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 ϵ with probability $1-\delta$ via samples

- (e.g., mean, variance, etc.)

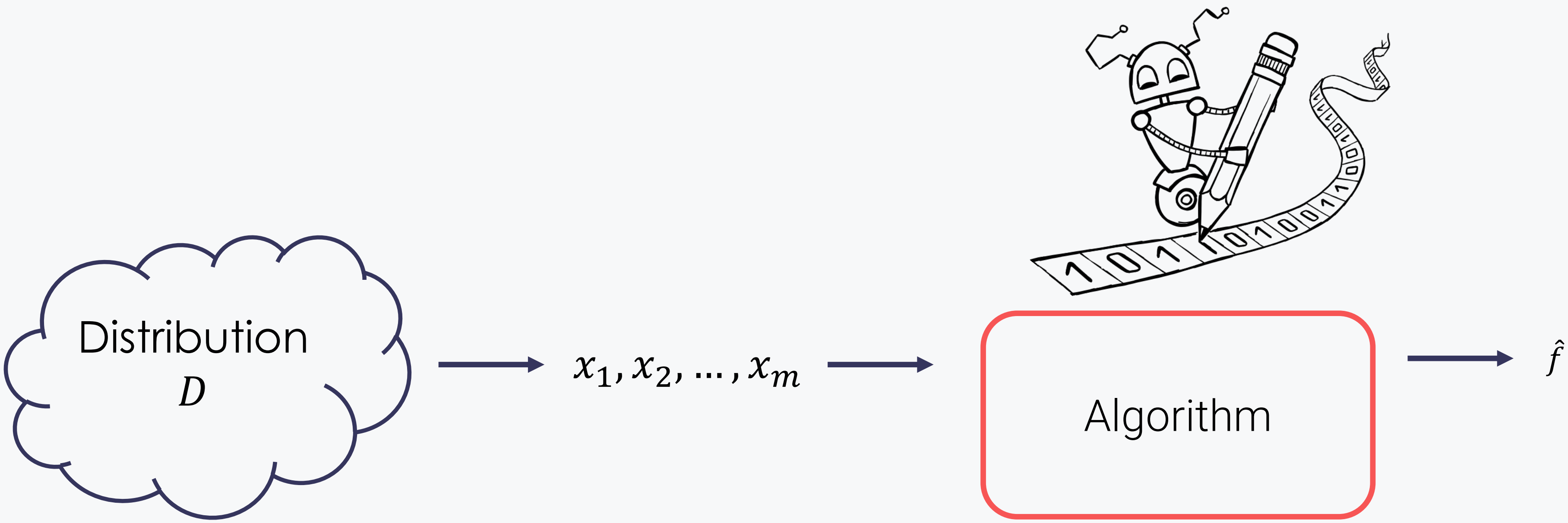


Image from: https://tilics.dmi.unibas.ch/the-turing-machine

Estimation with memory constraints

Unknown distribution D

Goal: Estimate $f(D)$ with error ϵ 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?

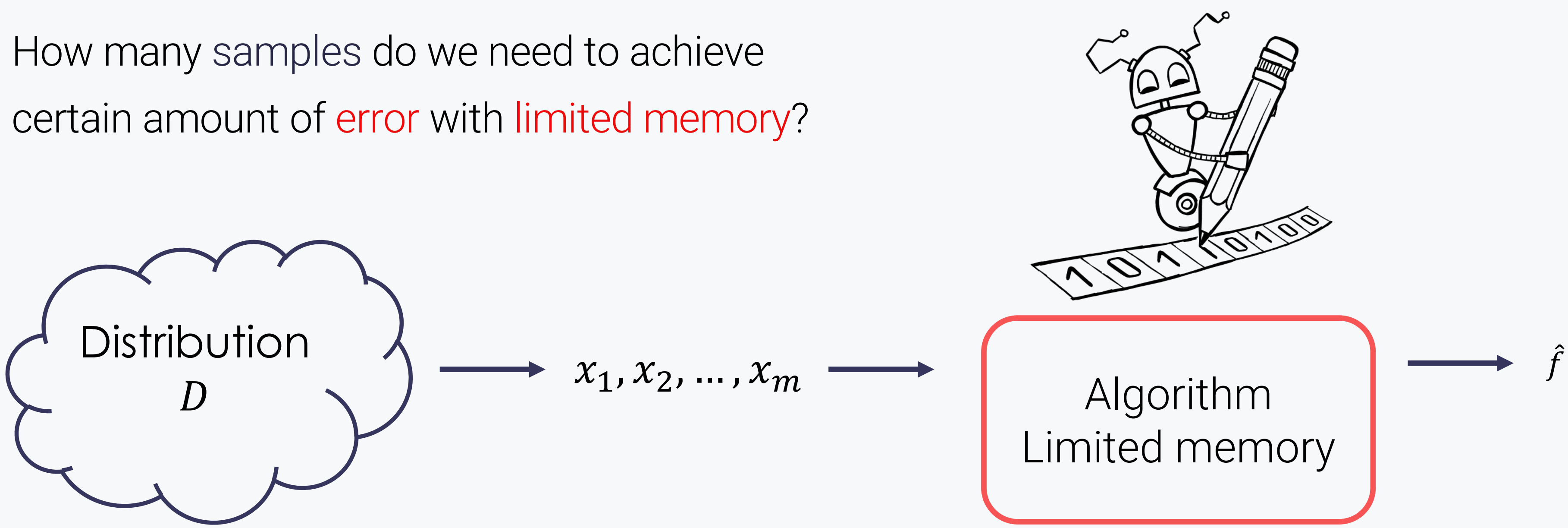
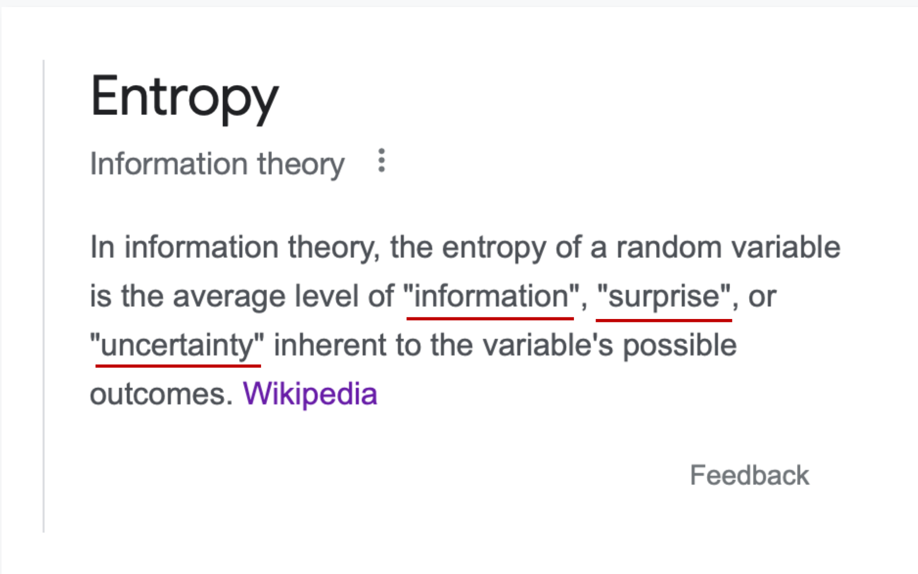


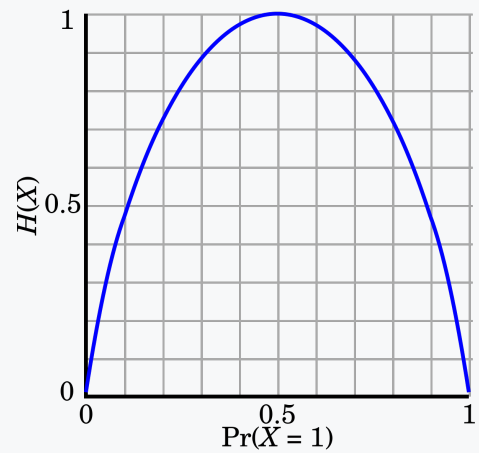
Image from: https://tilics.dmi.unibas.ch/the-turing-machine

This work: estimating entropy

Shannon's entropy of $D = (p_1, p_2, \dots, p_n)$:

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





Entropy of a binary random variable

This work: estimating entropy

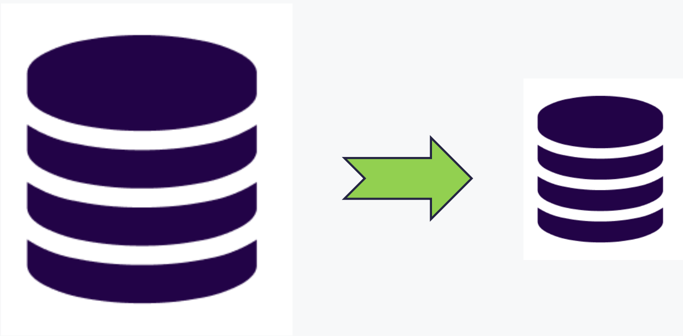
Shannon's entropy of $D = (p_1, p_2, \ldots, p_n)$:

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

Used in practice to measure randomness

Applications:

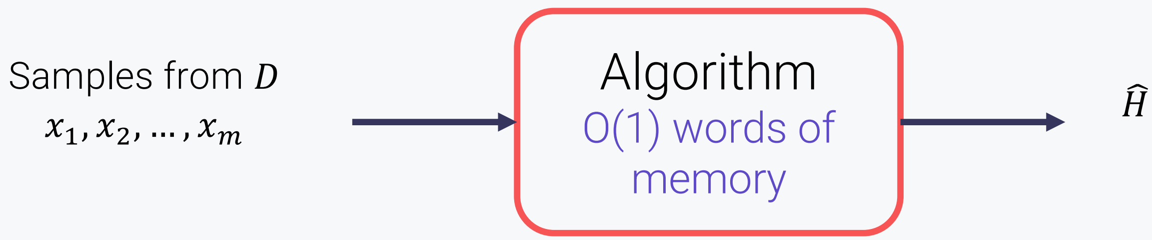
- Dataset summarization
- Data compression
- Evaluating language models
- Clustering and classification



Problem definition

Shannon's entropy of $D = (p_1, p_2, \dots, p_n)$:

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



Goal:

$$\Pr\left[\left|\hat{H} - H(D)\right| \le \epsilon\right] \ge 0.9$$

Memory constraint: $O(1)$ words of memory ($Polylog(n, 1/\epsilon)$ bits)

n = domain size
ϵ = error

Our results

Theorem [A, McGregor, Nelson, Waingarten'22]

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

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

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

[Batu, Dasgupta, Kumar, Rubinfeld. STOC 2002] [Paninski 2003] [Valiant 2008] [Valiant, Valiant. FOCS 2011] [Valiant, Valiant. JACM 2017] [Wu, Yang. IEEE Trans. IT 2016] [Jiao et al. IEEE Trans. IT 2015] .... (and many more)

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

[Acharya, Bhadane, Indyk, Sun, NeurIPS 2019]

A closely related model: streaming algorithms

Distribution D over $[n]$ → $x_1, x_2, \ldots, x_m$ → Algorithm with limited memory → $\hat{f}$

This talk: Properties of the distribution

$x_1, x_2, \ldots, x_m$ → Algorithm with limited memory → $\hat{f}$

Properties of the data stream itself

n = domain size
ϵ = error

Our results

Theorem

[A, McGregor, Nelson, Waingarten'22]

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

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

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

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

[Chakrabarti, Cormode, McGregor'10]
[Jayaram Woodruff'19]

Techniques

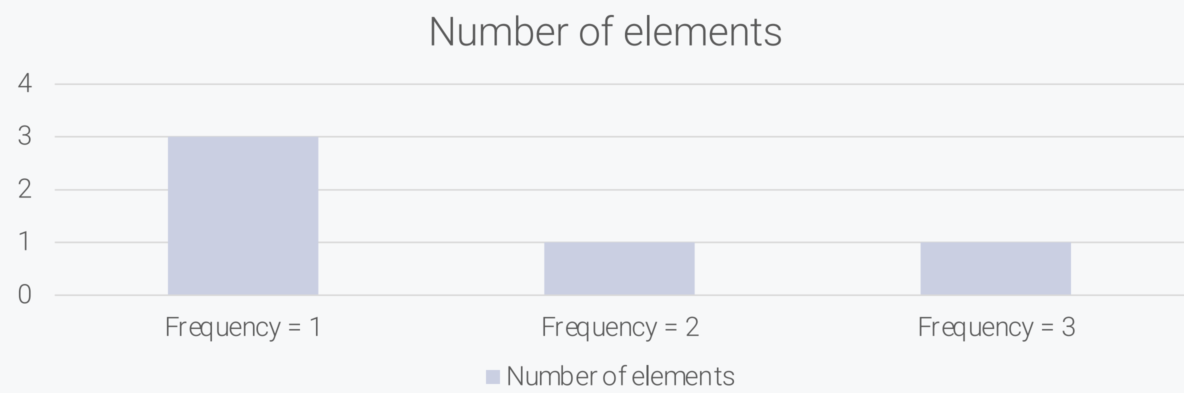
No memory constraint

Algorithm [Valiant, Valiant'11]:

1. Compute the fingerprint of the samples

List

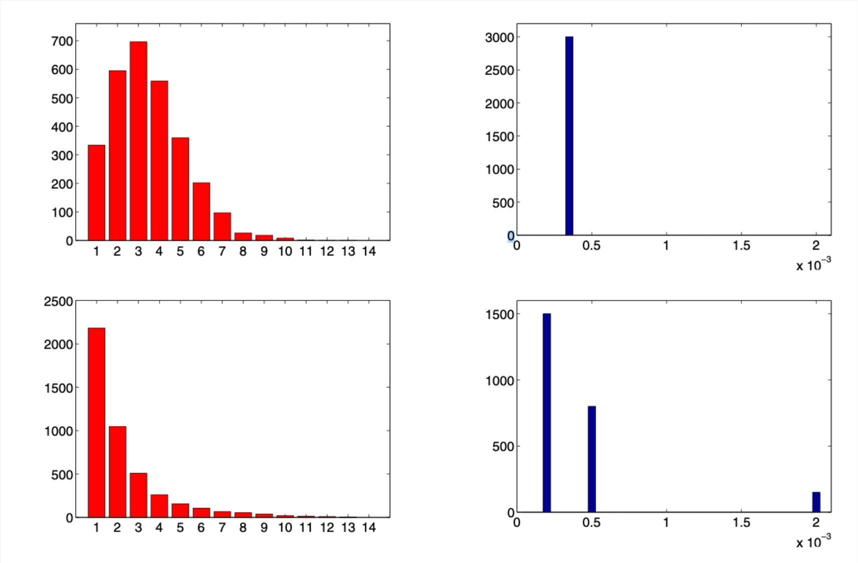
3 1 3 8 7 3 1 5



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

n = domain size
ϵ = error

How? Take average

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

x_1

p_{x_i}'s are unknown 😔

x_r

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

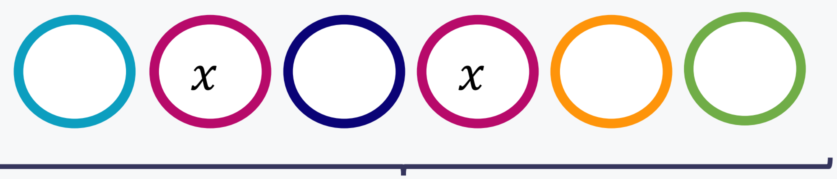
$$\frac{1}{r} \sum_{i=1}^{r} \log \frac{1}{p_{x_i}} \xrightarrow{\text{large } r} \mathrm{E}_{x \sim D}\left[\log \frac{1}{p_x}\right] = \mathrm{H(D)}$$

Estimate probabilities

Fix m. Count i's in next m samples.

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

#instances of $x \sim \text{Bin}\,(m, p_x)$



In the example: $\frac{2}{6}$

n = domain size
ϵ = error

How? Take average

$$H(D) \coloneqq \sum_{x=1}^{n} p_x \cdot \log \frac{1}{p_x} = \mathrm{E}_{x \sim D}\left[\log \frac{1}{p_x}\right]$$

x_1 x_2 ... x_r

$$\log \frac{1}{\hat{p}_{x_1}} \quad \log \frac{1}{\hat{p}_{x_2}} \quad \cdots \quad \log \frac{1}{\hat{p}_{x_r}}$$

$$\frac{1}{r} \sum_{i=1}^{r} \log \frac{1}{\hat{p}_{x_1}} \xrightarrow{\text{large r}} \mathrm{H(D)}$$

n = domain size
ϵ = error

How? Take average

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

Error of estimation (first two expressions)

Bias (second and third expressions)

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

❌

n = domain size
ϵ = error
m = number of samples to estimate p_i
r = number of rounds

Analysis of error

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

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

$$\leq \underbrace{\left|\mathrm{E}_{i\sim D}\left[\log\frac{1}{p_i}\right] - \mathrm{E}_{i\sim D}\left[\log\frac{1}{\hat{p}_i}\right]\right|}_{\text{Bias}} + \underbrace{|\mathrm{E}[\hat{H}_i] - \hat{H}|}_{\text{Error of estimation}}$$

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

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

n = domain size
ϵ = error

Simple algorithm [Plug-in estimator]

$$H(D) \coloneqq \sum_{i=1}^{n} p_i \cdot \log 1/p_i = \mathrm{E}_{i \sim D}[\log 1/p_i]$$

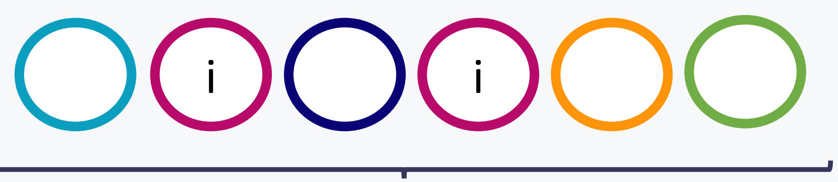
1. Repeat r times
   1. Draw $i \sim D$.
   2. $\hat{p}_i \leftarrow$ Estimate p_i →
   3. $\widehat{H}_i \leftarrow \log 1/\hat{p}_i$
2. Output: $\widehat{H} := \frac{1}{\mathrm{r}} \sum_{i=1}^{r} \widehat{H}_i$

Fix m. Count i's in next m samples.

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

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

In the example: $\frac{2}{6}$



Simple algorithm

n = domain size
ϵ = error
m = number of samples to estimate p_i
r = number of rounds

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

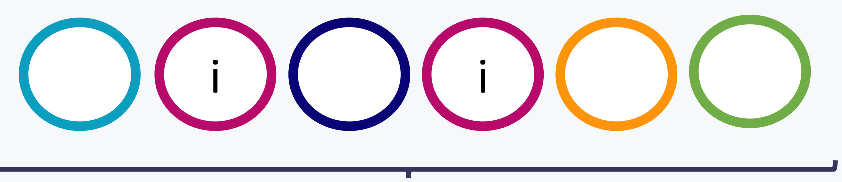
1. Repeat r times
   1. Draw $i \sim D$.
   2. $\hat{p}_i \leftarrow$ Estimate p_i
   3. $\widehat{H}_i \leftarrow \log 1/\hat{p}_i$
2. Output: $\widehat{H} := \frac{1}{\mathrm{r}} \sum_{i=1}^{r} \widehat{H}_i$

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

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

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

In the example: $\frac{2}{6}$



Idea I: Estimate via negative binomials

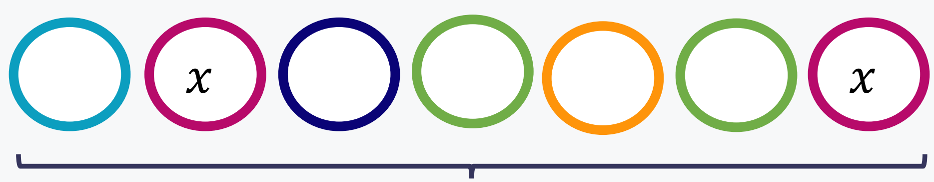
Count the number of samples until t instances of x are observed.

$$\#\text{samples} \sim \text{Negative Bin}\,(t, p_x)$$

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

$$\mathrm{E}[X_x] = 1/p_x$$

In the example for $t = 2$: $\mathrm{X}_x = \frac{7}{2}$



Analysis of error

n = domain size of the distribution
ϵ = error parameter
r = number of rounds
t = number of observed instance of i

Error: $\left|H(D) - \hat{H}\right| \overset{?}{\leq} \epsilon$

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

$$\leq \underbrace{\left|\mathrm{E}_{i\sim D}\left[\log\frac{1}{p_i}\right] - \mathrm{E}_{i\sim D}\left[\log\frac{1}{\hat{p}_i}\right]\right|}_{\text{Bias}} + \underbrace{\left|\mathrm{E}[\hat{H}_i] - \hat{H}\right|}_{\text{Error of estimation}}$$

$t = \Theta(1/\epsilon)$ implies bias $< \epsilon/2$ $\quad r = \Theta(\log^2 n/\epsilon^2)$ implies that error $< \epsilon/2$

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

Idea II: Remove bias

n = domain size of the distribution
ϵ = error parameter
r = number of rounds
t = number of observed instance of i
X_i = number of samples to see t instance of i
$\mathrm{E}[X_i] = 1/p_i$

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

Let $Y_i \leftarrow p_i X_i$

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

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

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

Idea II: Remove bias

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

Bias < $\mathrm{E}\left[\right.$ ... $\left. Y_i - 1)^{s+1}\right]$

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

concentrated

Polynomial of degree s of p_i

$\Pr[\text{k samples are equal}] = p_i^k$

Idea III: Remove $\log n$ factors

n = domain size of the distribution
ϵ = error parameter
r = number of rounds
t = number of observed instance of i
X_i = number of samples to see t instance of i
$\mathrm{E}[X_i] = 1/p_i$

Idea: Bucketing

Partition the range of X_i into L intervals

$$\mathrm{E}_{i\sim D}[\log X_i] = \sum_{\ell=1}^{L} \underbrace{\Pr[X_i \in I_\ell]}_{q_\ell} \cdot \underbrace{\mathrm{E}[\log X_i \mid X_i \in I_\ell]}_{H_\ell}$$

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

n = domain size of the distribution
ϵ = error parameter
r = number of rounds
t = number of observed instance of i
X_i = number of samples to see t instance of i
$\mathrm{E}[X_i] = 1/p_i$

Idea III: Remove $\log n$ factors

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

Bucke … curacy.

Removing $O(\log n)$.