REAL-TIME DATA SELECTION AND ORDERING FOR COGNITIVE BIAS MITIGATION[0,1]

The University of Texas at Austin
Electrical and Computer Engineering
Cockrell School of Engineering

WNCG

Sara Mourad, Ahmed Tewfik

Biased Information Processing Model

❑ **Cognitive bias:** Deviation from rational judgement yielding potentially incorrect or damaging inference/decision (Tversky and Kahneman 1972).

❑ **Sources of bias:** Limited human mental processing capacity, cognitive shortcuts (heuristics), social context, emotions, etc.

❑ **Examples:** -*Confirmation bias*: searching for or interpreting information in a way that supports a preconception. Observed in many settings, e.g., trials.

-*Anchoring bias*: Over reliance on one piece of information (usually the first). Observed in many settings, e.g., negotiations.

-*Framing:* drawing conclusion from information depending on how it is presented. Observed in many settings, e.g., marketing, media, politics.

❑ **Experiment[2]:**

-Humans receive training on how to classify objects from two classes.

-Human classification performance depends on order in which items are presented to humans.

Problem formulation

❑ **Binary hypothesis detection problem:**

$$H_0\colon Y_n = W_n$$
$$H_1\colon Y_n = m + W_n$$

where $W_n \sim N(0, \delta^2)$ are i.i.d. and m is the difference in the means under the two hypothesis .

❑ **Proposed Model for human decision-making under cognitive biases:**

$$L_k = L_{k-1} + p_k l_k \quad (1)$$

where $l_k = log(\frac{f(Y_k|H_1)}{f(Y_k|H_0)}) = (\frac{2mY_k - m^2}{2\delta})$ and p_k the adjustment weight that the subject gives to the new observation.

Alice
- **Ideal reference**
- Unbiased agent
- Has access to N observations

Bob
- **Actual decision maker**
- Biased agent uses model (1)
- Has only access to N' out of N observations

Goal: *Select N' out of N total observations to show to Bob so that his decision performance is within a desired distance from Alice's.*

Find in polynomial time a subset $K \subset [N], |K| \leq N'$, such that $|T - L_{N'}|$ is minimized, where $T = \frac{\sqrt{N'}(L_N - E[L_N|H_0])}{\sqrt{N}} + E[L_{N'}|H_0]$ is the target and $L_{N'} = \sum_{i \in K} l_i$ the biased cumulative log-likelihood ratio of Bob according to equation (1).

Proposed approximate solution: an extension of the approximate subset sum algorithm

❑ **The extension of the approximate subset sum algorithm is used to find the closest L_N, to the target T**

Algorithm 1 Modified approximate subset algorithm ($\mathcal{S}_1$,T,ϵ)

```
n ← |S_1|
R_0 ← {0}
G_0 ← {0}
for i ← 1 to n do
    R_i ← MergeLists(R_{i−1}, R_{i−1} + w_j l_i)
    G_i ← MergeLists(G_{i−1}, G_{i−1} + 1)
    (R_i,G_i) ← Trim(R_i, G_i, N', ε/2n)
end for
return the closest element in R_n to T with size in G_n
    less than or equal to N'
```

❑ S={l_1, l_2,..., l_N}. S1={S,S,...,S} N' times-> To account for permutations of same subset

❑ G tracks sizes of each element in R

❑ Trimming function modified to keep only smallest size subset

❑ **Returned subset sum has restricted size $\leq N'$**

Performance guarantee – Running time

Let $L_{N'}{}^*$ be the closest subset sum to T, such that

$$\mathrm{T} - \delta/2 \leq L_{N'}{}^* \leq \mathrm{T} + \delta/2\,,$$

$$Q\left(\frac{\lambda_A - E[L_N|H_1] + \frac{\delta\sqrt{N}}{2\sqrt{N'}}}{Var[L_N]}\right) \leq P_d(\mathrm{B})\ \&\ P_f(\mathrm{B})$$

$$\leq Q\left(\frac{\lambda_A - E[L_N|H_0] - \frac{\delta\sqrt{N}}{2\sqrt{N'}}}{Var[L_N]}\right)$$

where P_d(B) (P_f(B)) is the probability of detection (false alarm) of Bob, λ_A the threshold used by Alice when using Neyman-Pearson test.

- **Proposed algorithm returns y^*:**

$$(1-\varepsilon)\, L_{N'}{}^* \leq y^* \leq (1+\varepsilon)\, L_{N'}{}^*$$

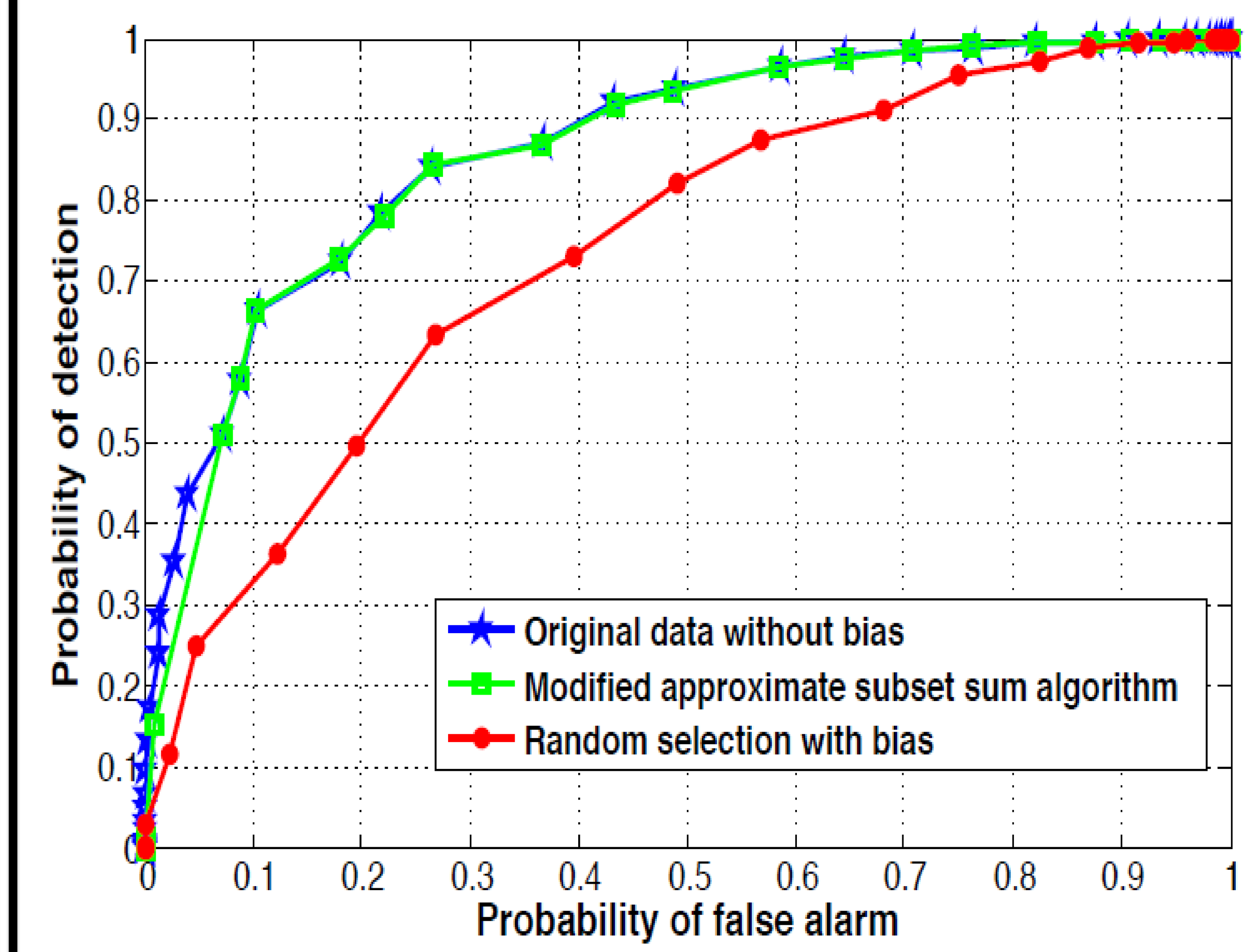
- **Running time of algorithm: $O(NN' \log N')$ and $O(1/\varepsilon)$**

[0] S. Mourad and A. Tewfik , "Cognitive Biases in Bayesian Updating and Optimal Information Sequencing" in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, April 2015.

[1] S. Mourad and A. Tewfik , "Real-Time Data Selection and Ordering for Cognitive Bias Mitigation" in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, March 2016.

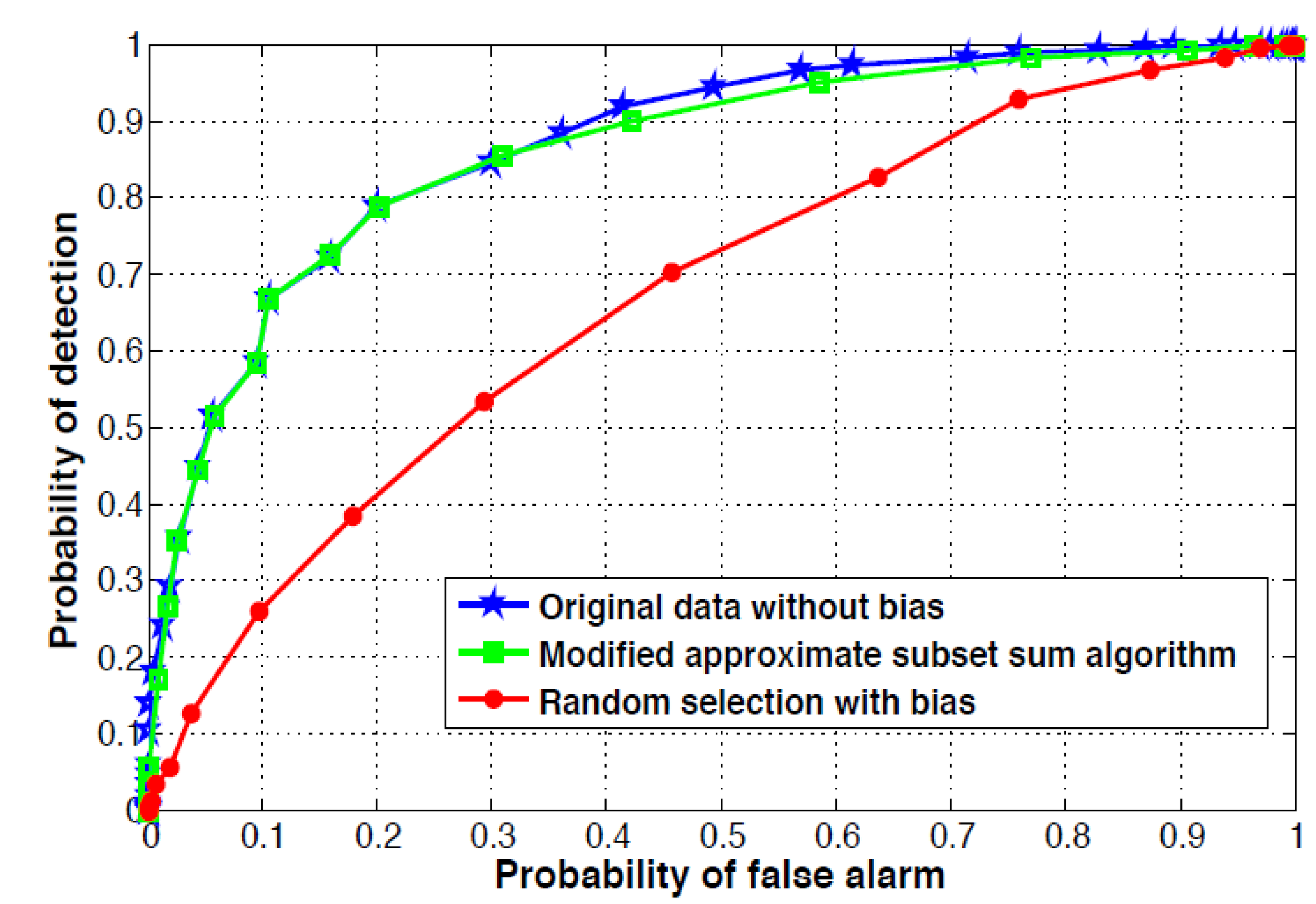
[2] Zhu, X., Gibson, B. R., Jun, K.-S., Rogers, T. T., Harrison, J., & Kalish, C. (2010). Cognitive Models of Test-Item Effects in Human Category Learning. In e 27th international conference on machine learning (ICML) (p. 158).

Results for confirmation bias towards H_0



- Near optimal performance of algorithm in region of interest
- For low probability of detection, accuracy of algorithm is more critical

Results for anchoring bias (Emphasis on first few observations)



- Near optimal performance in region of interest
- For low probability of detection and high probability of false alarm , accuracy of algorithm is more critical