Unified Sample-Optimal Property Estimation in Near-Linear Time

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 $x \in \mathcal{X}$
  $p_x \ge 0$ $\sum_{x\in\mathcal{X}} p_x = 1$
  $p = (p_h, p_t)$ $p_h = .6,\ p_t = .4$
- $\mathcal{P}$ collection of distributions
- $\mathcal{P}_\mathcal{X}$ all distributions over $\mathcal{X}$
  $\mathcal{P}_{\{h,\,t\}} = \{(p_h, p_t)\} = \{(.6, .4), (.4, .6), (.5, .5), (0, 1), \ldots\}$

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	$S(p)$	$\sum_x 1_{p_x>0}$
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:

$$|f^{\text{est}}(X^n) - f(p)| \le \varepsilon$$

  with probability $\ge 1 - \delta$, for all $p \in \mathcal{P}$

Empirical and Optimal Complexity

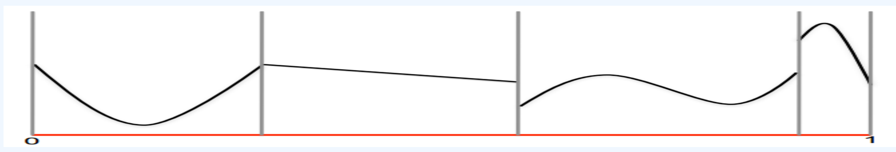
- $\mathcal{P}_k$ all k-symbol distributions, $\varepsilon \gtrsim n^{-0.1}, \delta = 1/3$

Property	$n_f(f^{\text{emp}}, \varepsilon)$	$n_f(f^{\text{opt}}, \varepsilon)$
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, $\mathcal{P}_{\ge 1/k} := \{p \mid p_x \ge 1/k, \forall x \in \mathcal{X}\}$
- Support size and coverage normalized by k and m
- *The empirical plug-in is suboptimal*

Piecewise Polynomials

- A function that is a (possibly different) polynomial on each of several sub-domains



- Extensively used in statistical inference tasks
  regression, density estimation, time series analysis

Additive Property Estimation

- Additive property $f(p) = \sum_x f_x(p_x)$
  entropy, support size, distance to uniformity …
- **Algorithm sketch**

1. Given X^n, compute its empirical distribution p^{emp}
2. For each symbol x, use p_x^{emp} to identify the $(1-1/n)$-confidence interval I_x for p_x
3. Approximate f_x over I_x by a proper polynomial g_x
4. Estimate $g_x(p_x)$ unbiasedly and sum up estimates

- *Implicitly uses piecewise polynomials*

Optimal (ε, δ)-Complexity

- For concreteness, entropy, other properties in paper
- Median trick: $\log(1/\delta)$ independent copies, take median to boost the confidence from $2/3$ to $1 - \delta$
- The median-trick complexity bound

$$n_f(f^{\text{med}}, \varepsilon, \delta) \lesssim \log\frac{1}{\delta} \cdot \frac{k}{\varepsilon \log k} + \log\frac{1}{\delta} \cdot \frac{\log^2 k}{\varepsilon^2}$$

Our estimator f^* achieves

$$n_f(f^*, \varepsilon, \delta) \lesssim \frac{k}{\varepsilon \log k} + \left(\log\frac{1}{\delta} \cdot \frac{1}{\varepsilon^2}\right)^{1.01}$$

- A nearly-matching lower bound with $1.01 \to 1$

Lipschitz Property Estimation

- $f(p) = \sum_x f_x(p_x)$ is L-Lipschitz
  All functions f_x have Lipschitz constants $\le L$
- **Optimal sub-linear complexity bound**

$$n_f(f^{\text{opt}}, \varepsilon, 1/3) \lesssim L^2 \cdot \frac{k}{\varepsilon^2 \log k}$$

- Previously, such a generic bound was proved only for symmetric and a few non-symmetric properties
- The estimator is highly concentrated, yielding near-optimal *differentially private* estimators

Poisson-McDiarmid Inequality

- Poisson sampling Sample size $n \to N \sim \text{Poi}(n)$
- Independent symbol counts
- Bounded difference property
  For all m and x^m, changing, adding, or deleting one symbol in x^m changes $f(x^m)$ by at most c
- **Inequality:** Let $c_* = 8\max\{c, n^{-1}\}$. $\forall\, \varepsilon > 0$,

$$\Pr\left(\left|f(X^N) - \mathbb{E}[f(X^N)]\right| > \varepsilon\right) \le 4\exp\left(-\frac{2\varepsilon^2}{nc_*^2}\right)$$