Sublinear Space for a Sequential Lightbulb Problem

Noah Bergam, Berkan Ottlik, Arman Özcan

Columbia University

Problem

Detecting correlations amidst noise is a fundamental problem in machine learning and statistics. The *lightbulb problem* [6] explores this problem in a simple Boolean vector setting:

Lightbulb Problem (LBP):
- Setup: Given $x_{:,1}, \dots, x_{:,n} \in \{-1,1\}^T$. All are independently and uniformly random except the planted pair of vectors, where $x_{:,i^*} \cdot x_{:,j^*} \geq \rho T$ for some $\rho \in (0,1)$.
- Goal: find the correlated pair (i^*, j^*) efficiently → *low runtime*.

Consider a **sequential version** of the problem, where the learner receives the entries of the vectors one-by-one.

Sequential Lightbulb Problem (SLBP):
- Setup: Every day, the learner observes on/off "lightbulbs" $x_{t,:} = (x_{t,1}, \dots, x_{t,n})$. Each $x_{i,t}$ lightbulbs are independent and uniformly random except two planted bulbs i^*, j^* for which $\mathbb{E}(x_{t,i^*} x_{t,j^*}) = \rho$ for each t.
- Goal: find the correlated pair (i^*, j^*) efficiently → *low space* think streaming.

Observation: Any algorithm for the LBP can be used to solve SLBP. However, existing algorithms for LBP require linear space, and all matrix mult-based approaches require super-linear space.

Question: *Can we solve SLBP with sublinear space? What is the tradeoff between space and number of rounds?*

Previous Work

LBP is a special case of Hamming closest pair, which can be solved using locality sensitive hashing [2] in $O(n^{2-\Theta(\rho)})$ time. The runtime advantage decays with the magnitude of planted correlation. Greg Valiant [5] was the first to develop a subquadratic algorithm for the lightbulb problem where the exponent is independent of ρ. The method was based on fast matrix multiplication. Further work by [1] and [3] reduced the exponent and derandomized the algorithm.

[4] discussed SLBP but didn't consider low-space algorithms.

Fall Fourier Talks 2024 Poster Session

Algorithm 0: Naive Counter, $O(n)$ space

Upshot: Simple way to achieve linear rounds.
Idea: Maintain counter for each lightbulb that counts how many lightbulbs it agrees with over time. Planted pair will likely have highest counts.

- Learner maintains n counters for each lightbulb, $c_1 = (c_{1,1}, \dots, c_{1,n}) = \vec{0}$.
- Every day $t = 1, 2, 3, \dots, T$:
  - Learner observes $y_{i,t} \in \{\pm 1\}$ for all $i \in [n]$. Let $S_t = \sum_{j=1}^{n} y_{t,j}$
  - For each i, update $c_{t+1,i} = c_{t,i} + y_{t,i}(S_t - y_{t,i})$
- Learner outputs $(\hat{i}, \hat{j})$ indices of highest absolute-value counters.

Lemma 1: After $O(n \log(n)/\rho^2)$ rounds, the highest-value counters correspond to the planted lightbulbs with high probability.

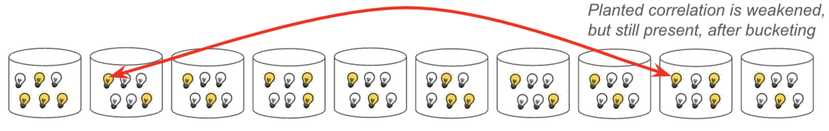
Algorithm 1: Elimination, $O(n^{1-\alpha})$ space

Upshot: Using a simple bucketing technique, we can toggle between space per round and number of rounds.

Idea: Group lightbulbs and run naive counter on the groups. Keep the top 2 groups and rerun until we can brute force.

- For every epoch $s = 1, \dots, S$:
  - Let n_s be the number of remaining lightbulbs.
    Split randomly into $m_s = n_s^{1-\alpha}$ groups of size n_s^{α}.
    Initialize m_s counters, one per group.
  - For $t = 1, \dots, T_s$: increment the counters as in the naive algorithm, but aggregated over the groups.
    At round T_s, eliminate all but the two highest-count groups.

Lemma 2: With high probability, in each "epoch," the chosen buckets keep the planted lightbulbs. Furthermore, the process terminates in $O(\log(\log n)/\log(1/\alpha))$ epochs, with $O(n^{1+\alpha})$ rounds per epoch.



Algorithm 2: Heavy Hitters, $O(\log n)$ space

Upshot: achieves constant space.
Idea: Suppose you maintain n^2 counters, one for each pair of vectors (studied in [4]). Then easily this is solved in $T = O(\log n)$ rounds, since w.h.p., planted pair will have dot product $> \rho T/2$ and the other entries $< \rho T/2$. In algorithm 2, run enough rounds until the planted pair becomes not just the maximum entry, but a ϕ-heavy hitter for some $\phi = \Theta(1)$.

Lemma 3: After $O(n^2 \log n)$ rounds, the planted pair index will be a 0.5-heavy hitter. A CountMin sketch will output this entry using $O(\log n/\phi)$ words of space.

Summary of Results

Algorithm	Space (words)	Rounds	Time/round
Alg 0: Naive	$O(n)$	$O(n \log(n)/\rho^2)$	$O(n)$
Alg 1: Elim(α)	$O(n^{1-\alpha})$	$O(n^{1+\alpha} \log(n)/\rho^2)$	$O(n)$
Alg 2: HH	$O(1)$	$O(n^2 \log(n)/\rho^2)$	$O(n)$
Alman [1]	$O(n^{1.43})$	$O(\log(n)/\rho^2)$	$O(n^{1.43})$
LSH	$O(n \log(n))$	$O(\log(n)/\rho^2)$	$O(n^{2-\Theta(\rho)})$

Table. Comparison of our online algorithms with naive applications of the standard offline algorithms.

Future directions

Both SLBP and LBP require $T = \Omega(\log(n)/\rho^2)$ rounds to achieve any fixed failure probability. This is related to established lower bounds for the Gap Hamming problem.

We would like to understand the tradeoff between space and number of rounds. Here is one basic conjecture to that end:

Conjecture: SLBP cannot be solved with $o(n)$ words of space and $O(\log n)$ rounds.

We could also try to give a lower bound on some function of the space and rounds. For instance, the tradeoffs observed in our algorithms suggests something like [space] × [rounds] = $\tilde{\Omega}(n^2)$.

Proving these tradeoffs would probably involve ideas from communication complexity. However, the standard two-player communication reduction used for the turnstile streaming model (Alice has the first half of the sequence, Bob has the second half) would not work due to the randomized nature of the data stream.

Acknowledgements

We thank Prof. Alexandr Andoni for his guidance and feedback.

References

[1] Josh Alman. An illuminating algorithm for the light bulb problem. *arXiv preprint arXiv:1810.06740*, 2018.

[2] Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, pages 604–613, 1998.

[3] Matti Karppa, Petteri Kaski, and Jukka Kohonen. A faster subquadratic algorithm for finding outlier correlations. *ACM Transactions on Algorithms (TALG)*, 14(3):1–26, 2018.

[4] Ramamohan Paturi, Sanguthevar Rajasekaran, and John Reif. The light bulb problem. *Information and Computation*, 117(2):187–192, 1995.

[5] Gregory Valiant. Finding correlations in subquadratic time, with applications to learning parities and juntas. In *2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*, pages 11–20. IEEE, 2012.

[6] Leslie G Valiant. Functionality in neural nets. In *Proceedings of the Seventh AAAI National Conference on Artificial Intelligence*, pages 629–634, 1988.