Profile Maximum Likelihood: An Optimal, Universal, Plug-and-Play Functional Estimator

Yi Hao and Alon Orlitsky, UCSD

Outline

Property estimation

Plug-in estimators

Prior results

Profile maximum likelihood

Results

Simple, unified, optimal, plug-in, estimators for four learning tasks

Proof elements: The fun theorem of maximum likelihood

Local heroes

Discrete support set $\mathcal{X}$

$\{\text{heads, tails}\} = \{\text{h, t}\}$ $\{\ldots, -1, 0, 1, \ldots\} = \mathbb{Z}$

Distribution p over $\mathcal{X}$, probability p_x for $x \in \mathcal{X}$

$p_x \geq 0$ $\sum_{x \in \mathcal{X}} p_x = 1$

$p = (p_{\text{h}}, p_{\text{t}})$ $p_h = .6,\ p_t = .4$

$\mathcal{P}$ collection of distributions

$\mathcal{P}_{\mathcal{X}}$ all distributions over $\mathcal{X}$

$\mathcal{P}_{\{\text{h, t}\}} = \{(p_{\text{h}}, p_{\text{t}})\} = \{(.6, .4), (.4, .6), (.5, .5), (0, 1), \ldots\}$

$f : \mathcal{P}_{\mathcal{X}} \to \mathbb{R}$

Maps distribution to real value

Shannon entropy	$H(p)$	$\sum_x p_x \log \frac{1}{p_x}$
Rényi entropy	$H_\alpha(p)$	$\frac{1}{1-\alpha} \log \left(\sum_x p_x^\alpha\right)$
Support size	$S(p)$	$\sum_x \mathbb{1}_{p_x>0}$
Support coverage	$S_m(p)$	$\sum_x (1-(1-p_x)^m)$
Expected # distinct symbols in m samples		
Distance to uniformity	$L_{\text{uni}}(p)$	$\sum_x \left\lvert p_x - \frac{1}{\lvert\mathcal{X}\rvert}\right\rvert$
Highest probability	$\max(p)$	$\max\{p_x : x \in \mathcal{X}\}$
...		

Many applications

Given: support set $\mathcal{X}$, property f

Unknown: $p \in \mathcal{P}_{\mathcal{X}}$

Estimate: $f(p)$

Entropy of English words

Given: $\mathcal{X} = \{\text{English words}\}$, unknown: p, estimate: $H(p)$

species in habitat

Given: $\mathcal{X} = \{\text{bird species}\}$, unknown: p, estimate: $S(p)$

How to estimate $f(p)$ when p is unknown?

Observe n independent samples $X^n = X_1, \ldots, X_n \sim p$

Reveal information about p

Estimate $f(p)$

Estimator: $f^{\text{est}} : \mathcal{X}^n \to \mathbb{R}$

Estimate for $f(p)$: $f^{\text{est}}(X^n)$

Simplest estimators?

Plug-in Estimators

Simple two-step estimators

Use X^n to derive estimate $p^{\text{est}}(X^n)$ of p

Plug-in $f(p^{\text{est}}(X^n))$ to estimate $f(p)$

Hope: As $n \to \infty$, $p^{\text{est}}(X^n) \to p$, then $f(p^{\text{est}}(X^n)) \to f(p)$

Simplest p^{est}?

n samples

N_x # times x appears

$p_x^{\text{emp}} := \frac{N_x}{n}$

$\mathcal{X} = \{a, b, c\}$ $p = (p_a, p_b, p_c) = (.5, .3, .2)$

Estimate p from $n = 10$ samples

$X^{10} = c, a, b, a, b, a, b, a, b, c$

$p_a^{\text{emp}} = \frac{4}{10}, \quad p_b^{\text{emp}} = \frac{4}{10}, \quad p_c^{\text{emp}} = \frac{2}{10}$

$p^{\text{emp}} = (.4, .4, .2)$

Empirical Plug-In Estimator

$f^{\text{emp}}(X^n) = f(p^{\text{emp}}(X^n))$

Entropy estimation

$X^{10} = c, a, b, a, b, a, b, a, b, c$

$p^{\text{emp}} = (.4, .4, .2)$

$H^{\text{emp}}(X^{10}) := H(.4, .4, .2)$

Advantages

- Plug-and-play: simple two steps
- Universal: applies to all properties
- Intuitive

Best-known, most-used distribution estimator

Performance?

Min-max Probably Approximately Correct (PAC) Formulation

Allowed additive approximation error $\epsilon > 0$

Allowed error probability $\delta > 0$

$n_f(f^{\text{est}}, p, \varepsilon, \delta)$: # samples f^{est} needs to approximate f well, $|f^{\text{est}}(X^n) - f(p)| \leq \varepsilon$ with probability $\geq 1 - \delta$

$n_f(f^{\text{est}}, \mathcal{P}, \varepsilon, \delta) := \max_{p \in \mathcal{P}} n_f(f^{\text{est}}, p, \varepsilon, \delta)$: # samples f^{est} needs to approximate every $p \in \mathcal{P}$

$n_f(\mathcal{P}, \varepsilon, \delta) := \min_{f^{\text{est}}} n_f(f^{\text{est}}, \mathcal{P}, \varepsilon, \delta)$ # samples the best estimator needs to approximate all distributions in $\mathcal{P}$

Empirical and Optimal Sample Complexity

$|\mathcal{X}| = k, \quad \mathcal{P}_{\mathcal{X}}|$ all distributions

Property	$n_f(f^{\text{emp}}, \varepsilon, 1/3)$	$n_f(\varepsilon, 1/3)$
Entropy	$k \cdot \frac{1}{\varepsilon}$	$\frac{k}{\log k} \cdot \frac{1}{\varepsilon}$
Supp. coverage	m	$\frac{m}{\log m} \cdot \log \frac{1}{\varepsilon}$
Dist. to uniform	$k \cdot \frac{1}{\varepsilon^2}$	$\frac{k}{\log k} \cdot \frac{1}{\varepsilon^2}$
Support size	$k \cdot \log \frac{1}{\varepsilon}$	$\frac{k}{\log k} \cdot \log^2 \frac{1}{\varepsilon}$

P03, VV11a/b, WY14/19, JVHW14/18, AOST14, OSW16, ADOS17, PW 19,...

For support size, $\mathcal{P}_{\geq 1/k} := \{p \mid p_x \geq 1/k, \forall x \in \mathcal{X}\}$

Regime where $\varepsilon \gtrsim n^{-0.1}$

Support size and coverage normalized by k and m respectively

Why is empirical plugin good? suboptimal? optimal plug-in?

i.i.d. $p \in \mathcal{P}_{\mathcal{X}}$, probability of observing $x^n \in \mathcal{X}^n$

$$p(x^n) := \Pr_{X^n \sim p}(X^n = x^n) = \prod_{i=1}^{n} p(x_i)$$

Maximum likelihood estimator: $x^n \to$ dist. p maximizing $p(x^n)$

$$p^{\mathsf{ml}}(x^n) = \arg\max_p p(x^n)$$

$p^{\mathsf{ml}}(h,t,h) = \arg\max_{p_h+p_t=1} p_h^2 \cdot p_t$

$p_h = 2/3,\ p_t = 1/3$

Identical to empirical estimator – always

Empirical good: Distribution that best explains observation

Work wells for small alphabets large sample

Overfits data when alphabet is large relative to sample size

Improve?

What Counts

iid: Do not care about order

Entropy, Rényi, support size, coverage: symmetric functionals

Do not care about labels

(h,h,t), (t,t,h), (h,t,h), (t,h,t), (t,h,h), (h,t,t) same entropy

Care only: # of elements appearing any given number of times

Three samples: 1 element appeared once, 1 element appeared twice

Profile: $\varphi = \{1, 2\}$

Profile maximum likelihood (PML)

Profile $\varphi(x^n)$ of x^n is the multiset of symbol frequencies

$$b\,a\,n\,a\,n\,a\,s \implies a \text{ appears thrice, } n \text{ twice, } b\,s \text{ once}$$
$$\implies \varphi(bananas) = \{3, 2, 1, 1\}$$

Probability of observing a profile φ when sampling from p is

$$p(\varphi) := \sum_{y^n:\varphi(y^n)=\varphi} p(y^n) = \sum_{y^n:\varphi(y^n)=\varphi} \prod_{i=1}^{n} p(y_i)$$

Profile maximum likelihood maps x^n to

$$p^{\mathsf{ml}}_{\varphi(x^n)} := \operatorname*{argmax}_{p \in \mathcal{P}_{\mathcal{X}}} p(\varphi(x^n))$$

Observe $x^3 = h, t, h$

Sequence ML: $p_h = 2/3$, $p_t = 1/3$

Profile: $\varphi = \{1, 2\}$

Profile ML: maximize probability of $\varphi = \{1, 2\}$

$p, q \quad p + q = 1$

$\Pr(\varphi = \{1, 2\}) = ppq + qqp + pqp + qpq + qpp + pqq = 3(p^2q + q^2p)$

$\max(p^2q + q^2p) = \max(qp \cdot (p + q)) = \max pq$

Profile ML: $p = q = \frac{1}{2}$

More logical

More interesting?

RESULTS

Summary

Profile maximum likelihood (PML) is a
unified, time- and sample-optimal approach
to four basic learning problems

- Additive property estimation
- Rényi entropy estimation
- Sorted distribution estimation
- Uniformity testing

Yi Hao and Alon Orlitsky
The Broad Optimality of Profile Maximum Likelihood
Arxiv, NeurIPS 2019

Additive functional: $f(p) = \sum_x f(p_x)$

Entropy, support size, coverage, distance to uniformity

For all symmetric, additive, Lipschitz*, functionals, for $n \geq n_f(|\mathcal{X}|, \varepsilon, 1/3)$ and $\varepsilon \geq n^{-0.1}$,

$$\Pr\left(\left|f\left(p^{\mathsf{ml}}_{\varphi(X^{4n})}\right) - f(p)\right| > 5\varepsilon\right) \leq \exp(-\sqrt{n})$$

With four times the optimal # samples for error probability $1/3$, PML plug-in achieves much lower error probability

Covers four functionals above

Can use near-linear-time PML approximation [CSS19]

Rényi Entropy

For integer $\alpha > 1$, PML plug-in has optimal $k^{1-1/\alpha}$ sample complexity

For non-integer $\alpha > 3/4$, (A)PML plug-in improves best-known results

Sorted Distribution Estimation

Under ℓ_1 distance, (A)PML yields optimal $\Theta(k/(\varepsilon^2 \log k))$ sample complexity for sorted distribution estimation

Actual distribution in ℓ_1 distance, $2(k-1)/(\pi\varepsilon^2)$ [KOPS '15]

Uniformity testing: $p = p_u$ v.s. $|p - p_u| \geq \varepsilon$; complexity $\Theta(\sqrt{k}/\varepsilon^2)$

Tester below is sample-optimal up to logarithmic factors of k

Input: parameters $k, \varepsilon,$ and a sample $X^n \sim p$ with profile φ

If any symbol appears $\geq 3 \max\{1, n/k\} \log k$ times, return 1

If $\left\| p_\varphi^{\mathsf{ml}} - p_u \right\|_2 \geq 3\varepsilon/(4\sqrt{k})$, return 1; else, return 0

Thank you!