Tolerating Adversarial Reviews via Active Learning

Joshua Turner

May 7, 2023

1 Introduction

1.1 Background

During the summer of 2021, I took part in an NSF REU program at the University of Maryland entitled "Combinatorics, Algorithms, and AI for Real Problems." Under the guidance of Professor Jonathan Katz and PhD candidate Benjamin Sela, my fellow researchers and I expanded on research directions suggested by Mr. Sela in his previous work. Though we all worked in the same problem setting, we each explored different directions.

This document is a summary of the problem setting and results of the particular research direction I pursued; its purpose is to outline my thought process and provide intuition for results, rather than to act as a comprehensive research paper.

1.2 Motivation

An important part of the human decision-making process is the evaluation of the quality of different options and the truth of different statements. For example, we purchase based on the perceived quality of products and act based on plausible news we hear about the world.

In evaluating the quality of a product or the truth of a statement, we might turn to the evaluations of others as a decision-making heuristic. By incorporating others' feedback, we might hope to benefit from the so-called "wisdom of the crowd." When those giving feedback are honest and share the decision-maker's preferences, this strategy can be a helpful shortcut.

However, the wisdom of the crowd can be poisoned by a malicious actor purposefully trying to lead such estimation astray. For example, consider the Russia's digital influence on the 2016 US presidential election. How might we make our evaluation process more robust against the intervention of malicious actors? In particular, how can learning more about a small subset of our objects help root out adversarial reviews of those objects, thus improving estimation? These questions guided my research. The hope is that we will be able to use the accuracy of particular reviews by the same reviewer in order to draw conclusions about whether that reviewer is malicious.

1.3 Model

Let n be the number of items being reviewed. We partially answer the above questions under the following assumptions.

Assumption 1. There is an unknown "ground truth" vector $\boldsymbol{\theta} \in \{0,1\}^n$, where $\boldsymbol{\theta}_i$ gives the true quality of item i.

In this model, we view an item as having a "true" quality of either good (1) or bad (0). Our task is to estimate $\boldsymbol{\theta}$ given a multiset of reviews R, where each element of R is a vector $\mathbf{r} \in \{0,1\}^n$. We interpret each review as being a reviewer's evaluation of each of the n items. Notice the additional assumption implicit in our definition of R that every review has something to say about every item.

For estimation to be possible, we need some subset of the reviews to correlate with the ground truth $\boldsymbol{\theta}$. This brings us to our next assumption. Suppose $p, \alpha \in [0,1]$ with $p > \frac{1}{2}$ and $\alpha < \frac{1}{2}$.

Assumption 2. A $(1-\alpha)$-fraction of the reviews in R are *honest*. An honest review $\mathbf{r}'$ satisfies $\mathbf{r}'_i = \boldsymbol{\theta}_i$ with independent probability p for all $i \leq n$.

In other words, most of the reviews will give the correct answer for most of the items. As we will see, this still leaves plenty of room for mischief by a malicious actor trying to poison the estimation process.

Assumption 3. An α-fraction of the reviews in R are *maliciously generated* by a strong adversary.

This strong adversary knows the estimation algorithm being used and the multiset of honest reviews R_H. With this knowledge and the goal of minimizing estimate correctness, it generates its reviews R_A. The estimation algorithm then receives all reviews $R = R_H \cup R_A$.

Assumption 4. The multiset of reviews R is large.

We make this final assumption so that sampling errors in the honest reviews become negligible. This lets us say that precisely a p-fraction of honest reviews will be correct on any given item, which makes later analysis easier. Though I make one more assumption later, this covers the basics of our problem setup.

1.4 Majority Vote Estimation

Since most of the reviews are correct on most of the items, it is tempting to estimate $\boldsymbol{\theta}$ by taking a majority vote on each item over all of our reviews. From assumption 2, we are guaranteed that a $p(1-\alpha)$-fraction of the reviews will be correct on each item. So, for a majority vote to work regardless of what the adversary does, we need $p > \frac{1}{2(1-\alpha)}$.
However, notice that if $\alpha \approx \frac{1}{2}$, we need $p \approx 1$ to successfully estimate $\boldsymbol{\theta}$. This is a very strong requirement of our honest reviews. How might we do better?

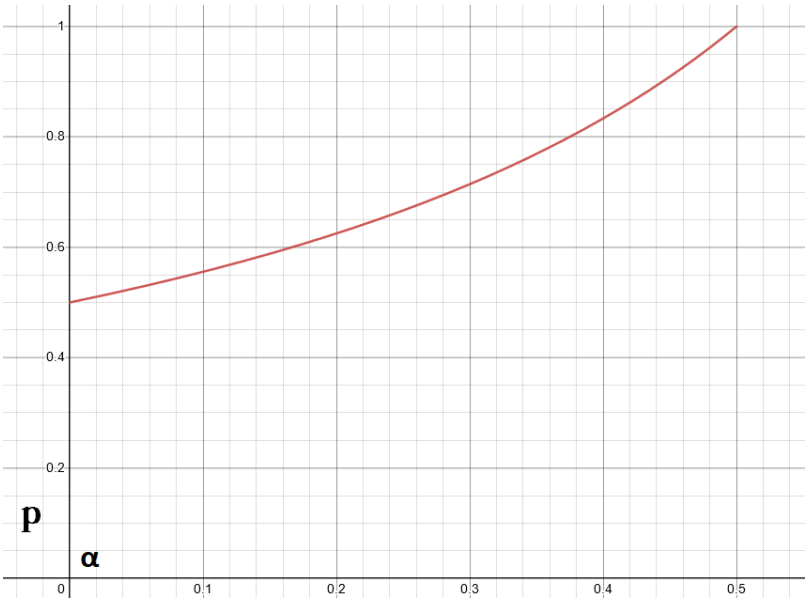

Figure 1: The graph of $p = \frac{1}{2(1-\alpha)}$ with horizontal α-axis. This is the minimum honest-review probability that guarantees successful majority estimation.

1.5 Active Learning

Active learning is a subfield of supervised learning in which an algorithm chooses which data to learn the labels of. In my research, I considered an active learning approach for improving the adversarial robustness of majority estimation. In particular, I investigated how learning a small number of the entries of $\boldsymbol{\theta}$ can help us improve our estimate and how we might choose which entries to query. Although I investigated multiple active learning rules, the one on which I spent the most time and which bore the most fruit was called the 50/50 Query.

2 50/50 Query

2.1 The Algorithm

The algorithm, which we will call `50/50`, is as follows. Given a multiset of reviews R with each review of length n,

- Compute $i := \arg\min_{j<n} |\text{Split}_R(j) - \frac{1}{2}|$. This finds the item that is closest to being a 50/50 split.
- Query the true value of that item $\boldsymbol{\theta}_i$ and define $R' := \{\mathbf{r}' \in R : \mathbf{r}'_i = \boldsymbol{\theta}_i\}$. This discards all reviews wrong on the queried item.
- Return $\text{Maj}(R')$. This takes a majority vote over the remaining reviews.

Why might we expect this to be more robust than a majority vote? First and most obvious, it counters the trivial strategy where the adversary uses all of its reviews to

give wrong answers on an item. Second, the item closest to a 50/50 split is the one we are most unsure about; by finding its true value, we are guaranteed to discover the most wrong answers and, for $p \gg \frac{1}{2}$, we are plausibly more likely to discover malicious reviews attempting to thwart estimation. With the knowledge that it is facing this algorithm, how might the adversary respond?

2.2 Adversary Strategies

How do we motivate this? I mean I can be honest and just say "this is the one I looked at." No need to continue research on this by finding out the general expression. With that in mind,

2.2.1 Sacrifice

Let q_A be the proportion of adversarial reviews that are incorrect on the sacrificial item i, and suppose that all adversarial reviews are incorrect everywhere else. For item i to actually be queried, we need the proportion of its answers which are incorrect to be closer to $\frac{1}{2}$ than for all other items (on which the adversary gives all incorrect answers), i.e.

$$|(q(1-\alpha)+q_A\alpha)-\frac{1}{2}| < |(q(1-\alpha)+\alpha)-\frac{1}{2}|. \tag{1}$$

The absolute values give two bounds on q_A, but only one is useful here[1], and that is

$$\begin{aligned}
\frac{1}{2}-q(1-\alpha)-q_A\alpha &< q(1-\alpha)+\alpha-\frac{1}{2} \\
-q_A\alpha &< 2q(1-\alpha)+\alpha-1 \\
q_A &> \frac{1-\alpha-2q(1-\alpha)}{\alpha} \\
&= \frac{(1-\alpha)(p-q)}{\alpha}.
\end{aligned}$$

For the remaining malicious reviews to successfully ruin the estimation of the $n-1$ remaining items, we need the number of remaining incorrect answers on the other items to be a majority. Since a q-fraction of honest reviews will be incorrect on i (and will thus be removed by the active learning rule), the proportion of remaining reviews which are honest is $h = p(1-\alpha)$. Therefore, the adversary needs to additionally choose q_A such that

$$\begin{aligned}
ph &< \alpha(1-q_A)+qh \\
h(p-q) &< \alpha(1-q_A) \\
q_A &< 1-\frac{h(p-q)}{\alpha} \\
&= 1-\frac{p(1-\alpha)(p-q)}{\alpha}.
\end{aligned} \tag{2}$$

[1]The other simply says $q_A < 1$, which tells us only that the adversary must give at least one correct answer on the sacrificial item.

Thus, for this attack to be possible, we need

$$\frac{(1-\alpha)(p-q)}{\alpha} < q_A < 1 - \frac{p(1-\alpha)(p-q)}{\alpha}. \tag{3}$$

By solving for equality of the upper and lower bounds, we have that, for a fixed p, this attack is only possible for

$$\alpha > 1 - \frac{1}{2p^2 + p}. \tag{4}$$

Solving in terms of p yields that, for a fixed α, this attack is only possible if

$$p < \frac{\sqrt{\frac{8}{1-\alpha} + 1} - 1}{4}.$$

From this, we find that the sacrificial attack becomes impossible (i.e. it works for no $\alpha < \frac{1}{2}$) for

$$p \geq \frac{\sqrt{17} - 1}{4} \approx 0.781. \tag{5}$$

This gives a significant improvement over simple majority estimation, which only guarantees success for $p \geq \frac{1}{2(1-\alpha)}$.

2.2.2 Multiple Sacrifice

We might expect that our estimation improves as we query more items. If it exists, can we quantify this improvement? How might the adversary's strategy change?

Essentially, the new algorithm repeats the "50/50 query then discard" step multiple times. More formally, let $n, K \in \mathbb{N}^+$ with $K \ll n$. Our new multiple query estimator is then

- Define $R_0 = R$.
- Loop K times, indexing by k starting at 0:
  - Over all unqueried indices j, compute $i_k := \arg\min_{j<n} |\text{Split}_R(j) - \frac{1}{2}|$. This finds the item that is closest to being a 50/50 split.
  - Query the true value of $\boldsymbol{\theta}_{i_k}$ and define $R_k := \{\mathbf{r}' \in R_{k-1} : \mathbf{r}'_{i_k} = \boldsymbol{\theta}_{i_k}\}$. This discards all reviews wrong on the queried item.
- Return $\text{Maj}(R_K)$. This takes a majority vote over the remaining reviews.

The goal of the adversary remains the same: to choose its reviews so that none of the final estimates are correct. In addition, the adversary must now choose the proportions of its remaining reviews incorrect on each item $q_1, q_2, \ldots, q_K$ such that q_k causes item k to be queried for all $k \leq K$. For what parameter settings $p, \alpha \in [0, 1]$ is this possible? Can we obtain bounds similar to equation (3)?

Let $k < K$ be arbitrary and define $p_k = 1 - q_k$ and $P_k = \prod_{j=1}^{k} p_j$. For item k to be queried, we compute an inequality in similar fashion to equation (1).

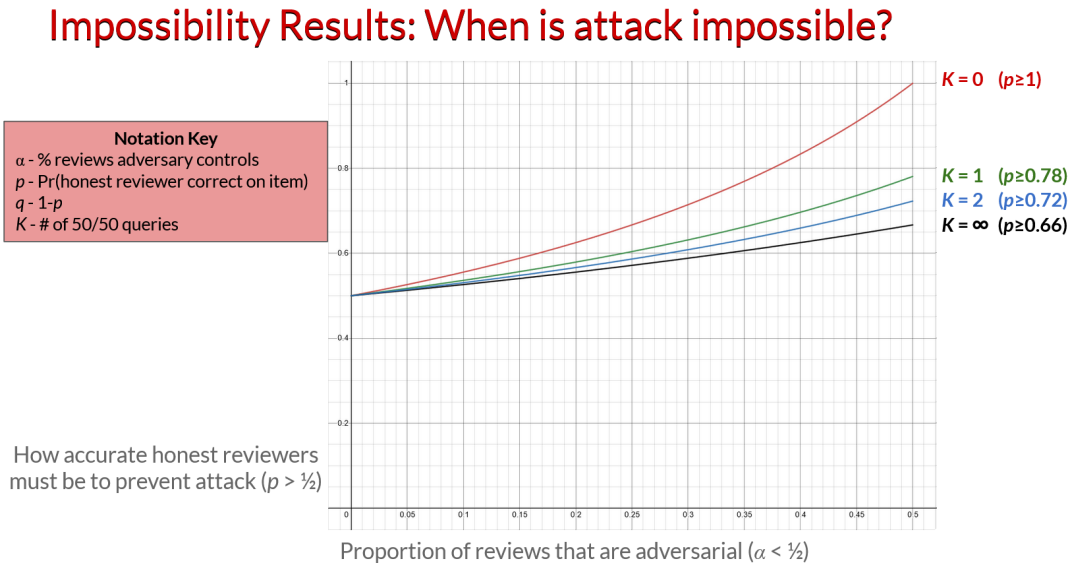

Figure 2: The analogue to Figure 1, for various values of K under the Multiple Sacrifice strategy

Note: This writeup is incomplete, so the derivation here is not included. However, empirical tests and rederivations of other bounds for $K \in \{0, 1\}$ give moderate evidence that the generalization is correct.

By reasoning similarly, finding upper and lower bounds on the choices of q_k, I found that, in order for attack to succeed, we must have

$$\alpha > 1 - \frac{q}{p(1 + p^K - 2p^{K+1})}.$$

In the limit of K, we then have that attack is impossible for $p \geq \frac{2}{3}$.