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

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

Usually, $\boldsymbol{\delta}$ constant, say **0.1**

Focus of the talk: **large** $\boldsymbol{k}$ (really large)

Sample complexity

$X_1, X_2, \ldots, X_n$: independent samples from p

$\hat{f}(X_1^n)$: estimate of $f(p)$

Sample complexity of $\hat{f}(X_1^n)$:

$$S(\hat{f}, \varepsilon, k, \delta) = \min\{n : \forall p, \Pr\left(|\hat{f}(X_1^n) - f(p)| \geq \varepsilon\right) \leq \delta\}$$

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

$$p_x^e = \frac{N_x}{n}$$

$$f^e(X_1^n) = f(p^e)$$

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]:

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

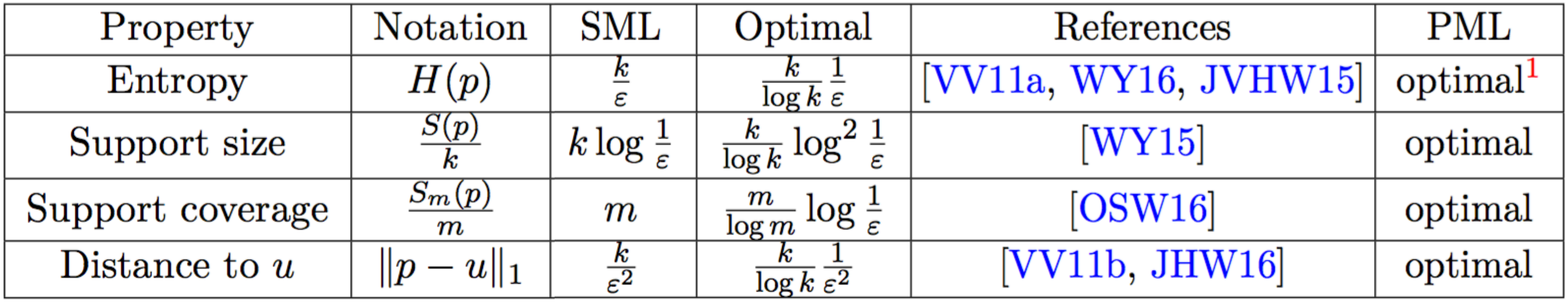
Known results summary

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.



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

Profiles

Profile of a sequence is the **multiset** of **multiplicities**:

$$\Phi(X_1^n) = \{N_x\}$$

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

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:

$$p(\Phi(X_1^n)) = \sum_{Y_1^n:\Phi(Y_1^n)=\Phi(X_1^n)} p(Y_1^n)$$

Maximize the profile probability:

$$p_{\Phi}^{PML} = \arg\max_{p} p(\Phi(X_1^n))$$

$X_1^n = (1H, 2T)$:

SML: $(2/3, 1/3)$

PML: $(1/2, 1/2)$

PML for symmetric properties

To estimate a symmetric $f(p)$:

- Find $p^{PML}(\Phi(X_1^n))$
- Output $f(p^{PML})$

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 $S\left(f\left(p^{PML}\right), 2\varepsilon, k, \delta \cdot \frac{e^{3\sqrt{n}}}{10}\right) \leq n$

If $\boldsymbol{n = S\left(f, \varepsilon, k, e^{-3\sqrt{n}}\right)}$, then $\boldsymbol{S\left(f\left(p^{PML}\right), 2\varepsilon, k, 0.1\right) \leq n}$

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 $p\big(\Phi(X^n)\big) > \delta$,

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

$p_{\Phi}^{PML}(\Phi) \geq p(\Phi) > \delta$

$\left|f\left(p_{\Phi}^{PML}\right) - f(p)\right| \leq \left|f\left(p_{\Phi}^{PML}\right) - \hat{f}(\Phi)\right| + \left|\hat{f}(\Phi) - f(p)\right| < 2\varepsilon$

- Profiles with $p\big(\Phi(X^n)\big) < \delta$,

$$p(p(\Phi(X^n) < \delta) < \delta \cdot \#profiles\ of\ length\ n$$

Ingredients

Better error probability guarantees.

Recall:

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

Stronger error guarantees using McDiarmid's inequality:

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

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

1. Acharya, J., Das, H., Orlitsky, A., & Suresh, A. T. (2016). A Unified Maximum Likelihood Approach for Optimal Distribution Property Estimation. *arXiv preprint arXiv:1611.02960*.
2. Paninski, L. (2003). Estimation of entropy and mutual information. *Neural computation*, *15*(6), 1191-1253.
3. Wu, Y., & Yang, P. (2016). Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Transactions on Information Theory*, *62*(6), 3702-3720.
4. Orlitsky, A., Suresh, A. T., & Wu, Y. (2016). Optimal prediction of the number of unseen species. *Proceedings of the National Academy of Sciences*, 201607774.
5. Orlitsky, A., Santhanam, N. P., Viswanathan, K., & Zhang, J. (2004, July). On modeling profiles instead of values. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence* (pp. 426-435). AUAI Press.
6. Valiant, G., & Valiant, P. (2011, June). Estimating the unseen: an n/log (n)-sample estimator for entropy and support size, shown optimal via new CLTs. In *Proceedings of the forty-third annual ACM symposium on Theory of computing* (pp. 685-694). ACM.
7. Jiao, J., Venkat, K., Han, Y., & Weissman, T. (2015). Minimax estimation of functionals of discrete distributions. *IEEE Transactions on Information Theory*, *61*(5), 2835-2885.
8. Vatedka, S., & Vontobel, P. O. (2016, July). Pattern maximum likelihood estimation of finite-state discrete-time Markov chains. In *Information Theory (ISIT), 2016 IEEE International Symposium on* (pp. 2094-2098). IEEE.
9. Vontobel, P. O. (2014, February). The Bethe and Sinkhorn approximations of the pattern maximum likelihood estimate and their connections to the Valiant-Valiant estimate. In *Information Theory and Applications Workshop (ITA), 2014* (pp. 1-10). IEEE.
10. Valiant, G., & Valiant, P. (2011, October). The power of linear estimators. In *Foundations of Computer Science (FOCS), 2011 IEEE 52nd Annual Symposium on* (pp. 403-412). IEEE.
11. Orlitsky, A., Sajama, S., Santhanam, N. P., Viswanathan, K., & Zhang, J. (2004). Algorithms for modeling distributions over large alphabets. In *Information Theory, 2004. ISIT 2004. Proceedings. International Symposium on* (pp. 304-304). IEEE.