554*, 2021.

Florian E Dorner, Momchil Peychev, Nikola Konstantinov, Naman Goel, Elliott Ash, and Martin Vechev. Human-guided fair classification for natural language processing. *arXiv preprint arXiv:2212.10154*, 2022.

Eve Fleisig, Aubrie Amstutz, Chad Atalla, Su Lin Blodgett, Hal Daumé III, Alexandra Olteanu, Emily Sheng, Dan Vann, and Hanna Wallach. 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.

Mitchell L Gordon, Michelle S Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S Bernstein. 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.

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

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

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

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, et al. Openassistant conversations–democratizing large language model alignment. *arXiv preprint arXiv:2304.07327*, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. On releasing annotator-level labels and information in datasets. *arXiv preprint arXiv:2110.05699*, 2021.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pp. 5389–5400. PMLR, 2019.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 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.

Ryutaro Tanno, Ardavan Saeedi, Swami Sankaranarayanan, Daniel C Alexander, and Nathan Silberman. 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.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 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 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 B.1. *Assuming mututally independent classifier and labeler errors (i.e. $p_w = p$, $p_b^0 = p_b^1 = p + \epsilon$, $q_w = q_b = q$), G can be written as follows:*

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

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

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

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 Proving theorem 2

Theorem 2. *For G as defined above, $m>1$ 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*

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

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

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

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

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

and

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

If we can prove that

$$-\Lambda_X^*(0) > -\Lambda_{X'}^*(0), \tag{2}$$

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 C.1. *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

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

□

C.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 C.1 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

$$\begin{aligned} &-\Lambda_X^*(0) + \Lambda_{X'}^*(0) \qquad (3)\\ &= \log\left(2\sqrt{M_m(q)(1 - M_m(q))\epsilon^2 + \epsilon(1 - p - \epsilon)p + ((1 - p - \epsilon)p)^2} + 1 - \epsilon - 2p(1 - p - \epsilon)\right)\\ &- m\log\left(2\sqrt{q(1 - q)\epsilon^2 + \epsilon(1 - p - \epsilon)p + ((1 - p - \epsilon)p)^2} + 1 - \epsilon - 2p(1 - p - \epsilon)\right). \end{aligned}$$

Because (2) 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 (3) 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

$$-\Lambda_X^*(0) + \Lambda_{X'}^*(0) = \log\left(2\sqrt{f^*(q)} + c\right) - m\log\left(2\sqrt{g^*(q)} + c\right). \qquad (4)$$

Differentiating yields

$$\begin{aligned} \frac{d}{dq}(-\Lambda_X^*(0) + \Lambda_{X'}^*(0)) &= \frac{d}{dq}\log\left(2\sqrt{f^*(q)} + c\right) - m\frac{d}{dq}\log\left(2\sqrt{g^*(q)} + c\right)\\ &= \frac{\frac{d}{dq}2\sqrt{f^*(q)}}{\left(2\sqrt{f^*(q)} + c\right)} - m\frac{\frac{d}{dq}2\sqrt{g^*(q)}}{\left(2\sqrt{g^*(q)} + c\right)}\\ &= \frac{\frac{f^{*'}(q)}{\sqrt{f^*(q)}}}{\left(2\sqrt{f^*(q)} + c\right)} - m\frac{\frac{g^{*'}(q)}{\sqrt{g^*(q)}}}{\left(2\sqrt{g^*(q)} + c\right)}\\ &= \frac{f^{*'}(q)}{\sqrt{f^*(q)}\left(2\sqrt{f^*(q)} + c\right)}\\ &- m\frac{g^{*'}(q)}{\sqrt{g^*(q)}\left(2\sqrt{g^*(q)} + c\right)}. \end{aligned}$$

Correspondingly using (3), (2) reduces to

$$f^{*'}(q) \geq mg^{*'}(q)\frac{\sqrt{f^*(q)}\left(2\sqrt{f^*(q)} + c\right)}{\sqrt{g^*(q)}\left(2\sqrt{g^*(q)} + c\right)} \qquad (5)$$

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, (5) holds if and only if

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

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

$$\frac{2M_m(q) - 1}{2q - 1} \binom{m-1}{\frac{m-1}{2}} q^{\frac{m-1}{2}} (1-q)^{\frac{m-1}{2}} \leq \frac{\sqrt{f^*(q)}\left(2\sqrt{f^*(q)} + c\right)}{\sqrt{g^*(q)}\left(2\sqrt{g^*(q)} + c\right)}. \tag{6}$$

Lemma C.2. *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 C.2 implies that it is sufficient to show

$$\frac{2M_m(q) - 1}{2q - 1} \binom{m-1}{\frac{m-1}{2}} q^{\frac{m-1}{2}} (1-q)^{\frac{m-1}{2}} \leq \frac{f^*(q)}{g^*(q)}. \tag{7}$$

Lemma C.3. *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 C.3 implies that (7) can be reduced to

$$\frac{2M_m(q)-1}{2q-1}\binom{m-1}{\frac{m-1}{2}}q^{\frac{m-1}{2}}(1-q)^{\frac{m-1}{2}} \leq \frac{\epsilon^2 M_m(q)(1-M_m(q))}{\epsilon^2 q(1-q)}. \tag{8}$$

Next, we need the following lemma:

Lemma C.4. *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. □

Lemma C.4 allows to rewrite (8) 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

$$(1+2\sigma(m,q))\binom{m-1}{\frac{m-1}{2}}q^{\frac{m+1}{2}}(1-q)^{\frac{m+1}{2}} \leq M_m(q)(1-M_m(q)). \tag{9}$$

We note that both sides approach zero from above as $q \to 1$, such that (9) 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.

$$\frac{d}{dq}\left((1+2\sigma(m,q))\binom{m-1}{\frac{m-1}{2}}q^{\frac{m+1}{2}}(1-q)^{\frac{m+1}{2}}\right) \geq \frac{d}{dq}(M_m(q)(1-M_m(q))). \tag{10}$$

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 $q < 1$, (10) 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

$$(1+2\sigma(m,q))\frac{m+1}{2}(1-2q) + q(1-q)2\frac{d}{dq}\sigma(m,q) \geq m(1-2M_m(q)). \tag{11}$$

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

$$\begin{aligned}
\frac{m-1}{2} &= m - \frac{m+1}{2} && (12)\\
&\geq -\frac{q(1-q)}{2M_m(q)-1}2\frac{d}{dq}\sigma(m,q) \\
&= -2\frac{q(1-q)}{2M_m(q)-1}(1-2q)\sum_{k \text{ uneven}}^{m-2}\frac{k+1}{2}\binom{k}{\frac{k+1}{2}}q^{\frac{k-1}{2}}(1-q)^{\frac{k-1}{2}} \\
&= 2\frac{2q-1}{2M_m(q)-1}\sum_{k \text{ uneven}}^{m-2}\frac{k+1}{2}\binom{k}{\frac{k+1}{2}}q^{\frac{k+1}{2}}(1-q)^{\frac{k+1}{2}} \\
&= 2\frac{1}{1+2\sigma(m,q)}\sum_{k \text{ uneven}}^{m-2}\frac{k+1}{2}\binom{k}{\frac{k+1}{2}}q^{\frac{k+1}{2}}(1-q)^{\frac{k+1}{2}} && (13)
\end{aligned}$$

We can upper bound

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

such that (12) reduces to

$$\frac{m-1}{2} \geq \frac{m-1}{2}\frac{2\sigma(m,q)}{1+2\sigma(m,q)}, \tag{14}$$

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

C.2 Correlated classifiers

We now analyze the case of correlated classifiers discussed in 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

$$\begin{aligned}&(2q-1)(1-p_w)p_b^0 - (2q-1)p_w\big(1-p_b^1\big)\\ &= (2q-1)\big((1-p_w)p_b^0 + p_w\big(p_b^1-1\big)\big) > 0.\end{aligned}$$

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

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

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 (4) with c_0 replacing c. A brief glance reveals that

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

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 $d_1M_m(q)(1-M_m(q)) + d_2$ and $d_1q(1-q) + d_2$ 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,

C.3 Correlated classifiers and labels

As in 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

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

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 (2) 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

$$\begin{aligned}f^*(\delta) &= c_1M_m(q+\delta)(1-M_m(q+\delta)) + c_2\Big(M_m(q)M_m(q+\delta) + (1-M_m(q))(1-M_m(q+\delta))\Big)\\ &\quad + c_4(1-M_m(q))M_m(q)\end{aligned}$$

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 (5), i.e.

$$f^{*'}(\delta) \geq mg^{*'}(\delta)\frac{\sqrt{f^*(\delta)}\Big(2\sqrt{f^*(\delta)} + c_0\Big)}{\sqrt{g^*(\delta)}\Big(2\sqrt{g^*(\delta)} + c_0\Big)}$$

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

$$\begin{aligned}f^{*'}(\delta) &= c_1(1-2M_m(q+\delta))m\binom{m-1}{\frac{m-1}{2}}(q+\delta)^{\frac{m-1}{2}}(1-q-\delta)^{\frac{m-1}{2}}\\ &\quad + c_2M_m(q)m\binom{m-1}{\frac{m-1}{2}}(q+\delta)^{\frac{m-1}{2}}(1-q+\delta)^{\frac{m-1}{2}}\\ &\quad - c_2(1-M_m(q))m\binom{m-1}{\frac{m-1}{2}}(q+\delta)^{\frac{m-1}{2}}(1-q+\delta)^{\frac{m-1}{2}}\\ &= (c_1(1-2M_m(q+\delta)) - c_2(1-2M_m(q)))m\binom{m-1}{\frac{m-1}{2}}(q+\delta)^{\frac{m-1}{2}}(1-q-\delta)^{\frac{m-1}{2}}.\end{aligned}$$

We note that

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

which is positive if

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

i.e.

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

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, (5) reduces to

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

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

Lemma C.5. *Let c_1, c_2 be positive and A, B, C, D be negative constants such that $c_1A - c_2B < 0$ and $c_1C - c_2D < 0$. Then $\frac{c_1A-c_2B}{c_1C-c_2D} \leq \frac{A}{C}$ is true if and only if $CB \geq DA$.*

Proof.

$$\begin{aligned}\frac{c_1A - c_2B}{c_1C - c_2D} \leq \frac{A}{C} &\iff c_1A - c_2B \geq \frac{A(c_1C - c_2D)}{C} \\ &\iff C(c_1A - c_2B) \leq A(c_1C - c_2D) \\ &\iff c_1CA - c_2CB \leq c_1CA - c_2DA \\ &\iff -c_2CB \leq -c_2DA \\ &\iff CB \geq DA\end{aligned}$$

□

We set

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

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 C.5 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, (5) reduces to

$$\begin{aligned}&\binom{m-1}{\frac{m-1}{2}}(q+\delta)^{\frac{m-1}{2}}(1-q-\delta)^{\frac{m-1}{2}}(1+2\sigma(m,q+\delta))\\&\leq\frac{\sqrt{f^*(\delta)}\Big(2\sqrt{f^*(\delta)}+c_0\Big)}{\sqrt{g^*(\delta)}\Big(2\sqrt{g^*(\delta)}+c_0\Big)}.\end{aligned}\tag{15}$$

To control this, we need another lemma:

Lemma C.6. *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

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

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

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

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

$$\begin{aligned}f_1 &= c_1M_m(q+\delta)(1-M_m(q+\delta)) + c_4(1-M_m(q))M_m(q),\\g_1 &= c_1(q+\delta)(1-q-\delta) + c_4(1-q)q,\end{aligned}$$

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 C.6, (15) will reduce to

$$\binom{m-1}{\frac{m-1}{2}}(q+\delta)^{\frac{m-1}{2}}(1-q-\delta)^{\frac{m-1}{2}}(1+2\sigma(m,q+\delta)) \leq \frac{f_1}{g_1}.\tag{16}$$

$f_1 \leq g_1$ 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.

$$\begin{aligned}&(q(q+\delta) + (1-q)(1-q-\delta))\\&\leq (M_m(q)M_m(q+\delta) + (1-M_m(q))(1-M_m(q+\delta))).\end{aligned}\tag{17}$$

At $\delta = 1 - q$, (17) 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 (17) holds at both extreme values for δ, it is sufficient for Lemma C.6 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 (16), we need one last lemma:

Lemma C.7. *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 $AD \leq BC$, (16) would reduce to

$$\begin{aligned} &\binom{m-1}{\frac{m-1}{2}}(q+\delta)^{\frac{m-1}{2}}(1-q-\delta)^{\frac{m-1}{2}}(1 + 2\sigma(m, q+\delta)) \\ &\leq \frac{M_m(q+\delta)(1 - M_m(q+\delta))}{(q+\delta)(1-q-\delta)}, \end{aligned} \tag{18}$$

which is equivalent to (8) and true by the calculations in section C.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.

$$\frac{d}{dq}\left((1+2\sigma(m,x))m\binom{m-1}{\frac{m-1}{2}}x^{\frac{m+1}{2}}(1-x)^{\frac{m+1}{2}}\right) \leq \frac{d}{dq}(M_m(x)(1-M_m(x))). \tag{19}$$

Note, that this equation is the reverse of (10), but with an additional factor of m on the left side. Repeating the calculations from Section C.1, (19) 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.