ECE 6980
An Algorithmic and Information-Theoretic Toolbox for Massive Data

Instructor: Jayadev Acharya — Lecture #1
Scribe: Huanyu Zhang — 23rd August, 2017

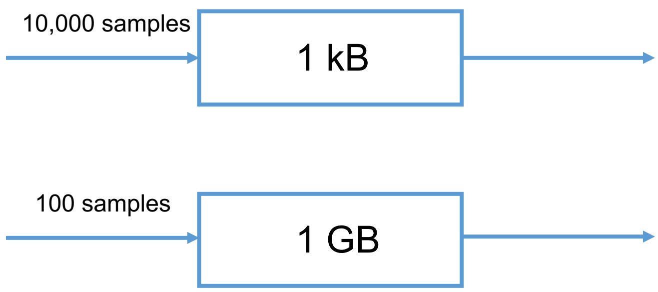
1 Introduction

Given samples from a probabilistic source, we want to infer something about the process. What resources do we need to solve a problem?

1 How much data (sample complexity)

2 How much time (computational complexity)

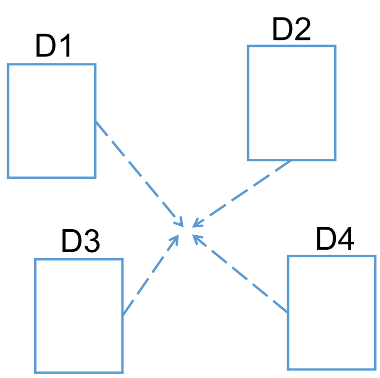
3 Space complexity



4 Robustness

5 Privacy

6 Communication



2 Learning discrete distributions

Definition 1. *A discrete distribution p over $\mathcal{X}$ is a function from a countable set $\mathcal{X}$ to $\mathbb{R}_+$, such that $\sum_{x\in\mathcal{X}} p(x) = 1$.*

Problem: Let p be an unknown distribution with alphabet size k ($|\mathcal{X}| = k$). Let $X_1, X_2, ...X_n$ be independent samples drawn from p, we want to output $\hat{p}$ such that p and $\hat{p}$ are close. In fact, we can never learn P exactly, even with infinite samples. The idea is that the set of all the sequence of samples is countable but the set of all the distribution is uncountable. It is impossible to deduce an uncountable set by a countable set.

2.1 Total Variation/ℓ_1 distance

For a subset $A \subseteq \mathcal{X}$, let $p(A) = \sum_{x\in A} p(x)$ be the probability of observing an element in A.

Definition 2. *The total variation distance between p, and q is*

$$d_{TV}(p, q) = \sup_{A\subseteq\mathcal{X}} |p(A) - q(A)|$$

Definition 3. *The ℓ_1 distance between p, and q is*

$$\ell_1(p, q) = \sum_{x\in\mathcal{X}} |p(x) - q(x)|$$

2.2 Problem Setting

Let $X_1, X_2, ...X_n$ be independent samples drawn from p, we want to output $\hat{p}$ s.t. with probability at least 0.9, $d_{TV}(p, \hat{p}) < \varepsilon$. Our question is how large n should be.

Suppose we observe $X_1^n \stackrel{\text{def}}{=} X_1, X_2, ...X_n$ from a distribution p over $\mathcal{X}$. Let

$$N_x \stackrel{\text{def}}{=} \{\#\ \textit{times symbol } x \textit{ appears in } X_1^n\}$$

We define the empirical estimator $\hat{p}(x) = \frac{N_x}{n}$.

For example, let $X_1^n = T\ H\ H\ T\ T\ T$, $N(H) = 2$, $N(T) = 4$ and $\hat{p}(H) = \frac{2}{6}$, $\hat{P}(T) = \frac{4}{6}$.

Theorem 4. *Empirical estimator solves the problem using $C \cdot \frac{k}{\varepsilon^2}$ number of samples, where C is some constant.*