[theorem]Corollary [theorem]Example [theorem]Definition [theorem]Conjecture

1 Stochastic Dominance

Definition 1.1 (Stochastic Order)**.** *For two disjoint sets of probabilities p_i and q_i,* $\mathbf{q}$ *is said to dominate* $\mathbf{p}$ *if for any $l \in \mathbb{N}$:*

$$\sum_{i}^{l} q_i \geq \sum_{i}^{l} p_i \tag{1}$$

2 Huffman Tree Searching

Given a set S, $|S| = n$ of items, containing an unknown subset of D, $|D| = k$ of defective items, a group testing procedure to find a single defective can be defined as follows:

Algorithm 1: Construction of the Huffman decision tree

Data: A Set of n items, $k << n$ of which are defective and a probability vector $\mathbf{p}$
Result: A decision tree
Define $l := n$ (length of the tree) **while** $l > 1$ **do**
    Find the two smallest nodes in probability;
    Create a parent node by adding together the probabilities of these nodes, and appending the items from the children into a group;
    Set $l := l - 1$
end

That is, the decision tree has as external nodes the set of items and internal nodes the groups of items queries will be performed against. These groups are not equal in size, but are instead weighted by probability.

Algorithm 2: Searching the Tree for a single defective

Data: A Set of n items, $k << n$ of which are defective
Result: A single defective item
Create a decision tree as above;
Define $G = n$ (Group-size);
if $G == 2$ **then**
    Test the left item **if** *Test is positive* **then**
        The left item is defective
    else
        The right item is defective
    end
else
    Test the left branch of root node of the tree;
    if *If the test is positive* **then**
        Set the left branch as the root of the tree;
        Set $G = \#$ items in left branch
    else
        Set the right branch as the root of the tree;
        Set $G = \#$ items in right branch
    end
end

Definition 2.1 (Defective Probability)**.** *An item in the set S is independently defective with probability*

This is equivalent to the positivity of

$$\left(\sum_{k=1}^{l} p_k\theta_k\right)\left(\sum_{i=1}^{M} p_k\right) - \left(\sum_{k=1}^{l} p_k\right)\left(\sum_{i=1}^{M} p_k\theta_k\right) \tag{5}$$

$$= \sum_{k=1}^{l}\left(\sum_{i=1}^{M} p_i\left(\theta_k - \theta_i\right)\right) \tag{6}$$

$$= \sum_{k=1}^{l}\left(\sum_{i=1}^{k} p_i\left(\theta_k - \theta_i\right)\right) + \sum_{k=1}^{l}\left(\sum_{i=k+1}^{M} p_i\left(\theta_k - \theta_i\right)\right) \tag{7}$$

The first term can be re-written (after a co-ordinate swap):

$$\sum_{i=1}^{l}\sum_{k=1}^{l} p_k p_i\left(\theta_k - \theta_i\right) = \sum_{k=1}^{l}\sum_{i=k+1}^{l} p_i p_k\left(\theta_i - \theta_k\right) \tag{8}$$

Now, adding this to the second term from (7) results in:

$$\sum_{k=1}^{l} p_k\left(\sum_{i=k+1}^{M} p_i\left(\theta_k - \theta_i\right) + \sum_{i=k+1}^{l} -p_i\left(\theta_k - \theta_i\right)\right) \tag{9}$$

$$= \sum_{k=1}^{l} p_k \sum_{l+1}^{M} p_i\left(\theta_k - \theta_i\right) \tag{10}$$

$$\geq 0 \tag{11}$$

and the result follows. □

Lemma 2.1. *For a set of items S, $|S| = n$ with non-identical probabilities p_i, $i \in \{1 \ldots n\}$, such that some unspecified number, k are defectivem layed out on a Huffman tree, the above procedure will return a defective (the 'leftmost') in a number of tests T bounded by:*

$$H\left(\bar{\mathbf{p}}\right) \leq T \leq H\left(\bar{\mathbf{p}}\right) + 1 \tag{12}$$

where $H\left(\circ\right)$ is the entropy function, and

$$\bar{\mathbf{p}} = \frac{\mathbf{p}}{\sum_{i=1}^{n} p_i} \tag{13}$$

Proof. The lemma follows the dominance of $\{q_i\}$ over $\{p_i\}$ and the optimality property of Huffman codes. □

2.1 Counting the total number of tests

Lemma 2.2 (Demcomposition of the entropy)**.** *The entropy has the property, that for any* $m \in \mathbb{N}$*:*

$$H(\mathbf{p}) = H([p_1 + \ldots + p_m, p_{m+1} + \ldots + p_n]) \tag{14}$$

$$+ (p_1 + \ldots + p_m) H\left(\frac{p_1}{\sum_{i=1}^{m} p_i}, \ldots, \frac{p_m}{\sum_{i=1}^{m} p_i}\right) \tag{15}$$

$$+ (p_{m+1} + \ldots + p_n) H\left(\frac{p_{m+1}}{\sum_{i=m+1}^{n} p_i}, \ldots, \frac{p_m}{\sum_{i=m+1}^{n} p_i}\right) \tag{16}$$

Definition 2.4. *A Group Testing problem* $M(n, k, \mathbf{p})$ *is the minimum number of tests* T *required to uniquely identify all* k *defectives in a set* S *of* $|S| = n$ *items, when each item is defective with an independent probability* p_i*.*

Corollary 2.1. $\sum_{i=1}^{n} p_i = k$

Corollary 2.2. *When* $p_i = \frac{k}{n}\ \forall\ i$*,* $M(n, k, \mathbf{p}) = M(n, k)$

Lemma 2.3. *For any* $m < n$*, and* $l < m$*:*

$$M(n, k, \mathbf{p}) \leq M(m, 1, \mathbf{p}_m) + M(n - l, k - 1, \mathbf{p}_{n-l}) \tag{17}$$

where

$$\mathbf{p}_m = \left(\frac{p_1}{\sum_{i=1}^{m} p_i}, \ldots, \frac{p_m}{\sum_{i=1}^{m} p_i}\right) \tag{18}$$

and:

$$\mathbf{p}_{n-m} = \left(\frac{p_{m+1}}{\sum_{i=m+1}^{n} p_i}, \ldots, \frac{p_m}{\sum_{i=m+1}^{n} p_i}\right) \tag{19}$$

Proof. The total number of tests must be the number of tests required to find a single defective in a subset of m items, plus the number tests to find the remaining defectives in the remaining $n - m$ items. □

Conjecture 2.1.

$$M(n, k, \mathbf{p}) \leq H(\mathbf{p}) \tag{20}$$

Sketch. Using Algorithm 2 find a single defective in a set of n items. By linearity of expectation:

$$\mathbb{E}T = \mathbb{E}M(n, k, \mathbf{p}) = \mathbb{E}M(m, 1, \mathbf{p}_m) + \mathbb{E}M(n - l, k - 1, \mathbf{p}_{n-l}) \tag{21}$$

$$\leq H(\mathbf{p}_m) + H(\mathbf{p}_{n-m}) \tag{22}$$

□

This proof didn't really go anywhere, it's not really that straightforward to estimate to second term $\mathbb{E}M(n - l, k - 1, \mathbf{p}_{n-l})$. Intuitively I'd expect:

$$\mathbb{E}M(n - l, k - 1, \mathbf{p}_{n-l}) \leq \mathbb{E}M(n - m, k - 1, \mathbf{p}_{n-m}) \tag{23}$$

as we're trying to solve a sparser problem ($n - l \geq n - m$ as $l < m$). However, this needs to be proved.

Some other bounds that come to mind (not optimal):

$$\mathbb{E}M\left(n,k,\mathbf{p}\right)\leq n \tag{24}$$

i.e. in the worst case, we have to check all n items.

If we divide the n items into k groups of size $\lceil\frac{n}{k}\rceil$ and condition on the event that there is a defective in each group we have:

$$\mathbb{E}\left(M\left(n,k,\mathbf{p}\right)|1\text{ defective per group}\right)\leq kH\left(\mathbf{p}_{\lceil\frac{n}{k}\rceil}\right) \tag{25}$$

2.2 Counting in Rounds

Divide the n items into κ_1 groups (labelled $S_1\ldots S_{\kappa_1}$) s.t.

$$\begin{aligned}\Delta_i &= \mathbb{P}\left(\text{set } S_i \text{ contains no defectives}\right) &(26)\\ &= \prod_{j\in S_i}\left(1-p_j\right) &(27)\\ &\sim 1/2 &(28)\end{aligned}$$

I.e. this defines an implicit function:

$$f^{(1)}:1\ldots n\to 1\ldots\kappa_1 \tag{29}$$

let

$$p_{j,s_i}=\frac{p_j}{P_i}\ ,\ P_i=\sum_{k\in S_k}p_k \tag{30}$$

For each group S_i, do a pilot test. We search further if

$$I_i=\mathbb{1}_{\{S_i\text{ contains a defective}\}}=1 \tag{31}$$

The total number of tests we will do is then

$$\sum_{i=1}^{\kappa_1}I_i\left[H\left(p_{j,s_i}\right)+1\right] \tag{32}$$

so,

$$\mathbb{E}T^{(1)}\leq\kappa_1+\sum_{i=1}^{\kappa_1}\left(1-\Delta_i\right)\left[H\left(p_{j,s_i}\right)+1\right] \tag{33}$$

Now,

$$\begin{aligned}H\left(p_{j,s_i}\right) &= \sum_{j\in S_i}-p_{j,S_i}\log p_{j,S_i} &(34)\\ &= \frac{1}{P_i}\left[\sum_{j\in S_i}-p_j\log p_j+\sum_{j\in S_i}p_j\log P_i\right] &(35)\\ &= \frac{1}{P_i}\sum_{j\in S_i}-p_j\log p_j+\log P_i &(36)\end{aligned}$$

Where the second and third lines follow from the definition of p_{j,S_i}. We have,

$$\Delta_i \sim e^{-P_i} \sim 1/2 \tag{37}$$

Which follows from the generalised Bernoulli inequality:

$$\prod_i (1 + x_i) \geq 1 + \sum_i x_i \tag{38}$$

So,

$$P_i \sim \ln 2 \tag{39}$$

$$\log P_i \sim \log \ln 2 \tag{40}$$

Putting this all together:

$$\mathbb{E}T^{(1)} \leq \kappa_1 + \sum_{i=1}^{\kappa_1} (1 - \Delta_i) \left[H\left(p_{j,s_i}\right) + 1 \right] \tag{41}$$

$$\leq \kappa_1 + \frac{1}{2 \ln 2} \left[\sum_{j=1}^{n} -p_j \log p_j + \kappa_1 \log \ln 2 + \kappa_1 \right] \tag{42}$$

2.3 Counting in Rounds II

We arrange the tests into k rounds, organised as a sequence of $J - 1$ negative tests before a positive test. As before, label each group in round i from $1 \ldots j$. Each test tests a groups of items chosen such that

$$\Delta_i = \mathbb{P}\left(\text{set } S_i \text{ contains no defectives}\right) \tag{43}$$

$$= \prod_{m \in S_m} (1 - p_m) \tag{44}$$

$$\sim 1/2 \tag{45}$$

where p_m is the probability that item m in group j is not defective.

Denote:

$$T_j = \sum_{\text{tests in round i}} \mathbb{1}_{\text{Test j is negative}} \tag{46}$$

The number of tests in round i is:

$$T^i_{tot} = T_j + \left[H\left(p_{j,s_j}\right) + 1 \right] \tag{47}$$

$$= T_j + \frac{1}{P_i} \sum_{j \in S_i} \left[-p_j \log p_j + \log P_i + 1 \right] \tag{48}$$

The total number of tests is then:

$$T_{tot} = \sum_{i=1}^{k} T_{tot}^{i} \tag{49}$$

$$= \sum_{i=1}^{k} \left[T_j + \left[H\left(p_{j,s_j}\right) + 1 \right] \right] \tag{50}$$

$$= \sum_{i=1}^{k} T_j + \sum_{i=1}^{k} \frac{1}{P_i} \sum_{j \in S_i} \left[-p_j \log p_j + \log P_i + 1 \right] \tag{51}$$

$$= \sum_{i=1}^{k} T_j + \sum_{i=1}^{k} \frac{1}{P_i} \sum_{j \in S_i} -p_j \log p_j + \sum_{i=1}^{k} \sum_{j \in S_i} \log P_i + \sum_{i=1}^{k} \sum_{j \in S_i} 1 \tag{52}$$

Using (40) this can be simplified to:

$$T_{tot} = \sum_{i=1}^{k} T_j + \frac{1}{\ln 2} \sum_{i=1}^{k} \sum_{j \in S_i} -p_j \log p_j + \tag{53}$$

$$\log \ln 2 \sum_{i=1}^{k} \sum_{j \in S_i} + \sum_{i=1}^{k} \sum_{j \in S_i} 1 \tag{54}$$

The last sum is $k/2$ - the inner sum is over items in a single positive test - which happens with probability p_j and so is equal to $\Delta_i \sim 1/2$.

The first sum is over items which are not in positive tests, this occurs with probability $(1 - \Delta_i) \sim 1/2$. So we are left with:

$$T_{tot} = k + \frac{1}{\ln 2} \sum_{i=1}^{k} \sum_{j \in S_i} -p_j \log p_j + \tag{55}$$

$$\log \ln 2 \frac{k}{2} \tag{56}$$

Condition 1. *Given $C \geq 1$, say that a collection of numbers $\{q_1, \ldots, q_m\}$ satisfies the C-bounded ratio condition if*

$$\max_{i,j} \frac{q_j}{q_i} \leq C.$$

(For example clearly if q_i are all identical, they satisfy the condition with $C = 1$).

Fix values $\{p_1, \ldots, p_m\}$ satisfying the C-bounded ratio condition and write $P = \sum_{j=1}^{m} p_j$ and $H(p) = -\sum_{j=1}^{m} p_j \log_2 p_j$.

Lemma 2.1. *If we create a Shannon–Fano tree for the probability distribution $\overline{p}_i := p_i/P$, then each item has length bounded by*

$$\ell \leq \frac{H(p)}{P} + \log_2 C + \log_2 P + 1. \tag{57}$$

Proof. Under the C-bounded ratio condition, for any i and j, we know that

$$-\log_2 p_i \leq -\log_2 p_j + \log_2 C.$$

Multiplying by p_j and summing, we obtain that

$$-P \log_2 p_i \leq H(p) + P \log_2 C.$$

Now, the Shannon–Fano length of the ith item is

$$\begin{aligned} \ell_i &= \lceil -\log_2 \overline{p}_i \rceil \leq -\log_2 p_i + \log_2 P + 1 \\ &\leq \left(\frac{H(p)}{P} + \log_2 C \right) + \log_2 P + 1. \end{aligned}$$

□

Lemma 2.2. *Consider items $1, \ldots, m$ where (independently) item i is defective with probability p_i, where $\{p_1, \ldots, p_m\}$ are C-bounded for some C. We can recover all defective items in the set using an expected number of*

$$(1 + P \log_2 P) + P(\log_2 C + 1) + H(p) \quad \textit{tests.}$$

tests.

Proof. The algorithm works as follows: first test the whole set. If this is negative, we are done. Otherwise, search using the Shannon–Fano tree which has bounded depth given by (57) – this is guaranteed to find one defective. Remove the defective from the set, and test all the remaining items together. If this is negative, we are done, otherwise we repeat the Shannon–Fano step.

Conditional on there being U defective items in the set, this will take

$$1 + U \left(\left(\frac{H(p)}{P} + \log_2 C \right) + \log_2 P + 1 \right)$$

steps. Since the expectation of U is P, the result follows. □

Hence, given items $1, \ldots, N$, if we can divide them into g groups $S_1, \ldots, S_g$, such that for some $C > 1$ the $\{p_i : i \in S_k\}$ satisfies the C-bounded ratio condition for each $k = 1, \ldots, g$, then the expected total number of tests is

$$\sum_{k=1}^{g} (1 + P_k \log_2 P_k) + \left(\sum_{j=1}^{N} p_j \right) (\log_2 C + 1) + \left(\sum_{j=1}^{N} -p_j \log_2 p_j \right), \tag{58}$$

where we write $P_k = \sum_{i \in S_k} p_i$.

Example 2.1. *Suppose all the $p_j = K/N$, and all groups have size $m = N/(K\alpha)$, so there are $K\alpha$ groups in total. Then each $P_k = 1/\alpha$, and (58) becomes (since $C = 1$):*

$$(K\alpha - K \log_2 \alpha) + K + K \log_2(N/K).$$

Differentiating wrt α gives that the optimal $\alpha = 1/\ln 2$, meaning that the bound becomes $K \log_2(N/K) + 1.93K$, so the lead order term is correct.

In general, if we write K_{eff} for $\sum_{j=1}^{N} p_j$, can we get (58) bounded by $\left(\sum_{j=1}^{N} -p_j \log_2 p_j \right) +$ const.K_{eff} ?