Maximum Likelihood Approach for Symmetric Distribution Property Estimation

Jayadev Acharya, Hirakendu Das, Alon Orlitsky, Ananda Suresh
Cornell, Yahoo, UCSD, Google

Information Theory and Applications (ITA), 2017
Property estimation

- $p$: unknown discrete distribution over $k$ elements
- $f(p)$: a property of $p$
- $\varepsilon$: accuracy parameter, $\delta$: error probability
- Given: access to independent samples from $p$
- Goal: estimate $f(p)$ to $\pm \varepsilon$ with probability $> 1 - \delta$

Usually, $\delta$ constant, say 0.1

Focus of the talk: large $k$ (really large)
Sample complexity

\(X_1, X_2, \ldots, X_n: \text{independent samples from } p\)

\(\hat{f}(X^n_1): \text{estimate of } f(p)\)

Sample complexity of \(\hat{f}(X^n_1)\):

\[ S(\hat{f}, \varepsilon, k, \delta) = \min\{n: \forall p, \Pr(|\hat{f}(X^n_1) - f(p)| \geq \varepsilon) \leq \delta\} \]

Sample complexity of \(f\):

\[ S(f, \varepsilon, k, \delta) = \min_{\hat{f}} S(\hat{f}, \varepsilon, k, \delta) \]
Symmetric properties

Permuting labels does not change $f(p)$

Examples:

• $H(p) \triangleq -\sum_x p_x \cdot \log(p_x)$

• $S(p) \triangleq \sum_x 1_{p_x>0}$

Renyi entropy, distance to uniformity, unseen symbols, divergences, etc
Sequence maximum likelihood

\( N_x : \) \# times \( x \) appears in \( X_1^n \) (multiplicity)

\( p^e : \) empirical distribution

\[
p^e_x = \frac{N_x}{n}
\]

\[
f^e(X_1^n) = f(p^e)
\]

Empirical estimators are maximum likelihood estimators

\[
p^e = \arg \max_p p(X_1^n)
\]

Call this sequence maximum likelihood (SML)
Empirical entropy estimation

Empirical estimator:

\[ H^e (X_1^n) = H(p^e) = \sum_x \frac{N_x}{n} \log \frac{n}{N_x}. \]

\[ S(H^e, \varepsilon, k, 0.1) = \Theta \left( \frac{k}{\varepsilon} \right) \]

Various corrections proposed:
Miller-Maddow, Jackknifed estimator, Coverage adjusted, ...

Sample complexity: \( \Omega(k) \)

[Paninski’03]: \( S(H, \varepsilon, k, 0.1) = o(k) \) (existential)

**Note:** For \( \varepsilon \ll 1/k \), empirical estimators are the best
Entropy estimation

[ValiantValiant’11a]: Constructive proofs based on LP:

\[ S(H, \varepsilon, k, 0.1) = \Theta \varepsilon \left( \frac{k}{\log k} \right) \]

• [YuWang’14, HanJiaoVenkatWeissman’14, ValiantValiant11b]: Simplified algorithms, and exact rates:

\[ S(H, \varepsilon, k, 0.1) = \Theta \left( \frac{k}{\varepsilon \log k} \right) \]
Support coverage

Expected number of symbols when $p$ is sampled $m$ times

$$S_m(p) = \sum_x (1 - (1 - p_x)^m)$$

Goal: Estimate $S_m(p)$ to $\pm (\varepsilon \cdot m)$

[OrlitskySureshWu’16, ZouValiantetal’16]:

$$n = \Theta \left( \frac{m}{\log m} \log \left( \frac{1}{\varepsilon} \right) \right) \text{ samples for } \delta = 0.1$$
Known results summary

Many symmetric properties: entropy, support size, distance to uniform, support coverage

• Different estimator for each property
• Sophisticated results from approximation theory
Main result

Simple, ML based plug-in method that is sample-optimal for entropy, support-coverage, distance to uniform, support size.

<table>
<thead>
<tr>
<th>Property</th>
<th>Notation</th>
<th>SML</th>
<th>Optimal</th>
<th>References</th>
<th>PML</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entropy</td>
<td>$H(p)$</td>
<td>$\frac{k}{\varepsilon}$</td>
<td>$\frac{k}{\log k} \frac{1}{\varepsilon}$</td>
<td>[VV11a, WY16, JVHW15]</td>
<td>optimal$^1$</td>
</tr>
<tr>
<td>Support size</td>
<td>$\frac{S(p)}{k}$</td>
<td>$k \log \frac{1}{\varepsilon}$</td>
<td>$\frac{k}{\log k} \log^2 \frac{1}{\varepsilon}$</td>
<td>[WY15]</td>
<td>optimal</td>
</tr>
<tr>
<td>Support coverage</td>
<td>$\frac{S_m(p)}{m}$</td>
<td>$m$</td>
<td>$\frac{m}{\log m} \log \frac{1}{\varepsilon}$</td>
<td>[OSW16]</td>
<td>optimal</td>
</tr>
<tr>
<td>Distance to $u$</td>
<td>$|p - u|_1$</td>
<td>$\frac{k}{\varepsilon^2}$</td>
<td>$\frac{k}{\log k} \frac{1}{\varepsilon^2}$</td>
<td>[VV11b, JHW16]</td>
<td>optimal</td>
</tr>
</tbody>
</table>
Profiles

Profile of a sequence is the multiset of multiplicities:

\[ \Phi(X_1^n) = \{N_x\} \]

\[ X_1^n = (1H, 2T), \text{ or } X_1^n = (2H, 1T), \Phi(X_1^n) = \{1,2\} \]

Symmetric properties depend on multiset of probabilities

Coins w/ bias 0.4, and 0.6 have same symmetric property

Optimal estimators have same output for sequences with same profile.

Profiles are sufficient statistic
Profile maximum likelihood [OSVZ’04]

Probability of a profile:
\[ p(\Phi(X^n_1)) = \sum_{Y^n_1: \Phi(Y^n_1) = \Phi(X^n_1)} p(Y^n_1) \]

Maximize the profile probability:
\[ p^{PML}_\Phi = \arg \max_p p(\Phi(X^n_1)) \]

\( X^n_1 = (1H, 2T) \):
SML: (2/3, 1/3)
PML: (1/2, 1/2)
PML for symmetric properties

To estimate a symmetric $f(p)$:
• Find $p^{\text{PML}}(\Phi(X_1^n))$
• Output $f(p^{\text{PML}})$

Advantages:
• No tuning parameters
• Not function specific
Main result

PML is **sample-optimal** for entropy, support coverage, distance to uniformity, and support size.
Ingredients

Guarantee for PML.

If \( n = S(f, \varepsilon, k, \delta) \), then \( S(f(P^{PML}), 2\varepsilon, k, \delta \cdot \frac{e^{3\sqrt{n}}}{10}) \leq n \)

If \( n = S(f, \varepsilon, k, e^{-3\sqrt{n}}) \), then \( S(f(P^{PML}), 2\varepsilon, k, 0.1) \leq n \)

# profiles of length \( n < \frac{e^{3\sqrt{n}}}{10} \)
Ingredients

\( n = S(f, \varepsilon, \delta) \), achieved by an estimator \( \hat{f}(\Phi(X^n)) \):

\( p \): underlying distribution.

• Profiles \( \Phi(X^n) \) such that \( p(\Phi(X^n)) > \delta \),

\[ p^{PML}(\Phi) \geq p(\Phi) > \delta \]

\[ p^{PML}_\Phi(\Phi) \geq p(\Phi) > \delta \]

\[ |f(p^{PML}_\Phi) - f(p)| \leq |f(p^{PML}_\Phi) - \hat{f}(\Phi)| + |\hat{f}(\Phi) - f(p)| < 2\varepsilon \]

• Profiles with \( p(\Phi(X^n)) < \delta \),

\[ p(p(\Phi(X^n) < \delta) < \delta \cdot \#\text{profiles of length } n \]
Ingredients

Better error probability guarantees.

Recall:

\[ S(H, \epsilon, k, 0.1) = \Theta \left( \frac{k}{\epsilon \cdot \log k} \right) \]

Stronger error guarantees using McDiarmid's inequality:

\[ S(H, \epsilon, k, e^{-n^{0.9}}) = \Theta \left( \frac{k}{\epsilon \cdot \log k} \right) \]

With twice the samples error drops **exponentially**

Similar results for other properties
Main result

PML is **sample-optimal** for entropy, unseen, distance to uniformity, and support size.

Even **approximate PML** is optimal for these.
Algorithms

• EM algorithm [Orlitsky et al]
• Approximate PML via Bethe Permanents [Vontobel]
• Extensions of Markov Chains [VatedkaVontobel]

Polynomial time algorithms for approximating PML
Summary

• Symmetric property estimation

• PML plug-in approach
  • Universal, simple to state
  • Independent of particular properties

• Directions:
  • Efficient algorithms – for approximate PML
  • Relies heavily on existence of other estimators
In Fisher’s words ... 

Of course nobody has been able to prove that maximum likelihood estimates are best under all circumstances. Maximum likelihood estimates computed with all the information available may turn out to be inconsistent. Throwing away a substantial part of the information may render them consistent.

R. A. Fisher
References


