CS5234 - Algorithms at Scale

Liew Zhao Wei

Semester 1, 2023-2024

---

1 Probability and Bounds

WIP.

1. Union bound, linearity of expectation 2. Markov, Chebyshev, Chernoff, Hoeffding 3. k-universal hash family (include space analysis)

2 Simple Techniques

WIP. 1. Reservoir sampling 2. Mean trick to drive down variance 3. Median trick to boost success Probability 4. Median of mean trick (2 and 3) to bound concetration

3 Sketches

1. combining sketches, linear sketches 2. Misra-Gries 3. Count-Min-Sketch and Count-Sketch

4 Dimensions and Distances

Lemma 4.1 (Johnson-Lindenstrauss Lemma)
For any set $S \subseteq \mathbb{R}^d$ of n-points, there is an embedding $f \colon \mathbb{R}^d \to \mathbb{R}^m$ for $m = O(\epsilon^{-2} \log n)$ such that

$$\forall u, v \in S \quad (1-\epsilon)\|u-v\|_2^2 \leq \|f(u)-f(v)\|_2^2 \leq (1+\epsilon)\|u-v\|_2^2 \tag{1}$$

In other words, we can embed S into a lower-dimensional space while approximately preserving ℓ_2 norms.

Some observations:

- The embedding has only a logarithmic dependence on n and *no* dependence on d.
- The embedding is can be generated using a Gaussian distribution.
- The embedding can be represented as a linear transformation, or in other words, a matrix.

Definition 4.2 (Locality Sensitive Hash)
A hash family $\mathcal{H} = \{h \colon \mathcal{U} \to S\}$ is a (r_1, r_2, p_1, p_2)-locally sensitive if for all points $p, p' \in \mathcal{U}$,

1. if $d(p, p') \leq r_1$, then $\Pr_{h \in \mathcal{H}}[h(p) = h(p')] \geq p_1$,
2. if $d(p, p') > r_2$, then $\Pr_{h \in \mathcal{H}}[h(p) = h(p')] \leq p_2$.

In other words, a *locality sensitive hash* (LSH) is a hash family where similar items are more likely to collide. Note that the definition makes sense only if $r_1 < r_2$ and $p_1 > p_2$.

WIP. 1. ANN, PLEB, how to solve them