Algorithm 1: Algorithm for the non-iid group testing problem

Data: A Set S of $|S| = n$ items, μ of which are actually defective in expectation, a probability vector $\mathbf{p}^{(n)}$ describing each item's independent probability of being defective, and a cutoff θ

Result: The set of defective items

Discard items with $p_i \leq \theta$

Sort the remaining items into B bins, collecting items together with $p_i \in \left[1/2C^r, 1/2C^{r-1}\right)$ in bin r.

Sort the items in each bin into sets s.t. the (normalised) probability of each set is less than $1/2$.

Test each set in turn

if *The test is positive* **then**

  Arrange the items in the set on a Shannon-Fano/Huffman Tree and search the set for all the defectives it contains

end

1 Results

The performance of the above algorithm (in terms of the sample complexity) were analysed by simulating 500 items, with a mean number of defectives equal to 8. I.e. $N = 500$ and $\mu = 8$.

The probability distribution $\mathbf{p}$ was generated by a Dirichlet distribution. We cannot simply choose a set of random numbers and normalise by the sum. Consider the case of two random numbers, (x, y), distributed uniformly on the square $[0, 1]^2$. Normalising by the sum $(x + y)$ projects the point (x, y) onto the line $x + y = 1$ and so favours points closer to $(0.5, 0.5)$ than the endpoints. The Dirichlet distribution avoids this by generating points directly on the simplex.

We then chose values of the cutoff parameter θ from 0.0001 to 0.01, and for each θ_i simulated the algorithm 1000 times. We plot the empirical distribution of tests, varying theta as well as the uniformity/concentration of the probability distribution (via the parameter α of the Dirichlet distribution). We also plot, the theoretical lower and upper bounds on the number of Tests required for successful recovery alongside the empirical number tests (all as a function of θ).

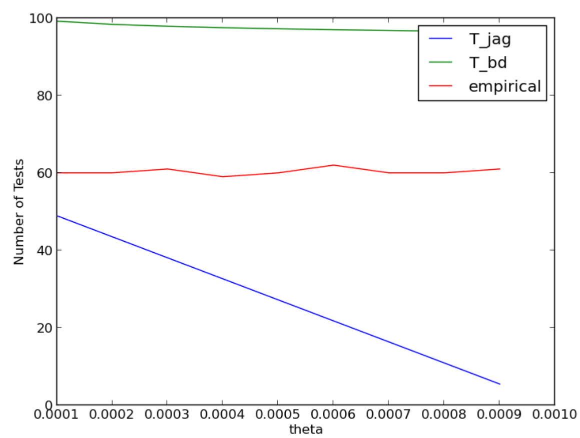

Figure 1: Theoretical lower and upper bounds and empirical Test frequencies as functions of θ

Note that the Upper bound is not optimal and there still is some room for improvement. Note also that the lower bound degrades with θ_i.

Figures (2) and (3) show that the performance is relatively insensitive to the cut-off θ, and more sensitive to the uniformity (or otherwise) of the probability distribution $\mathbf{p}$. Heuristically, this is for because distributions which are highly concentrated on a few items algorithms can make substantial savings on the testing budget by testing those highly likely items first (which is captured in the bin structure of the above algorithm).

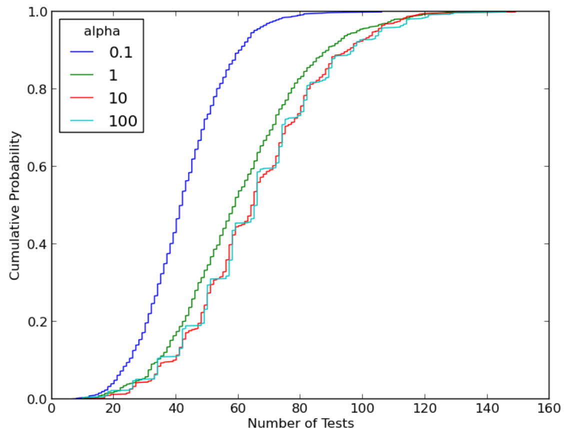

Figure 2: Cumulative distribution curves of the modified Hwang algorithm with fixed $\theta = 0.0001$ and α varying

The insensitivity to the cutoff θ is due to items below θ being overwhelmingly unlikely to be defective - which for small θ means that few items (relative to the size of the problem) get discarded.

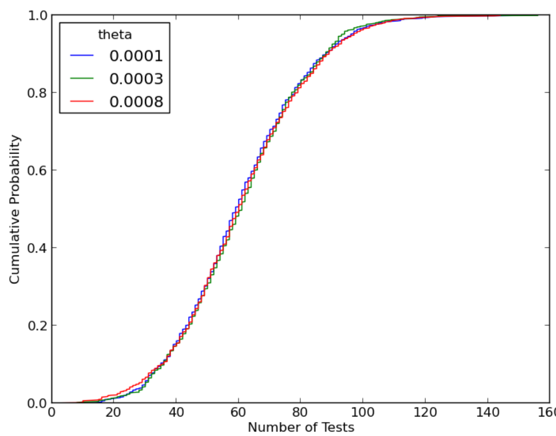

Figure 3: Cumulative distribution curves for fixed $\alpha = 1$ and varying θ