The Broad Optimality of Profile Maximum Likelihood

Yi Hao and Alon Orlitsky

University of California San Diego {yih179, alon}@ucsd.edu

Discrete Distributions

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

Distribution Property (Functional)

- $f : \mathcal{P}_{\mathcal{X}} \to \mathbb{R}$
- 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	$L_{\text{uni}}(p)$	$\sum_x \left\lvert p_x - \frac{1}{\lvert\mathcal{X}\rvert}\right\rvert$

Property Estimation

- 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?
- Many applications:
  vocabulary and population estimation, database similarity, graphical model learning, neural spike trains, property testing …

Learn from Examples

- Observe n independent samples
  $X^n = X_1,\ldots,X_n \sim p$
- Estimate $f(p)$
- Estimator: $f^{\text{est}} : \mathcal{X}^n \to \mathbb{R}$
- Estimate for $f(p)$: $f^{\text{est}}(X^n)$

Empirical Estimator

- n samples – X^n
- $p_x^{\text{emp}}(X^n) := (N_x\text{: \# times } x \text{ appears in } X^n)/n$
  $\mathcal{X} = \{a,b,c\}$ $p = (p_a,p_b,p_c) = (.5,.3,.2)$
  $X^{10} = c,a,b,a,b,a,b,a,b,c$
  $p_a^{\text{emp}} = \frac{4}{10}$, $p_b^{\text{emp}} = \frac{4}{10}$, $p_c^{\text{emp}} = \frac{2}{10}$
- Maximum likelihood estimator (MLE): $p^{\text{ml}} = p^{\text{emp}}$

Empirical Plug-In Estimator

- $f^{\text{emp}}(X^n) = f(p^{\text{emp}}(X^n))$
- Advantages:
  Plug-and-play: simple two steps
  Universal and Intuitive: applies to all properties
- Best-known, most-used property estimator

Sample Complexity

- Probably Approximately Correct (PAC)
- Allowed additive approximation error $\epsilon > 0$
- Allowed error probability $\delta > 0$
- $n_f(f^{\text{est}},\varepsilon,\delta)$: number of samples f^{est} needs to approximate property f well:
  $$|f^{\text{est}}(X^n) - f(p)| \leq \varepsilon$$
  with probability $\geq 1-\delta$, for all $p \in \mathcal{P}$

Empirical and Optimal Complexity

- $\mathcal{P}_k$ all k-symbol distributions, $\varepsilon \gtrsim n^{-0.1}, \delta = 1/3$

Property	$n_f(f^{\textbf{emp}},\varepsilon)$	$n_f(f^{\textbf{opt}},\varepsilon)$
Entropy	$k\cdot\frac{1}{\varepsilon}$	$\frac{k}{\log k}\cdot\frac{1}{\varepsilon}$
Support coverage	m	$\frac{m}{\log m}\cdot\log\frac{1}{\varepsilon}$
Distance to uniform	$k\cdot\frac{1}{\varepsilon^2}$	$\frac{k}{\log k}\cdot\frac{1}{\varepsilon^2}$
Support size	$k\cdot\log\frac{1}{\varepsilon}$	$\frac{k}{\log k}\cdot\log^2\frac{1}{\varepsilon}$

- For support size, $\mathcal{P}_{\geq 1/k} := \{p \mid p_x \geq 1/k, \forall x \in \mathcal{X}\}$
- Support size and coverage normalized by k and m

Profiles

- iid: order doesn't matter
- Symmetric properties: labels don't matter
- Profile: # elements appearing any given # times
- (h,h,t), (t,t,h), (h,t,h), (t,h,t), (t,h,h) same entropy
  1 element appeared once, 1 element twice
  Profile: $\varphi = \{1,2\}$
- $\varphi(x^n)$: multiset of symbol frequencies in x^n

Profile Probability

- Probability of observing φ when sampling from p
  $$p(\varphi) := \sum_{y^n:\varphi(y^n)=\varphi} p(y^n) = \sum_{y^n:\varphi(y^n)=\varphi} \prod_{i=1}^n p(y_i)$$
- p, q $p+q=1$
  $\Pr(\varphi = \{1,2\}) = 3(p^2q + q^2p)$

Profile Maximum Likelihood (PML)

- Distribution maximizing profile probability
- Maps x^n to $p^{\text{pml}}_{\varphi(x^n)} := \operatorname{argmax}_{p\in\mathcal{P}} p(\varphi(x^n))$

Broad optimality of PML

PML – unified, time- and sample-optimal for

- Additive property estimation
- Non-additive property – Rényi entropy estimation
- Sorted distribution estimation
- Identity/Uniformity testing

Additive Property Estimation

- Additive property: $f(p) = \sum_x f_x(p_x)$
  entropy, support size, distance to uniformity …
- *For **all** symmetric, additive, properly Lipschitz, properties, for $n \geq n_f(|\mathcal{X}|,\varepsilon,1/3)$ and $\varepsilon \gtrsim n^{-0.1}$,*
  $$\Pr\left(\left|f\left(p^{pml}_{\varphi(X^{4n})}\right) - f(p)\right| > 5\varepsilon\right) \leq \exp(-\sqrt{n})$$
- With **four** times the **optimal** # samples for error probability 1/3, PML plug-in achieves **much lower** error probability
- Near **linear-time** (A)PML approximation [CSS19]

Rényi, Distribution, Testing

- **Rényi entropy**
  Integer $\alpha > 1$, PML has optimal $k^{1-1/\alpha}$ complexity
  Non-integer $\alpha > 3/4$, PML improves best-known
- **Sorted distribution estimation**
  (A)PML yields optimal $\Theta(k/(\varepsilon^2\log k))$ sample complexity under ℓ_1 distance
- **Uniformity testing**: $p = p_u$ v.s. $|p - p_u| \geq \varepsilon$; optimal $\Theta(\sqrt{k}/\varepsilon^2)$ up to logarithmic factors of k:

Input: params $k,\varepsilon,$ and a sample with profile φ
If $\exists N_x \geq 3\max\{1,n/k\}\log k$, return 1
If $\left\lVert p^{\text{pml}}_\varphi - p_u\right\rVert_2 \geq 3\varepsilon/(4\sqrt{k})$, return 1; else, 0