CS5120: Probability & Computing **Jan-May 2020**

Lecture 13: Mean Estimation from Samples

Lecturer: N.R.Aravind *Scribe: N.R.Aravind*

1 The Problem

We have a random variable X, whose distribution we don't know. We'd like to find $\mu = E[X]$ from a set of n independent random samples $X_1, \ldots, X_n$.

Our goal is to output a number z such that:

$$Pr[|z - \mu| > \varepsilon\mu] < \delta.$$

We shall call this an (ε, δ) approximation. For given ε, δ, we'd like to know:

1. What is the number n of samples that suffices?
2. How should we use these sampes to find the approximate value?

The second question appears to have an easier answer: we use the average value of the samples as our estimate, that is: $\frac{X_1 + \ldots + X_n}{n}$. In the next section, we calculate the number of samples needed with this answer, as a function of the approximation and error guarantees.

2 The Mean

We prove two bounds on the sample size, based on assumptions about X.

Proposition 1 *If $E[X] = \mu > 0$ and X takes values in an interval of length $c\mu$, then the average of $\lceil c^2 \frac{1}{\varepsilon^2} \log\left(\frac{2}{\delta}\right) \rceil$ samples is an (ε, δ)-approximation.*

For the second bound, instead of assuming that the range of X is bounded, we assume that $Var[X]$ is bounded. Note that if X takes values in an interval of length $c\mu$, then $Var[X] \leq c^2\mu^2/4$ (but the converse is not necessarily true); thus we now make a weaker assumption.

Proposition 2 *If $\sigma[X] \leq cE[X]$, then the average of $\lceil c\frac{1}{\varepsilon^2}\frac{1}{\delta}\rceil$ samples is an (ε, δ) approximation.*

Note that the dependence on δ is too large here, compared to that in Proposition 1. Our goal in the next two sections is to show that even with the weaker assumption of Proposition 2, it is sufficient to have as few samples as in Proposition 1. We now prove the two propositions.

Proof of Proposition 1 Let $Y = \frac{X_1 + \ldots + X_n}{n}$. We have $E[Y] = \mu$. Suppose that X is defined on an interval of length L. From the Chernoff bound (Theorem 5) for the average of n independent variables, we have:

$$Pr[|Y - \mu| > \varepsilon\mu] < 2e^{-2n\varepsilon^2\mu^2/L^2}.$$

We want the RHS to be less than δ. We can rewrite the desired inequality as: $-2n\varepsilon^2\mu^2/L^2 < \log(\delta/2)$, which yields:

$$n > \frac{L^2}{\varepsilon^2\mu^2}\log(\frac{2}{\delta}).$$

By setting $L = c\mu$, we get the desired bound. ■

,

Proof of Proposition 2 We have $Var[Y] = \frac{Var[X]}{n} \leq \frac{c^2\mu^2}{n}$. Thus, by Chebyshev's inequality, we have:

$$Pr[|Y - \mu| > \varepsilon\mu] \leq \frac{c}{\varepsilon^2 n}.$$

Thus, for this probability to be less than δ, the value stated in the proposition suffices. ■

3 The Median

Let M be the median of $X_1, \ldots, X_n$. Can M be a good estimate of the mean? We will see that if X is well-concentrated, in the sense of the definition below, then M is a good estimate.

Definition We say that X is ε-concentrated if $P[|X - E[X]| > \varepsilon E[X]] \leq 1/4$.

In the above definition, instead of 1/4, any fixed number less than 1/2 would also work, for an estimate such as below.

Lemma 3 *If X is ε-concentrated and $X_1, \ldots, X_n$ are independent samples of X, then the median of $X_1, \ldots, X_n$ is an (ε, δ)-approximation of $E[X]$ for $n \geq 32 \log(\frac{1}{\delta})$.*

Proof Let $\mu = E[X]$ and M be the median of $X_1, \ldots, X_n$. We define random variables $Y_1, \ldots, Y_n$ such that $Y_i = 1$ if $|X_i - \mu| \leq \varepsilon\mu$ and $Y_i = 0$ otherwise. Let $Y = Y_1 + \ldots + Y_n$. Then $|M - \mu| > \varepsilon\mu$ if and only if $Y \leq n/2$.

We are given that $Pr(Y_i = 1) \geq 1/2 + \alpha$ for $\alpha = 1/4$. Then we have, from the Chernoff-bound analysis of majority voting:

$Pr[Y \leq n/2] < \delta$ for $n = O\left(\frac{1}{\alpha^2} \log(1/\delta)\right)$.

Since $\alpha = 1/4$ is a constant, the lemma follows. ■

We remark that Lemma 3 is useful to reduce the probability of error while maintaining the approximation guarantee.

4 Median-of-means

We now return to our original problem, and note that the given variable X may not be ε-concentrated. Thus, our goal is to produce a random variable Y which has the same mean as X and is ε-concentrated. We will then take $O\left(\log(\frac{1}{\delta})\right)$ independent copies of Y and output the median of these copies. For this, we simply choose Y to be the average of a number of samples. Note that our goal is only to get an ε-approximation with probability of error at most 1/4. Thus, by Proposition 2, $\lceil \frac{16c}{\varepsilon^2} \rceil$ samples suffice for this. Thus, we obtain the following.

Proposition 4 *Let X be a random variables such that $\sigma[X] \leq cE[X]$. Let $r = \left\lceil 32 \log\left(\frac{1}{\delta}\right)\right\rceil$, $s = \left\lceil \frac{16c}{\epsilon^2} \right\rceil$, and let $\{Y_{i,j} : 1 \leq i \leq r, 1 \leq j \leq s\}$ be independent samples of X. Let $Y_i = \frac{Y_{i,1} + \ldots + Y_{i,s}}{s}$ for $i = 1, \ldots, r$ and let Y be the median of $Y_1, \ldots, Y_r$. Then Y is a (ε, δ)-approximation of $E[X]$.*

5 The Chernoff Bound

Theorem 5 *Let $\{X_i\}_{i=1}^n$ be a set of independent random variables in $[a_i, b_i]$ and $L = \sum_{i=1}^n (b_i - a_i)^2$. Let $X = \sum_{i=1}^n X_i$ with $E[X] = \mu$. Then, for every $t > 0$, we have:*

$$Pr[|X - \mu| > t] < 2e^{-2t^2/L}.$$

Equivalently, $Pr[|X/n - \mu/n| > t] < 2e^{-2n^2t^2/L}$.