CSCE 689: Special Topics in Modern Algorithms for Data Science Fall 2023

Lecture 7 — September 6, 2023

Prof. Samson Zhou *Scribe: Bokun Wang*

In this lecture, we first wrap up the probability section by revisiting the query time and memory slots needed to avoid collisions in the hashing problem. After that, we introduce the "coupon collector" problem. Last, move to the dimensionality reduction section and show some preliminaries.

1 Wrapping up the Probability Section

1.1 Revisiting collision avoidance in the Hashing problem

By the "birthday paradox", we require $\Theta(n^2)$ memory slots[1] to avoid collisions with high probability (say 0.99) when hashing n items. The query runtime $\times$ memory of this strategy is $\Theta(n^2) \times O(1)$. A natural question is

Can we do better than $\Theta(n^2) \times O(1)$?

It is possible to trade query runtime for memory. In Lecture 2, we propose another strategy to avoid collisions via the linked list: We use only $\Theta(n)$ memory slots and we store multiple items in the same slot as a linked list. So far, it is still unclear what is the query runtime (i.e., maximum number of collisions in a location) of this linked-list strategy. A wild guess is that the query time is $O(n)$ such that the smaller query time $\times$ memory is not improved. Next, we show a *sharper* upper bound of the query runtime by recalling the "Max Load" problem discussed in Lecture 6.

In Lecture 6, we showed that the largest number of collisions in each slot is $O(\log n)$ with high probability. Thus, query runtime $\times$ memory of this linked-list strategy is $\Theta(n) \times O(\log n)$.

1.2 The "Coupon Collector" problem

Suppose we have a fair n-sided die and we roll n times. How many times should we roll the die before we see all possible outcomes among the rolls?

For a fixed $k \in [n]$, define the random variable $X_i = 1$ if the i-th roll is k and $X_i = 0$ otherwise, so that $\mathbb{E}[X_i] = \frac{1}{n}$. The total number of rolls with value k is $X = X_1 + \ldots + X_n$. Note that $\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X_i] = 1$. We can apply the Chernoff bound to X. For any $\delta > 2$,

$$\Pr[X \geq (1+\delta)] \leq 2\exp\left(-\frac{\delta^2}{2+\delta}\right) < 2\exp(-\delta/2).$$

Choose $\delta = 4\log n + 2$. Then, $\Pr[X \geq 4\log n + 3] \leq \frac{1}{n^2}$. By the union bound, no outcome will be rolled more than $4\log n + 3$ times with probability $1 - \frac{1}{n}$. So the answer to the "coupon collector" problem is $O(n \log n)$.

[1] Each slot is not the head of a linked list.

2 Dimensionality Reduction

Challenges for algorithms in the "Big Data Era": (1) A large amount of data points; (2) Many measurements per data point, i.e., very high dimensional data. This section focuses on reducing the dimension of data points.

Consider the data matrix $X \in \mathbb{R}^{n \times d}$, where n is the number of data points and d is the dimension of each data point. Each row $x_i \in \mathbb{R}^d$, $i \in [n]$ corresponds to a data point. Dimensionality reduction methods transform the data points so that they have much smaller dimension, i.e.,

$$x_1, \ldots, x_n \in \mathbb{R}^d \to y_1, \ldots, y_n \in \mathbb{R}^m \quad \text{for} \quad m \ll d.$$

Next, we introduce a notion that evaluates the quality of a dimensionality reduction.

Definition (Low distortion embedding)**.** Given data points $x_1, \ldots, x_n$, a distance function D, and an accuracy parameter $\varepsilon \in [0, 1)$, a low-distortion embedding of $x_1, \ldots, x_n$ is a set of points $y_1, \ldots, y_n$, and a distance function D' such that for all $i, j \in [n]$

$$(1 - \varepsilon) D(x_i, x_j) \leq D'(y_i, y_j) \leq (1 + \varepsilon) D(x_i, x_j).$$

An example of embedding that does not have any distortion: Suppose that data points $x_1, \ldots, x_n$ lie on the first axis. Let D, D' be the Euclidean distance $\|\cdot\|_2$ and y_i be the first coordinate of x_i such that $\|x_i - x_j\|_2 = \|y_i - y_j\|$ for any $i, j \in [n]$ ($\varepsilon = 0$).