COMP 677: Estimation of Entropy in Constant Space

Lecture 2
Maryam Aliakbarpour
Fall 2023
Today’s lecture

• House keeping items
• Concentration of random variables
• Estimation of Entropy in Constant Space
• Feedback form
Class project

- Projects types:
  - Survey (4 papers)
  - Research
- Abstract: Due 9/13 (in two weeks)
  - One page
  - The topic of focus
- Progress report: Due 10/18
  - Mid-point evaluation
  - 3-page report
- Final project: Due 11/29
  - 8-page final report

- Project presentation
Next week

Paper:

When is Memorization of Irrelevant Training Data Necessary for High-Accuracy Learning?

Reading assignment: Due 9/6 before 4pm.
Concentration of random variables
Entropy estimation in constant space

Joint work with Andrew McGregor (Umass Amherst), Jelani Nelson (UC Berkeley), Erik Waingarten (Penn)
Estimation with memory constraints

Unknown distribution $D$

Goal: Estimate $f(D)$ with error $\epsilon$ with probability $1 - \delta$ via samples
- (e.g., mean, variance, etc.)

Image from: https://tilics.dmi.unibas.ch/the-turing-machine
Estimation with memory constraints

Unknown distribution $D$

Goal: Estimate $f(D)$ with error $\varepsilon$ with probability $1 - \delta$ via samples
- (e.g., mean, variance, etc.)

How many samples do we need to achieve certain amount of error with limited memory?

Image from: https://tilics.dmi.unibas.ch/the-turing-machine
This work: estimating entropy

Shannon’s entropy of $D = (p_1, p_2, \ldots, p_n)$:

$$H(D) := \sum_{x=1}^{n} p_x \log_2 \frac{1}{p_x}$$

Entropy of a binary random variable

In information theory, the entropy of a random variable is the average level of "information", "surprise", or "uncertainty" inherent to the variable's possible outcomes. Wikipedia
This work: estimating entropy

Shannon’s entropy of $D = (p_1, p_2, ..., p_n)$:

$$H(D) := \sum_{x=1}^{n} p_x \log_2 \frac{1}{p_x}$$

Used in practice to measure randomness

Applications:
- Dataset summarization
- Data compression
- Evaluating language models
- Clustering and classification
Problem definition

Shannon’s entropy of $D = (p_1, p_2, ..., p_n)$:

$$H(D) := \sum_{x=1}^{n} p_x \log_2 \frac{1}{p_x}$$

Goal:

$$\Pr[|\hat{H} - H(D)| \leq \epsilon] \geq 0.9$$

Memory constraint: $O(1)$ words of memory ($\text{Polylog}(n, 1/\epsilon)$ bits)
Our results

There exists an algorithm for the entropy estimation problem that uses $O(1)$ words ($Polylog(n, 1/\epsilon)$ bits) of memory and

$$0\left(\frac{n \log(1/\epsilon)^4}{\epsilon^2}\right)$$ samples.

$\Theta\left(\frac{n}{\epsilon \log n} + \frac{\log^2 n}{\epsilon^2}\right)$ samples with no memory constraint

$0\left(\frac{n \log (1/\epsilon)^3}{\epsilon^3}\right)$ samples with $O(1)$ words of memory

$n = \text{domain size}$
$\epsilon = \text{error}$

A closely related model: streaming algorithms

Distribution $D$ over $[n] \quad \rightarrow \quad x_1, x_2, \ldots, x_m \quad \rightarrow \quad \hat{f}$

This talk: Properties of the distribution

$x_1, x_2, \ldots, x_m \quad \rightarrow \quad \hat{f}$

Algorithm with limited memory

Properties of the data stream itself
Our results

There exists an algorithm for the entropy estimation problem that uses $O(1)$ words ($\text{Polylog}(n, 1/\epsilon)$ bits) of memory and $O\left(\frac{n \log(1/\epsilon)^4}{\epsilon^2}\right)$ samples.

Note: Estimating the empirical entropy of the stream can NOT be done in $O(1)$ words of memory.

$\Omega\left(\frac{1}{\epsilon^2} \cdot (\log \log n + \log 1/\epsilon)\right)$ bits

$n =$ domain size
$\epsilon =$ error

[A, McGregor, Nelson, Waingarten’22]

[Chakrabarti, Cormode, McGregor’10]

[Jayaram Woodruff’19]
Techniques
No memory constraint

Algorithm [Valiant, Valiant’11]:
1. Compute the fingerprint of the samples

List: 

- Frequency = 1: 3 elements
- Frequency = 2: 1 element
- Frequency = 3: 3 elements
- Frequency = 4: 8 elements
- Frequency = 5: 7 elements
- Frequency = 6: 3 elements
- Frequency = 7: 1 element
- Frequency = 8: 5 elements
No memory constraint

Algorithm [Valiant, Valiant’11]:
1. Compute the fingerprint of the samples
2. Come up with a histogram of a distribution that is likely to generate

Plots from [Valiant, Valiant’11]
No memory constraint

Algorithm [Valiant, Valiant’11]:
1. Compute the fingerprint of the samples
2. Come up with a histogram of a distribution that is likely to generate
3. Output a distribution that is compatible with the histogram

Works well ignoring the labels!  
Entropy
Support size

Requires memorizing all the samples
Entropy estimation with no memory constraint

A simple approach
How? Take average

$$H(D) := \sum_{x=1}^{n} p_x \cdot \log \frac{1}{p_x} = E_{x \sim D} \left[ \log \frac{1}{p_x} \right]$$

$n = \text{domain size}$

$\epsilon = \text{error}$

$p_{x_i}$'s are unknown 😞

$$\log \frac{1}{p_{x_1}} \quad \log \frac{1}{p_{x_2}} \quad \ldots \quad \log \frac{1}{p_{x_r}}$$

$$\frac{1}{r} \sum_{i=1}^{r} \log \frac{1}{p_{x_i}} \quad \text{large } r \quad \rightarrow \quad E_{x \sim D} \left[ \log \frac{1}{p_x} \right] = H(D)$$
Fix $m$. Count $i$’s in next $m$ samples.

Set $\hat{p}_x = \frac{\text{# instances}}{m}$

In the example: $\frac{2}{6}$
How? Take average

\[ H(D) := \sum_{x=1}^{n} p_x \cdot \log \frac{1}{p_x} = E_{x \sim D} \left[ \log \frac{1}{p_x} \right] \]

\[
\begin{align*}
\log \frac{1}{\hat{p}_{x_1}} & \quad \log \frac{1}{\hat{p}_{x_2}} & \quad \ldots & \quad \log \frac{1}{\hat{p}_{x_r}} \\
\end{align*}
\]

\[
\frac{1}{r} \sum_{i=1}^{r} \log \frac{1}{\hat{p}_{x_i}} \quad \text{large } r \quad H(D)
\]

\[
\epsilon = \text{error}
\]

\[
\frac{H(D)}{n} = \text{domain size}
\]
How? Take average

\[
\frac{1}{r} \sum_{i=1}^{r} \log \frac{1}{\hat{p}_{x_1}} \xrightarrow{\text{large } r} E_{x \sim D} \left[ \log \frac{1}{\hat{p}_{x}} \right] \xrightarrow{\text{large } m} E_{x \sim D} \left[ \log \frac{1}{p_x} \right] = H(D)
\]

Error of estimation

Bias

\[
E[\text{#samples}] = \Theta(r \cdot m) = \Theta \left( \frac{n \log \left( \frac{n}{\epsilon} \right)}{\epsilon^3} \right)
\]

\(n = \text{domain size}\)

\(\epsilon = \text{error}\)
Analysis of error

Error: $|H(D) - \hat{H}| \leq \epsilon$

$|H(D) - \hat{H}| \leq |H(D) - E[\hat{H}_i]| + |E[\hat{H}_i] - \hat{H}|

$\leq |E_{i \sim D} \left[ \log \frac{1}{p_i} \right] - E_{i \sim D} \left[ \log \frac{1}{\hat{p}_i} \right]| + |E[\hat{H}_i] - \hat{H}|

Bias  Error of estimation

$m > \Omega(n/\epsilon)$ implies bias $< \epsilon/2$  $r = \Theta(\log m/\epsilon^2)$ implies that error $< \epsilon/2$

$E[\#\text{samples}] = \Theta(r \cdot m) = \Theta \left( \frac{n \log \left( \frac{n}{\epsilon} \right)}{\epsilon^3} \right)$
Simple algorithm [Plug-in estimator]

\[ H(D) := \sum_{i=1}^{n} p_i \cdot \log \frac{1}{p_i} = \mathbb{E}_{i \sim D}[\log \frac{1}{p_i}] \]

1. Repeat \( r \) times
   1. Draw \( i \sim D \).
   2. \( \hat{p}_i \leftarrow \text{Estimate } p_i \)
   3. \( \hat{H}_i \leftarrow \log \frac{1}{\hat{p}_i} \)

2. Output: \( \hat{H} := \frac{1}{r} \sum_{i=1}^{r} \hat{H}_i \)

Fix \( m \). Count \( i \)'s in next \( m \) samples.

\#instances of \( i \sim \text{Bin}(m, p_i) \)

Set \( \hat{p}_i = \frac{\# \text{ instances}}{m} \)

In the example: \( \frac{2}{6} \)

\[ i \quad i \quad i \quad \text{ and others} \]
Simple algorithm

\[ H(D) := \sum_{i=1}^{n} p_i \cdot \log \frac{1}{p_i} = \mathbb{E}_{i \sim D}[\log 1/p_i] \]

1. Repeat \( r \) times
   1. Draw \( i \sim D \).
   2. \( \hat{p}_i \leftarrow \text{Estimate } p_i \)
   3. \( \hat{H}_i \leftarrow \log 1/\hat{p}_i \)

2. Output: \( \hat{H} := \frac{1}{r} \sum_{i=1}^{r} \hat{H}_i \)

Fix \( m \). Count the number of instances of \( i \) in the next \( m \) samples.

\[ \#\text{instances of } i \sim \text{Bin}(m, p_i) \]

Set \( \hat{p}_i = \frac{\# \text{ instances}}{m} \)

In the example: \( \frac{2}{6} \)

[Diagram of circles with instances marked with 'i']
Idea I: Estimate via negative binomials

Count the number of samples until $t$ instances of $x$ are observed.

\[ \#\text{samples} \sim \text{Negative Bin} \left( t, p_x \right) \]

Set $X_x = \frac{\#\text{samples}}{t}$

$\mathbb{E}[X_x] = \frac{1}{p_x}$

In the example for $t = 2$ : $X_x = \frac{7}{2}$
Analysis of error

Error: $|H(D) - \hat{H}| \leq \epsilon$

$|H(D) - \hat{H}| \leq |H(D) - E[\hat{H}_i]| + |E[\hat{H}_i] - \hat{H}|$

$\leq E_{i \sim D} \left[ \log \frac{1}{p_i} \right] - E_{i \sim D} \left[ \log \frac{1}{\hat{p}_i} \right] + |E[\hat{H}_i] - \hat{H}|$

Bias

Error of estimation

$t = \Theta(1/\epsilon) \implies \text{bias} < \epsilon/2$

$r = \Theta(\log^2 n/\epsilon^2) \implies \text{error} < \epsilon/2$

$E[\#\text{samples}] = \Theta(r \cdot t \cdot n) = \Theta(n \log^2 n/\epsilon^3)$

$n = \text{domain size of the distribution}$

$\epsilon = \text{error parameter}$

$r = \text{number of rounds}$

$t = \text{number of observed instance of } i$
Idea II: Remove bias

Idea: Estimate bias and subtract it from $\hat{H}$.

Let $Y_i \leftarrow p_i X_i$

Bias = $|E_{i \sim D}[\log 1/p_i] - E_{i \sim D}[\log X_i]| = |E_{i \sim D}[\log Y_i]|$

$E_{i \sim D}[Y_i] = 1$. Taylor expansion around $Y = 1$:

Bias = $E_{i \sim D}[\log Y_i] = E \left[ Y_i - 1 - \frac{(Y_i-1)^2}{2} + \frac{(Y_i-1)^3}{3} - \ldots \right]$

$n =$ domain size of the distribution
$\epsilon =$ error parameter
$r =$ number of rounds
$t =$ number of observed instance of $i$
$X_i =$ number of samples to see $t$
instance of $i$
$E[X_i] = 1/p_i$
Idea II: Remove bias

Idea: Truncated Taylor expansion. Keep the first $s = \log(1/\epsilon)$ terms.

$$\text{Bias} < \mathbb{E}\left[ (Y_i - 1)^{s+1} \right]$$

Reduce $t$ to $O(\text{polylog}(1/\epsilon))$. ✔️

Polynomial of degree $s$ of $p_i$

Pr[k samples are equal] = $p_i^k$
Idea III: Remove log $n$ factors

Idea: Bucketing

Partition the range of $X_i$ into $L$ intervals

$$E_{i \sim D} \left[ \log X_i \right] = \sum_{\ell=1}^{L} \Pr[X_i \in I_\ell] \cdot E[\log X_i | X_i \in I_\ell]$$

Estimate $\hat{q}_L$ and $\hat{H}_L$

$n = \text{domain size of the distribution}$
$\epsilon = \text{error parameter}$
$r = \text{number of rounds}$
$t = \text{number of observed instance of } i$
$X_i = \text{number of samples to see } t \text{ instance of } i$
$E[X_i] = 1/p_i$
Idea III: Remove \( \log n \) factors

\[
\text{Error} \leq \left| \sum_{\ell=1}^{L-1} (\hat{q}_\ell - q_\ell) \cdot (H_\ell - H_L) \right| + \left| \sum_{\ell=1}^{L} q_\ell \cdot (H_\ell - \hat{H}_\ell) \right|
\]

Buckets of large \( X \) can be computed with less accuracy.

Removing \( O(\log n) \).

- \( n \) = domain size of the distribution
- \( \epsilon \) = error parameter
- \( r \) = number of rounds
- \( t \) = number of observed instance of \( i \)
- \( X_i \) = number of samples to see \( t \) instance of \( i \)
- \( E[X_i] = 1/p_i \)