Attenuation: Optimal Probability Estimation

Ben Kim, Ye Liu, Rong Wei

From the paper "Always Good Turing: Asymptotically Optimal Probability Estimation"

Alon Orlitsky, Narayana P. Santhanam, Junan Zhang

Topics

Introduction

- Say we are choosing pebbles from a bag, with only a couple of tries.
- We choose one yellow pebble, and one green pebble.
- How do we estimate the true probability distribution?

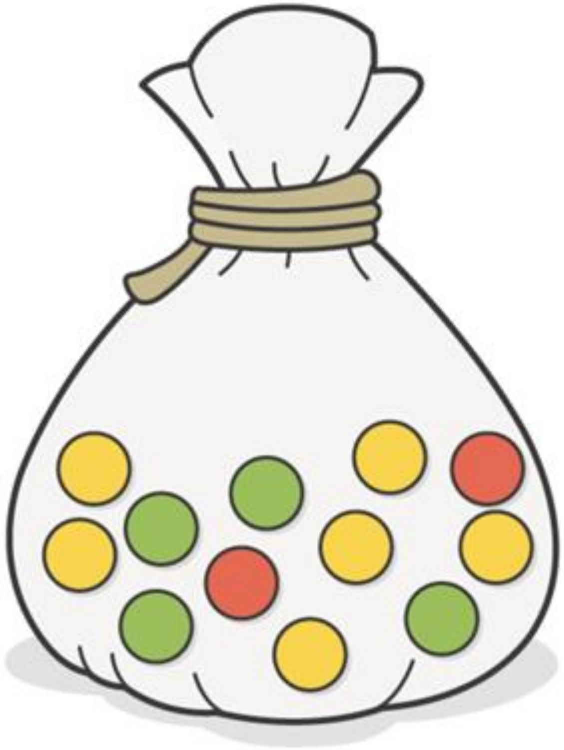
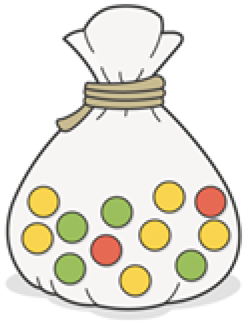


Image Credits: Bamboozle Math Games

How do we estimate the probability distribution?



- Naive Empirical: half green, half red

$$P(\text{Yellow}) = 0.5, \quad P(\text{Green}) = 0.5$$

- Laplace estimator: Addition of one to every possibility: 1 red, 2 green, 2 yellow

$$P(\text{Yellow}) = \frac{2}{5}, \quad P(\text{Green}) = \frac{2}{5}, \quad P(\text{Red}) = \frac{1}{5}$$

  - Other add constant estimators have taken a similar approach
  - This approach is weak when the number of possibilities is too large compared to sample size

- Alan Turing and I.J. Good had the same problem deciphering encrypted messages during WWII. (Good Turing Estimator!!)

Good Turing Estimator

- I.J. Good and Alan Turing had obtained the German Cipher Book wanted to apply the cipher book for a cryptanalysis to help decipher messages

- Derived the “Good Turing Estimator”
  - Conceptually: “smooths” probability distribution and reallocating probability to rare events
  - Useful for small sample size, or many events with small possibilities
  - Turing and Good had a small sample size of German intercepted ciphers

- Since publication, has had useful applications in information retrieval, spelling correction, speech recognition

Image Credits: Science Museum, Virginia Tech Science Magazine

Main Contributions

- This paper introduces a novel framework which can be used to evaluate probability estimators based on their attenuation

- The authors derive diminishing attenuation estimators, which approach optimal performance as there is an increase in samples

- They then evaluate the performance of all these estimators by bounding them as well as analyzing simple examples

Definitions

- Estimator - Assigns probability distribution to observed samples
- Patterns - abstract the sequence of observations, replaces each unique element with its order of first appearance.

  Valid for each new outcome i, $\quad i > 1 \quad$ occurs after the $(i-1)$-th index.

  - Ex: $(121), (132)$
  - Denoted by $\Psi(\overline{x}), \Psi(a, a, b, c) = 1123$

- Probability of Patterns - the probability that a sequence generates a pattern when sampled from a distribution

  $$p^{\Psi}(\bar{\psi}) \stackrel{\text{def}}{=} p\{\bar{x} \in A^n : \Psi(\bar{x}) = \bar{\psi}\}$$

  - Ex: $\Psi(11)$

Definitions

- Maximum pattern probability:
  - Highest probability assigned to the pattern by any distribution. $\hat{p}^{\Psi}\left(\overline{\psi}\right) = \max_{p} p^{\Psi}\left(\overline{\psi}\right)$
  - Ex: Constant Distribution $\hat{p}^{\Psi}(1 \ldots 1) = 1$ and Continuous Distribution $\hat{p}^{\Psi}(12 \ldots n) = 1$
- We denote a pattern $\psi_1^n = \psi_1 \psi_2 \ldots \psi_n$, and the number of distinct symbols appearing in the pattern $m = |\{\psi_1, \ldots, \psi_n\}|$.
- Sequential Estimators:
  - A mapping q that associates with every pattern ψ_1^n a probability distribution over $[m+1] = \{1, 2, \ldots, m+1\}$
  - Chain Rule:

$$q(\psi_1^n) = \prod_{i=0}^{n-1} q(\psi_{i+1} | \psi_1^i)$$

Definitions

- Ex: Add-one estimator producing '1213'

$$
\begin{aligned}
q_{+1}(1|1) &= \frac{1+1}{3} = \frac{2}{3}, \\
q_{+1}(2|1) &= \frac{0+1}{3} = \frac{1}{3}, \\
q_{+1}(1213) &= q_{+1}(1|\Lambda) \cdot q_{+1}(2|1) \cdot q_{+1}(1|12) \cdot q_{+1}(3|121) \\
&= \frac{1}{1} \cdot \frac{1}{3} \cdot \frac{2}{5} \cdot \frac{1}{6} \\
&= \frac{1}{45}.
\end{aligned}
$$

Definitions

- Sequence attenuation of an estimator q for a pattern ψ_1^n :

$$R(q, \psi_1^n) = \frac{\hat{p}^{\Psi}(\psi_1^n)}{q(\psi_1^n)} \qquad \frac{\text{highest probability assigned to } \psi_1^n \text{ by any distribution}}{\text{probability assigned to it by q}}$$

- Ex: Estimator q assigns a probability of 0.1 to pattern ψ_1^n
  - True probability p assigns it 0.3 to pattern ψ_1^n
  - $$R(q, \psi_1^n) = \frac{\hat{p}^{\Psi}(\psi_1^n)}{q(\psi_1^n)} = \frac{0.3}{0.1} = 3$$
  - q 's probability for this pattern is three times smaller than the best possible probability distribution

Definitions

- worst-case sequence attenuation of q (largest sequence attenuation of q for any length-n pattern):

  $$R^n(q) = max_{\psi_1^n \in \Psi^n} R(q, \psi_1^n)$$

- worst-case symbol attenuation of q for length-n patterns: $(R^n(q))^{\frac{1}{n}}$

- (asymptotic, worst-case, symbol) attenuation of q: $R^*(q) = \lim\ \sup_{n \to \infty} (R^n(q))^{\frac{1}{n}}$

- Diminishing attenuation estimator, as samples increase we approach optimal distribution estimation

A Preliminary Result

- multiplicity of ψ in ψ_1^n (number of times ψ appears in pattern):

$$\mu_\psi = \mu_\psi(\psi_1^n) = |1 \le i \le n : \psi_i = \psi|$$

- prevalence of the multiplicity μ (number of symbols appearing μ times in pattern):

$$\varphi_\mu = \overline{\varphi_\mu} = |\psi : \mu_\psi = \mu|$$

- Example

For pattern ψ_1^n = 1213,

$\mu_1 = 2, \mu_2 = \mu_3 = 1$: 1 appears twice, 2 and 3 each appear once

$\varphi_2 = 1, \varphi_1 = 2$: 2 symbols including 2 and 3 appear once, 1 symbol including 1 appear twice

A Preliminary Result

Number of distinct patterns with prevalences ψ_1, ψ_2, … , ψ_n:

$$\frac{n!}{\prod_{\mu=1}^{n} (\mu!)^{\varphi_\mu} \varphi_\mu!} \stackrel{\text{def}}{=} N(\varphi_1, \ldots, \varphi_n). \quad \text{where} \quad n = \sum_\mu \mu \varphi_\mu$$

Since maximum probability is achieved by having a distribution with the same probability,

$$\hat{p}^{\Psi}(\overline{\psi}) \leq \frac{1}{N(\varphi_1, \ldots, \varphi_n)}.$$

Unbounded- and Constant-Attenuation Estimators

Add-constant estimators have unbounded attenuation.

A modified version of the add-one estimator and the Good-Turing estimator have constant, albeit non-diminishing, attenuations.

Add-One Estimator

Add-constant estimators have unbounded attenuation.

Theorem 1 $R^*(q_{+1}) = \infty$

Ex: For pattern 123…n, $q_{+1}(123\cdots n) = \frac{1}{1}\frac{1}{3}\cdots\frac{1}{2n+1} = \frac{246\cdots 2n}{(2n+1)!} = \frac{2^n(12\cdots n)}{(2n+1)!} = \frac{2^n n!}{(2n+1)!}$ (14.1)

$p^{\psi}(12\cdots n) = 1$ since a string of positive integers is pattern iff the first appearance of any i ≥ 2 occurs after that of i − 1

$$R^*(q_{+1}) = \lim\ \sup_{n\to\infty}(R^n(q))^{\frac{1}{n}} = \lim\ \sup_{n\to\infty}\frac{p^{\psi}(12\cdots n)}{q_{+1}(123\cdots n)} = \lim\ \sup_{n\to\infty}\frac{(2n+1)!}{2^n n!} \quad (14.2)$$

$$\frac{(2n+1)!}{2^n n!} = 2n\frac{(2n+1)!}{2^{n+1}nn!} \geq 2n\frac{(2n+1)!}{2^{n+1}(n+1)!} \geq 2n\frac{(2n+1)!}{(n+1)!(n+1)!} = 2n\frac{(2n+1)(2n)\cdots(n+2)}{(n+1)n\cdots 21}$$

$\geq 2n \geq 2n\frac{1}{e} = \frac{2n}{e}$ by using the fact 2^n grows slower than (n+1)!

As n goes to infinity, $\frac{(2n+1)!}{2^n n!}$ goes to infinity.

Therefore, the attenuation of add one estimator is infinity so that unbounded.

Modified Add-one Estimator

The estimator uses the add-one rule to estimate the probability of the next symbol being new or repeated, and for repeated symbols it assigns a probability proportional to the number of occurrences of the symbol.

m: number of distinct symbols appearing in a pattern ψ_1^n

μ_ψ: multiplicity of ψ in ψ_1^n for 1 ≤ ψ ≤ m

Then estimator assigns probability as :

$$q_{+1'}(\psi_{n+1}|\psi_1^n) = \begin{cases} \frac{m+1}{n+2} & \text{if } \psi_{n+1} = m+1 \\ \frac{n-m+1}{n+2}\frac{\mu_{\psi_{n+1}}}{n} & \text{if } 1 \leq \psi_{n+1} \leq m \end{cases} \quad (15.1)$$

If the next symbol has never been seen, define the probability as the add-one rule; if the next symbol has been seen from 1 to m, define the probability multiply the proportion of number of times of the symbol to length n.

Modified Add-one Estimator

Theorem 2 $1.69 < R^*(q_{+1'}) \leq 2.85$

Ex: pattern $\overline{\psi} = 12\cdots\frac{n}{2}12\cdots\frac{n}{2}$ estimator assigns probability

$$q_{+1'}(\overline{\psi}) = \frac{((\frac{n}{2})!)^2(\frac{n}{2}-1)!}{(n+1)!(n-1)!} \quad (16.1) \qquad \approx 0.58^n n^{-n/2}$$

by using Stirling's approximation $n! \approx \sqrt{2\pi n}(\frac{n}{e})^n$ and approximate terms like n-1 to n for large n

uniform distribution over an alphabet of size 0.628n assigns to ψ the probability $0.98^n n^{-n/2}$

$$R^*(q_{+1'}) = \lim\ \sup_{n\to\infty}(R^n(q_{+1'}))^{\frac{1}{n}} \geq (\frac{\hat{p}^{\Psi}(\psi_1^n)}{q(\psi_1^n)})^{\frac{1}{n}} = (\frac{0.98^n n^{-n/2}}{0.58^n n^{-n/2}})^{\frac{1}{n}} > 1.69$$

Modified Add-one Estimator

sequence attenuation of any length-n pattern ψ with m distinct symbols is bounded by

$$R(q_{+1}, \overline{\psi}) = \frac{\hat{p}^{\Psi}(\overline{\psi})}{q_{+1}(\overline{\psi})} \leq \frac{\frac{1}{N(\varphi_1, \ldots, \varphi_n)}}{q_{+1}(\overline{\psi})} = 2^{nH(\frac{m}{n}) - mlog(\frac{m}{n})}$$

Then try to maximize by m to let $\frac{d}{dm} nH(\frac{m}{n}) - m\frac{m}{n} = 0$

$$\frac{d}{dm} nH(\frac{m}{n}) - mlog(\frac{m}{n}) = log(1 - \frac{m}{n}) - log(\frac{m}{n}) + log(\frac{n}{m}) - 1 = log((1 - \frac{m}{n})\frac{n^2}{m^2}) - 1 = 0 \quad (18.1)$$

By solving the equation above, we get n = 2m. Take this back to $R(q_{+1}, \overline{\psi})$, we have $2^{1.5n}$.

Then the attenuation of estimator is bounded by

$$(R^n(q_{+1}))^{\frac{1}{n}} = (2^{1.5n})^{\frac{1}{n}} \approx 2.85$$

Good-Turing Estimator

$r = \mu_{\varphi_{n+1}}(\varphi_1^n)$: number of φ_{n+1} appearing in ψ_1^n

$$q(\psi_{n+1}|\psi_1^n) = \begin{cases} \frac{\varphi_1'}{n}, & r = 0 \\ \frac{r+1}{n}\frac{\varphi_{r+1}'}{\varphi_r'}, & r \geq 1 \end{cases} \quad (19.1)$$

where φ_μ' is a smoothed value $\varphi_\mu' = \max(\varphi_\mu, 1)$: simplest smoothing technique

$$q_{\mathrm{GT1}}(\psi_{n+1}|\psi_1^n) \stackrel{\mathrm{def}}{=} \begin{cases} \frac{\max(\varphi_1,1)}{S_{\mathrm{GT1}}(\psi_1^n)}, & r = 0 \\ \frac{r+1}{S_{\mathrm{GT1}}(\psi_1^n)}\frac{\max(\varphi_{r+1},1)}{\varphi_r}, & r \geq 1, \end{cases} \quad (19.2)$$

where

$$S_{\mathrm{GT1}}(\psi_1^n) \stackrel{\mathrm{def}}{=} \max(\varphi_1, 1) + \sum_{\mu:\varphi_\mu>0} \varphi_\mu \cdot (\mu+1)\frac{\max(\varphi_{\mu+1}, 1)}{\varphi_\mu}$$ try to ensure probability sum to 1

is a normalization factor.

Good-Turing Estimator

Theorem 3 $1.39 < R^*(q_{GT1}) \le 2$

Ex: for the pattern $12(132)^{n/3} \stackrel{\text{def}}{=} 12132132\ldots132$.

Reason to choose this pattern: there are always some symbols appearing different times than others

$q_{\text{GT1}}(\overline{\psi}) = \Theta(72^{-n/3})$ by considering probability associate with pattern 132 with 3, 4, 6 possible values

$\hat{p}^{\Psi}(\overline{\psi}) = \Theta(3^{-n})$ by having uniform distribution assign to ψ

$$R^*(q_{GT1}) = \lim\ \sup_{n\to\infty}(R^n(q_{GT1}))^{\frac{1}{n}} = \left(\frac{\hat{p}^{\Psi}(\overline{\psi})}{q_{GT1}(\overline{\psi})}\right)^{1/n} \ge \left(\frac{3^{-n}}{72^{-n/3}}\right)^{1/n} = \frac{72^{1/3}}{3} > 1.39$$

Good-Turing Estimator

Upper bound

$$r(i) \stackrel{\text{def}}{=} \mu_{\psi_{i+1}}(\psi_1^i) \qquad \varphi_\mu^i \stackrel{\text{def}}{=} \varphi_\mu(\psi_1^i)$$

$$q_{\text{GT1}}(\psi_1^n) = \frac{\prod_{\mu=1}^n (\mu!)^{\varphi_\mu}}{\prod_{i=1}^{n-1} S_{\text{GT1}}(\psi_1^i)} \cdot \prod_{i=1}^{n-1} \frac{\max(\varphi_{r(i)+1}^i, 1)}{\varphi_{r(i)}^i} \quad (21.1)$$

$$\hat{p}^\Psi(\overline{\psi}) \leq \frac{1}{N(\varphi_1, \ldots, \varphi_n)} = \frac{\prod_{\mu=1}^n (\mu!)^{\psi}\varphi_\mu^n!}{n!} \quad (21.2)$$

$$R^n(q) = \max_{\psi_1^n \in \Psi^n} \frac{\hat{p}^\Psi(\psi_1^n)}{q(\psi_1^n)} <= \frac{\prod_{\mu=1}^n (\mu!)^{\psi_\mu}\varphi_\mu^n!}{n!} \Big/ \left(\frac{\prod_{\mu=1}^n (\mu!)^{\varphi_\mu}}{\prod_{i=1}^{n-1} S_{\text{GT1}}(\psi_1^i)} \cdot \prod_{i=1}^{n-1} \frac{\max(\varphi_{r(i)+1}^i, 1)}{\varphi_{r(i)}^i} \right)$$

$$= \left(\max_{\psi_1^n} \frac{\prod_{\mu=1}^n \varphi_\mu^n!}{\prod_{i=1}^{n-1} \max(\varphi_{r(i)+1}^i, 1)/\varphi_{r(i)}^i} \right) \cdot \left(\max_{\psi_1^n} \frac{\prod_{i=1}^{n-1} S_{\text{GT}'}(\psi_1^i)}{n!} \right) \stackrel{\text{def}}{=} R_G^n \cdot R_S^n$$

Reason to separate to Rg and Rs: make calculation of upper bound much easier

Good-Turing Estimator

$$\left(\max_{\psi_1^n} \frac{\prod_{\mu=1}^{n} \varphi_\mu^n!}{\prod_{i=1}^{n-1} \max(\varphi_{r(i)+1}^i, 1)/\varphi_{r(i)}^i}\right) \cdot \left(\max_{\psi_1^n} \frac{\prod_{i=1}^{n-1} S_{\mathrm{GT}'}(\psi_1^i)}{n!}\right) \stackrel{\text{def}}{=} R_G^n \cdot R_S^n$$

According to the definition, we observe that $\prod_{i=1}^{n-1} \frac{\varphi_{r(i)+1}^i + 1}{\varphi_{r(i)}^i} = \prod_{\mu=1}^{n} \varphi_\mu^n!$ (22.1)

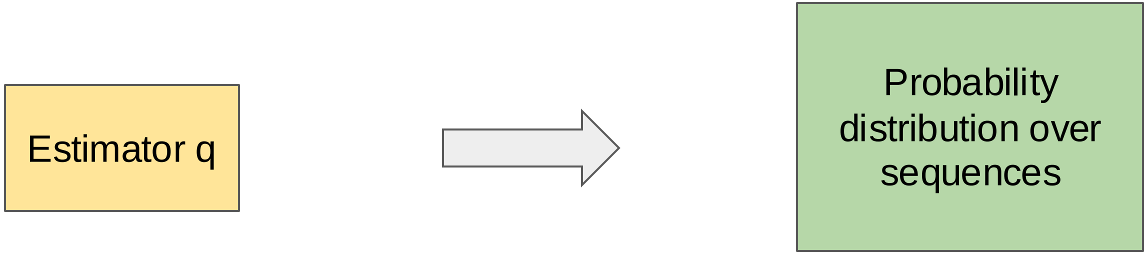
Place it into Rg, we could get $R_G^n \leq 2^{n-1}$.

Also, because $S_{\mathrm{GT1}}(\psi_1^n) \leq n + \sqrt{8n}$, $R_S^n \leq \left(1 + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right)\right)^{n-1} \cdot \frac{1}{n}$

Multiplying Rs and Rg together we could get upper bound 2.

Diminishing-attenuation Estimator

Diminishing-attenuation Estimator



To evaluate the performance of an estimator, We compare their sequence (symbol) attenuation.

$$R^*(q) = \lim \sup_{n\to\infty} (R^n(q))^{\frac{1}{n}}$$

Diminish-attenuation estimator: $R^*(q) \to 1$

Per-symbol probability assigned by the estimator is asymptotically the best possible.

Diminishing-attenuation Estimator

$q_{\frac{2}{3}}$ Computationally more efficient
(requires only a constant number of operations per symbol)

$q_{\frac{1}{2}}$ Attenuation approaches 1 more quickly

A Low Complexity Estimator

A Low Complexity Estimator

Definition

We define the estimator as the following:

$$q_{\frac{2}{3}}(\psi_{n+1}|\psi_1^n) = \frac{1}{S_{c_{n+1}}(\psi_1^n)} \times \begin{cases} f_{c_{n+1}}(\varphi_1 + 1), & r = 0 \\ (r+1)\frac{f_{c_{n+1}}(\varphi_{r+1}+1)}{f_{c_{n+1}}(\varphi_r)}, & r > 0 \end{cases}$$

Here μ_ψ is multiplicity of ψ and φ_μ is the prevalence of μ. Recall the symbols:

$$f_c(\varphi) = \max(\varphi, c), c_n = \lceil n^{\frac{1}{3}} \rceil$$
$$r = \mu_{\psi_{n+1}}(\psi_1^n)$$

Additionally, $S_{c_{n+1}}(\psi_1^n)$ is a normalization factor.

$$S_{c_{n+1}}(\psi_1^n) = f_{c_{n+1}}(\varphi_1 + 1) + \sum_{\mu=1}^{n} \varphi_\mu(\mu + 1)\frac{f_{c_{n+1}}(\varphi_{\mu+1} + 1)}{f_{c_{n+1}}(\varphi_\mu)}$$

A Low Complexity Estimator

Theorem

The upper bound on the $q_{\frac{2}{3}}$ estimator's attenuation is at most $2^{O(n^{\frac{2}{3}})}$.

$$R^n(q_{\frac{2}{3}}) = 2^{O(n^{\frac{2}{3}})}$$

Where the implied constant is at most 10.

Remark

The symbol attenuation diminishes to 1 at a rate of at least $2^{O(n^{-\frac{1}{3}})}$.

A Low Complexity Estimator

Proof.

We denote $g_c(\varphi) = \prod_{k=1}^{\varphi} f_c(k) = \begin{cases} c^{\varphi}, & 0 \leq \varphi \leq c \\ \frac{c^c}{c!}\varphi!, & \varphi \geq c \end{cases}$. Then the sequence estimator can be calculated through induction on n:

$$q_{\frac{2}{3}}(\psi_1^n) = \frac{\prod_{\mu=1}^{n}((\mu!)^{\varphi_\mu^n} g_{c_n}(\varphi_\mu^n))}{\prod_{i=2}^{n} S_{c_i}(\psi_1^{i-1})} \prod_{i=1}^{n-1}(\prod_{\mu=1}^{i} \frac{g_{c_i}(\varphi_\mu^i)}{g_{c_{i+1}}(\varphi_\mu^i)})$$

Recall the maximum probability of a pattern is:

$$\hat{p}^{\Psi}(\overline{\psi}) \leq \frac{\prod_{\mu=1}^{n}(\mu!)^{\varphi_\mu}\varphi_\mu!}{n!}$$

We can get the upperbound for the sequence attenuation:

$$R^n(q_{\frac{2}{3}}) \leq \max_{\psi_1^n} \prod_{\mu=1}^{n} \frac{\varphi_\mu^n!}{g_{c_n}(\varphi_\mu^n)} \cdot \max_{\psi_1^n} \frac{\prod_{i=1}^{n-1} S_{c_i}(\psi_1^i)}{n!} \cdot \max_{\psi_1^n} \prod_{i=1}^{n-1}(\prod_{\mu=1}^{i} \frac{g_{c_{i+1}}(\varphi_\mu^i)}{g_{c_i}(\varphi_\mu^i)})$$

A Low Complexity Estimator

- Since $g_c(\varphi) \geq \varphi!$, then $R_G^n = \max_{\psi_1^n} \prod_{\mu=1}^{n} \frac{\varphi_\mu^n!}{g_{c_n}(\varphi_\mu^n)} \leq 1$.
- Lemma 21[2] indicates that:

$$S_\gamma(\psi_1^n) \leq (1+\frac{1}{\gamma})n + \sqrt{\frac{2n(2\gamma+1)^2}{\gamma}}$$

Through Arithmetic Mean-Geometric Mean Inequality, we get:

$$R_S^n = \max_{\psi_1^n} \frac{\prod_{i=1}^{n-1} S_{c_i}(\psi_1^i)}{n!}$$

$$\leq (\frac{1}{n-1}\sum_{i=1}^{n-1}((1+\frac{1}{c_{i+1}}) + \sqrt{\frac{2(2c_{i+1}+1)^2}{ic_{i+1}}}))^{n-1} \cdot \frac{1}{n}$$

- Lemma 20[2] includes the inequality:

$$R_L^n = \max_{\psi_1^n}(\prod_{\mu=1}^{i} \frac{g_{c_{i+1}}(\varphi_\mu^i)}{g_{c_i}(\varphi_\mu^i)}) \leq \prod_{i=1}^{n-1}(\frac{c_{i+1}}{c_i})^{\sqrt{2ic_{i+1}}}$$

[2]Orlitsky, A., Santhanam, N.P. and Zhang, J., 2004. Universal compression of memoryless sources over unknown alphabets. *IEEE Transactions on Information Theory*, *50*(7), pp.1469-1481.

A Low Complexity Estimator

Finally, we incorporate inequalities above and get:

$$R^n(q_{\frac{2}{3}}) \leq \prod_{i=1}^{n-1} (\frac{c_{i+1}}{c_i})^{\sqrt{2ic_{i+1}}}$$

$$\cdot (\frac{1}{n-1} \sum_{i=1}^{n-1} ((1 + \frac{1}{c_{i+1}}) + \sqrt{\frac{2(2c_{i+1}+1)^2}{ic_{i+1}}}))^{n-1} \cdot \frac{1}{n}$$

Plug $c_n = \lceil n^{\frac{1}{3}} \rceil$ into above inequality, we get the upperbound of $R^n(q_{\frac{2}{3}})$ is $2^{O(n^{\frac{2}{3}})}$

■

A Low Complexity Estimator

Remark

The number of operations required to compute all of $q_{\frac{2}{3}}(\psi_1), q_{\frac{2}{3}}(\psi_2|\psi_1), \ldots, q_{\frac{2}{3}}(\psi_n|\psi_1^{n-1})$ *grows linearly with n. It means that it requires only a constant number of operations per symbol. Recall the construction of estimator:*

$$q_{\frac{2}{3}}(\psi_{n+1}|\psi_1^n) = \frac{1}{S_{c_{n+1}}(\psi_1^n)} \times \begin{cases} f_{c_{n+1}}(\varphi_1 + 1), & r = 0 \\ (r+1)\frac{f_{c_{n+1}}(\varphi_{r+1}+1)}{f_{c_{n+1}}(\varphi_r)}, & r > 0 \end{cases}$$

Since $c_n = \lceil n^{\frac{1}{3}} \rceil$*, then compute* $c_1, \ldots, c_n$ *requires only* $O(n^{\frac{1}{3}})$ *multiplications and* $O(n)$ *comparisons. It suffices to evaluate the complexity to calculate* $S_{c_i}(\psi_1^{i-1})$*. The proof is done by separating* $i \in$ *perfect cubes* $Z^3 = \{1^3, 2^3, \ldots\}$ *and* $i \notin Z^3$*, and then discuss the computation complexity of* $S_{c_i}(\psi_1^{i-1})$ *under two sets.*

[2]Orlitsky, A., Santhanam, N.P. and Zhang, J., 2004. Universal compression of memoryless sources over unknown alphabets. *IEEE Transactions on Information Theory*, *50*(7), pp.1469-1481.

A Low Attenuation Estimator

A Low Attenuation Estimator

Definition

We define the estimator as the following:

$$q_{\frac{1}{2}}(\psi_{n+1}|\psi_1^n) = \frac{\sum_{\overline{y}\in\Psi^{t_n}(\varphi_1^n\cdot\psi_{n+1})}\tilde{p}(\overline{y})}{\sum_{\overline{y}\in\Psi^{t_n}(\varphi_1^n)}\tilde{p}(\overline{y})}$$

Here

$$z(\psi_1^n) = \frac{1}{N(\psi_1,\ldots,\psi_n)} = \frac{\prod_{\mu=1}^n(\mu!)^{\varphi_\mu}(\varphi_\mu!)}{n!}$$

$$\tilde{p}(\psi_1^n) = \frac{z(\psi_1^n)}{\sum_{\overline{y}\in\Psi^n} z(\overline{y})}$$ is the distribution over Ψ^n

$$t_n = 2^{\lfloor\log n\rfloor+1}$$ is the smallest power of 2 that is larger than n

$$\Psi^{t_n}(\psi_1^n) = \{y_1^{t_n}\in\Psi^{t^n} : y_1^n = \psi_1^n\}$$

denotes the set of patterns of length t_n with prefix ψ_1^n

A Low Attenuation Estimator

We observe that the construction of the $q_{\frac{1}{2}}$ estimator is closely related to the partition of an integer.

$$\sum_{\overline{y}\in\Psi^n} z(\overline{y}) = \sum_{\overline{y}\in\Psi^n} \frac{1}{N(\overline{y})} = |\Phi^n|$$

Here $|\Phi^n|$ is the total partitions of the integer n.
For example, $n = 4$, the number of length-n patterns is $|\Psi^4| = 15$, the $4th$ Bell number, and $|\Phi^4| = 5$

- $4 = 4 + 0 \Rightarrow \{1111\}$
- $4 = 3 + 1 \Rightarrow \{1112, 1121, 1211, 1222\}$
- $4 = 2 + 2 \Rightarrow \{1122, 1212, 1221\}$
- $4 = 1 + 1 + 2 \Rightarrow \{1123, 1213, 1231, 1223, 1232, 1233\}$
- $4 = 1 + 1 + 1 + 1 \Rightarrow \{1234\}$

A Low Attenuation Estimator

Recall for any pattern $\psi_1^n \in \Psi^n$ of profile $\overline{\varphi} \in \Psi^n$, since every i.i.d. distribution assigns the same probability to all patterns of the same profile, the maximum probability of a pattern is upperbounded by:

$$\hat{p}(\psi_1^n) \leq \frac{1}{N(\psi_1, \ldots, \psi_n)}$$

Inspired by this upperbound, we construct following distributions:

$$\tilde{p}(\psi_1^n) = \frac{z(\psi_1^n)}{\sum_{\overline{y} \in \Psi^n} z(\overline{y})} = \frac{\frac{1}{N(\psi_1, \ldots, \psi_n)}}{\sum_{\overline{\psi} \in \Psi^n} \frac{1}{N(\psi_1, \ldots, \psi_n)}} = \frac{1}{N(\psi_1, \ldots, \psi_n)|\Phi^n|}$$

We denote $\tilde{p}^k(\psi_1^n) = \tilde{p}(\Psi^k(\psi_1^n)) = \sum_{\overline{y} \in \Psi^k(\psi_1^n)} \tilde{p}(\overline{y})$ Therefore,

$$q_{\frac{1}{2}}(\psi_{n+1}|\psi_1^n) = \frac{\sum_{\overline{y} \in \Psi^{t_n}(\varphi_1^n \cdot \psi_{n+1})} \tilde{p}(\overline{y})}{\sum_{\overline{y} \in \Psi^{t_n}(\varphi_1^n)} \tilde{p}(\overline{y})} = \frac{\tilde{p}^{t_n}(\psi_1^{n+1})}{\tilde{p}^{t_n}(\psi_1^n)}$$

A Low Attenuation Estimator

Theorem

The upper bound on the $q_{\frac{1}{2}}$ estimator's attenuation is bounded by:

$$R^n(q_{\frac{1}{2}}) \leq exp(\frac{4\pi}{\sqrt{3}(2-\sqrt{2})}\sqrt{n})$$

Remark

$q_{\frac{1}{2}}$ achieves a sequence attenuation of $2^{O(n^{\frac{1}{2}})}$, hence a symbol attenuation diminishes to 1 at a rate of at least $2^{O(n^{-\frac{1}{2}})}$.

A Low Attenuation Estimator

Proof.

For $n = 1$, the theorem holds trivially.

We rewrite the attenuation:

$$R^n(q_{\frac{1}{2}}) = \frac{\hat{p}^{\Psi}(\psi_1^n)}{q_{\frac{1}{2}}(\psi_1^n)} = \frac{\hat{p}^{\Psi}(\psi_1^n)}{\tilde{p}^{t_n}(\psi_1^n)} \cdot \frac{\tilde{p}^{t_n}(\psi_1^n)}{q_{\frac{1}{2}}(\psi_1^n)}$$

In next several slides, we will show bounds for each ratio. Then combine two parts together, we get the upperbound for sequence attenuation of $q_{\frac{1}{2}}$.

[2] Orlitsky, A., Santhanam, N. P., & Zhang, J. (2004). Universal compression of memoryless sources over unknown alphabets. *IEEE Transactions on Information Theory*, *50*(7), 1469-1481.

A Low Attenuation Estimator

For the first part, we first observe that:

$$\hat{p}^{\Psi}(\psi_1^n) = \sup_{\psi_1^n} \sum_{\overline{y} \in \Psi^{t_n}(\psi_1^n)} p(\overline{y}) \leq \sum_{\overline{y} \in \Psi^{t_n}(\psi_1^n)} \hat{p}_{\overline{y}}(\overline{y}) \leq \sum_{\overline{y} \in \Psi^{t_n}(\psi_1^n)} \frac{1}{N(y)}$$

$$= \sum_{\overline{y} \in \Psi^{t_n}(\psi_1^n)} |\Phi^n| \cdot \tilde{p}(y) \leq exp(\pi\sqrt{\frac{2}{3}}\sqrt{t_n})(\sum_{\overline{y} \in \Psi^{t_n}(\psi_1^n)} \tilde{p}(y))$$

$$= \tilde{p}^{t_n}(\psi_1^n) \cdot exp(\pi\sqrt{\frac{2}{3}}\sqrt{t_n})$$

The last inequality comes from Hardy and Ramanujan[3], which shows that the number of unordered partitions of n is:

$$exp(\pi\sqrt{\frac{2}{3}}\sqrt{n}(1 - o(1))) \leq |\Phi^n| \leq exp(\pi\sqrt{\frac{2}{3}}\sqrt{n})$$

Then we claim that:

$$\frac{\hat{p}^{\Psi}(\psi_1^n)}{\tilde{p}^{t_n}(\psi_1^n)} \leq exp(\pi\sqrt{\frac{2}{3}}\sqrt{t_n})$$

[3] Hardy, G. H., & Ramanujan, S. (1918). Asymptotic formulaæ in combinatory analysis. Proceedings of the London Mathematical Society, 2(1), 75-115.

A Low Attenuation Estimator

For the second part, we prove by induction on $i \geq 0$. We claim that for all $2^i < n < 2^{i+1}$,

$$\frac{\tilde{p}^{2^{i+1}}(\psi_1^n)}{q_{\frac{1}{2}}(\psi_1^n)} \leq exp(\pi\sqrt{\frac{2}{3}}\frac{\sqrt{2^{i+1}}}{\sqrt{2}-1})$$

First,

$$q_{\frac{1}{2}}(\psi_1^2) = \tilde{p}(\psi_1^2) = \frac{1}{2}$$

Then for i which satisfies $2^i \leq n \leq 2^{i+1}$, we have

$$q_{\frac{1}{2}}(\psi_1^n) = q_{\frac{1}{2}}(\psi_1^{2^i})q_{\frac{1}{2}}(\psi_1^n|\psi_1^{2^i}) = q_{\frac{1}{2}}(\psi_1^{2^i})\frac{\tilde{p}^{2^{i+1}(\psi_1^n)}}{\tilde{p}^{2^{i+1}(\psi_1^{2^i})}}$$

Hence,

$$\frac{\tilde{p}^{2^{i+1}}(\psi_1^n)}{q_{\frac{1}{2}}(\psi_1^n)} = \frac{\tilde{p}^{2^{i+1}}(\psi_1^{2^i})}{q_{\frac{1}{2}}(\psi_1^{2^i})} = \frac{\tilde{p}^{2^{i+1}}(\psi_1^{2^i})}{\tilde{p}(\psi_1^{2^i})} \cdot \frac{\tilde{p}(\psi_1^{2^i})}{q_{\frac{1}{2}}(\psi_1^{2^i})}$$

A Low Attenuation Estimator

By the induction hypothesis,

$$\frac{\tilde{p}(\psi_1^{2^i})}{q_{\frac{1}{2}}(\psi_1^{2^i})} \leq exp(\pi\sqrt{\frac{2}{3}}\frac{\sqrt{2^i}}{\sqrt{2-1}})$$

By definition of $\tilde{p}$, we get:

$$\begin{aligned}
\frac{\tilde{p}^{2^{i+1}}(\psi_1^{2^i})}{\tilde{p}(\psi_1^{2^i})} &= \frac{\tilde{p}^{2^{i+1}}(\psi_1^{2^i})}{\frac{1}{N(\psi_1^{2^i})|\Phi^{2^i}|}} \\
&\leq (N(\psi_1^{2^i}) \cdot \tilde{p}^{2^{i+1}}(\psi_1^{2^i})) \cdot exp(\pi\sqrt{\frac{2}{3}}\sqrt{2^i}) \\
&\leq (\sum_{\overline{y}\in\Psi^{2^{i+1}}} \tilde{p}(\overline{y})) \cdot exp(\pi\sqrt{\frac{2}{3}}\sqrt{2^i}) \\
&= exp(\pi\sqrt{\frac{2}{3}}\sqrt{2^i})
\end{aligned}$$

∎

Lower Bound on Attenuation

Lower bound on attenuation

Can we make the sequence attenuation arbitrarily small?

Theorem

For every estimator q, the sequence attenuation of any estimator grows at least exponentially in the cube root of the sequence length.

$$R^n(q) \geq exp\{\frac{3}{2}n^{\frac{1}{3}}[1 - o(1)]\}$$

[4] Orlitsky, A., & Santhanam, N. P. (2003, March). Performance of universal codes over infinite alphabets. In Data Compression Conference, 2003. Proceedings. DCC 2003 (pp. 402-410). IEEE.
[5] Jevtić, N., Orlitsky, A., & Santhanam, N. P. (2005). A lower bound on compression of unknown alphabets. Theoretical computer science, 332(1-3), 293-311.

Performance Examples

- Consider the low complexity estimator $q_{1/3}$ utilized for simple sequences
- Repeating Sequence ‘aaaa’
  - Estimates $1-\Theta(1/n)$ that the next symbol is ‘a’, $\Theta(1/n)$ that it is new
- Alternating sequences ‘ababa…’
  - $\Theta(1/n)$ That it is new, splits remaining $1-\Theta(1/n)$ between ‘a’ and ‘b’
- Unique symbols ‘abcdef’
  - $1-\Theta(1/n^{2/3})$ That the next symbol is new
- Doubled symbols ‘aabbcc…’
  - $1/4$ that the next symbol is new, $3/2n$ that the symbol is a preceding one

- The estimator generally aligns with one’s intuition for simple patterns

Applications of Good Turing Estimation

- Distribution estimating in Machine Learning [6]
  - Good-Turing estimators is near optimal for discrete distributions

- Life sciences [7], [8]
  - Applied to estimate the unseen species in a habitat
  - Occurrence of genetic variants

- Language Processing [9]
  - Applied in speech recognition and computational linguistics

[6] Orlitsky, A., & Suresh, A. T. (2015). Competitive distribution estimation: Why is Good-Turing good. *Advances in Neural Information Processing Systems* (pp. 2143-2151).
[7] Chao, A., & Lee, S. M. (1992). Estimating the number of classes via sample coverage. *Journal of the American Statistical Association*, 87(417), 210-217.
[8] Ionita-Laza, I., Lange, C., & Laird, N. M. (2009). Estimating the number of unseen variants in the human genome. *Proceedings of the National Academy of Sciences*, 106(13), 5008-5013.
[9] Gale, W. A., & Sampson, G. (1995). Good-Turing frequency estimation without tears. *Journal of Quantitative Linguistics*, 2(3), 217-237.

Conclusion

- This paper introduces a novel framework which can be used to evaluate probability estimators based on their attenuation

- The authors derive diminishing attenuation estimators, which approach optimal performance as there is an increase

- They then evaluate the performance of these estimators by bounding them as well as analyzing simple examples

References

Main Reference: [1] Orlitsky, A., Santhanam, N. P., & Zhang, J. (2003). Always Good Turing: Asymptotically Optimal Probability Estimation. *Science*, 302(5644), 427–431. DOI: 10.1126/science.1086078.

[2] Orlitsky, A., Santhanam, N.P. and Zhang, J., 2004. Universal compression of memoryless sources over unknown alphabets. *IEEE Transactions on Information Theory*, *50*(7), pp.1469-1481.

[3] Hardy, G. H., & Ramanujan, S. (1918). Asymptotic formulaæ in combinatory analysis. *Proceedings of the London Mathematical Society*, 2(1), 75-115.

[4] Orlitsky, A., & Santhanam, N. P. (2003, March). Performance of universal codes over infinite alphabets. In *Data Compression Conference, 2003. Proceedings. DCC 2003 (pp. 402-410). IEEE.*

[5] Jevtić, N., Orlitsky, A., & Santhanam, N. P. (2005). A lower bound on compression of unknown alphabets. *Theoretical computer science, 332(1-3), 293-311.*

[6] Orlitsky, A., & Suresh, A. T. (2015). Competitive distribution estimation: Why is Good-Turing good. *Advances in Neural Information Processing Systems* (pp. 2143-2151).

[7] Chao, A., & Lee, S. M. (1992). Estimating the number of classes via sample coverage. *Journal of the American Statistical Association*, 87(417), 210-217.

[8] Ionita-Laza, I., Lange, C., & Laird, N. M. (2009). Estimating the number of unseen variants in the human genome. *Proceedings of the National Academy of Sciences*, 106(13), 5008-5013.

[9] Gale, W. A., & Sampson, G. (1995). Good-Turing frequency estimation without tears. *Journal of Quantitative Linguistics*, 2(3), 217-237.

Thanks for Listening !