# Estimating Symmetric Distribution Properties
# Maximum Likelihood Strikes Back!

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

Shannon Channel
March 5, 2018

Based on:

*A Unified Maximum Likelihood Approach for Estimating Symmetric Properties of Discrete Distributions*,

International Conference on Machine Learning (ICML), 2017

http://proceedings.mlr.press/v70/acharya17a.html

Slides available at:

https://people.ece.cornell.edu/acharya/talks/shannon-18acharya.pdf

# Outline

1. Motivation
2. Methods
3. Proofs
4. Future directions

# Chapter 1: **Motivation**

# Distribution properties

$\mathcal{P}$: a collection of discrete distributions

- $\mathcal{P} = \Delta_k$: all distributions over $[k] = \{1, \ldots, k\}$
- $\Delta_6$: distributions over $[6]$

**Property**

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

- $p(3) = ?$
- Is it fair? Is $p(i) = 1/6$ for all $i$?

# Property estimation

$p$ **unknown** distribution in $\mathcal{P}$

Given independent samples $X_1^n = X_1, X_2, \ldots, X_n \sim p$

Estimate $f(p)$

*Sample complexity* $S(f, \mathcal{P}, \varepsilon, \delta)$

Minimum $n$ necessary to

Estimate $f(p) \pm \varepsilon$

With error probability $< \delta$ (usually constant)

# Symmetric properties

$f$ **symmetric** if unchanged under input permutations

Entropy $H(p) \triangleq \sum_x p(x) \log \frac{1}{p(x)}$

Support size $S(p) \triangleq \sum_i \mathbb{I}_{\{p(x)>0\}}$

Many others: Renyi entropy, support coverage, …

Coins with bias 0.4, and with bias 0.6 have same entropy!

# Entropy

$$H(p) \triangleq -\sum_{x} p(x) \log p(x)$$

- Most popular measure of randomness
- Central quantity in information theory [Shannon'48]

How many samples to estimate $H(p)$ to $\pm\varepsilon$?

Long line of work: [Empirical, Miller-Maddow, Jackknifed, Coverage adjusted, BUB (paninski'03), …]

# Estimating entropy

- Randomness of neural spike trains
- Feature selection in decision trees
- Graphical models (Chow-Liu)

Traditional setting for property estimation:

$\mathcal{P} = \Delta_k$: distributions on $[k]$ (small $k$)

Obtain many samples (large $n$)

Genetics, neural spikes, text, computer vision, ecology:

**$k$ large, possibly infinite, perhaps unknown**

# Estimating the unseen: Corbet's butterflies

2 years trapping butterflies in Malay peninsula:

| **Frequency** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Species** | **118** | **74** | **44** | **24** | **29** | **22** | **20** | **19** | **20** | **15** | **12** | **14** | **6** | **12** | **6** |

Asked Fisher:

**how many new species if he goes for two more years?**

# Estimating the unseen: formulation

$p$: **unknown** discrete distribution

$$S_m(p) \triangleq \mathbb{E}[\#\textbf{distinct symbols} \text{ in } m \text{ ind. samples} \sim p]$$

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

Normalized coverage: $\frac{S_m(p)}{m}$

How many samples to estimate $\frac{S_m(p)}{m}$ to $\pm\varepsilon$?

# Estimating the unseen: applications

- Estimating vocabulary size
- Ecological diversity
- Microbial diversity on skin

Well studied [Good Toulmin, Efron Thisted] For constant $\varepsilon$:

$$\text{Requires } \frac{m}{2} \text{ samples}$$

More recently, [Zou Valiant Valiant Chan …'16, Orlitsky Suresh Wu '16]:

$$\text{Requires } \frac{m}{\log m} \text{ samples}$$

# Chapter 2: **Methods**

# Plug-in estimation

Using $X_1^n$, find an estimate $\hat{p}$ of $p$

Estimate $f(p)$ by $f(\hat{p})$

How to estimate $p$?

# Sequence maximum likelihood (SML)

$$p_{x^n}^{\text{sml}} \triangleq \arg\max_p \; p(x^n) = \arg\max_p \; \prod_i p(x_i)$$

$X_1^3 = h, h, t$

$$p^{sml} = \arg\max \; p^2(h) \cdot p(t)$$

$$p^{sml}(h) = \frac{2}{3}, \; p^{sml}(t) = \frac{1}{3}$$

Same as the *empirical-frequency* distribution

Multiplicity $N_x$ - # times $x$ appears in $X_1^n$

$$p^{\text{sml}}(x) = \frac{N_x}{n}$$

# SML for entropy

Sample complexity of SML to estimate $H(p)$ over $\boldsymbol{\Delta}_k$ :

$$S^{sml}(H, \boldsymbol{\Delta}_k, \varepsilon) = \Theta\left(\frac{k}{\varepsilon}\right)$$

In the asymptotic $n \to \infty$, SML is optimal

# Sample complexity of entropy

Sample complexity of $H(p)$:

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

[…, Paninski'03, Valiant Valiant'11, Han Jiao Venkat Weissman'15, Wu Yang'15]

[Valiant Valiant '11]: plug-in, **sub-optimal** in $\varepsilon$

[Han Jiao Weissman '18]: optimal in $\varepsilon$ by tweaking VV'11

# Optimal estimators

General recipe:

1. Approximate $H(p)$ with a polynomial in $p$
2. Estimate the polynomial

Different non-plug-in estimator for each property

Sophisticated approximation theory results

# Prior work

For several important properties

Optimal is a **logarithmic factor** better than empirical

| Property | $\mathcal{P}$ | SML | Optimal | References |
|---|---|---|---|---|
| $H(\boldsymbol{p})$ | $\boldsymbol{\Delta}_k$ | $\frac{k}{\varepsilon}$ | $\frac{k}{\varepsilon \cdot \log k}$ | Valiant Valiant '11, Han Jiao Venkat Weissman '15, Wu Yang '15 |

# Prior work

For several important properties

Optimal is a **logarithmic factor** better than empirical

| Property | $\mathcal{P}$ | SML | Optimal | References |
|---|---|---|---|---|
| $H(\boldsymbol{p})$ | $\boldsymbol{\Delta}_k$ | $\frac{k}{\varepsilon}$ | $\frac{k}{\varepsilon \cdot \log k}$ | Valiant Valiant '11, Han Jiao Venkat Weissman '15, Wu Yang '15 |
| $\frac{S_m(\boldsymbol{p})}{m}$ | $\boldsymbol{\Delta}_\infty$ | $m$ | $\frac{m}{\log m} \log \frac{1}{\varepsilon}$ | Orlitsky Suresh Wu'16 |

# Prior work

For several important properties

Optimal is a **logarithmic factor** better than empirical

| Property | $\mathcal{P}$ | SML | Optimal | References |
|---|---|---|---|---|
| $H(\boldsymbol{p})$ | $\boldsymbol{\Delta}_k$ | $\frac{k}{\varepsilon}$ | $\frac{k}{\varepsilon \cdot \log k}$ | Valiant Valiant '11, Han Jiao Venkat Weissman '15, Wu Yang '15 |
| $\frac{S_m(\boldsymbol{p})}{m}$ | $\boldsymbol{\Delta}_\infty$ | $m$ | $\frac{m}{\log m}\log\frac{1}{\varepsilon}$ | Orlitsky Suresh Wu'16 |
| $\frac{S(\boldsymbol{p})}{k}$ | $\boldsymbol{\Delta}_k$ | $k\log\frac{1}{\varepsilon}$ | $\frac{k}{\log k}\log^2\frac{1}{\varepsilon}$ | Wu Yang '16 |
| $\| \boldsymbol{p} - u \|_1$ | $\boldsymbol{\Delta}_k$ | $\frac{k}{\varepsilon^2}$ | $\frac{k}{\varepsilon^2 \cdot \log k}$ | Han Jiao Weissman '16 |

# Our results [Acharya Das Orlitsky Suresh '17]

Unified, simple, sample-optimal approach for all above problems

- Plug-in estimator
- **Maximum likelihood** principle:

~~sequence maximum likelihood (SML)~~

**profile** maximum likelihood (PML)

# Chapter 3: PML

# Profiles

Profile is the multi-set of multiplicities

$$\Phi(X_1^n) \triangleq \{N_x : x \in X_1^n\}$$

$$\Phi(h,h,t) = \Phi(t,h,t) = \{1,2\}$$

$$\Phi(\alpha,\gamma,\beta,\gamma) = \{1,1,2\}$$

# Probability multiset

Symmetric properties determined by

- Probability multiset: $\{p(1), p(2), \dots\}$

$p(h) = 0.4 \Longrightarrow \{0.4, 0.6\}$

$p(h) = 0.6 \Longrightarrow \{0.4, 0.6\}$

**Profiles are sufficient statistic for symmetric properties**

$h, h, t$, **OR** $t, h, t \Longrightarrow$ same estimate

# Estimating probability multiset

Orlitsky Santhanam Viswanathan Zhang ‘04:

“On modeling profiles instead of values”, UAI

**More extensively:**

OSVZ: “On estimating a probability multiset”, online

# Profile maximum likelihood (PML)

Profile probability

$$p(\Phi) = \sum_{x^n:\, \Phi(x^n)=\Phi} p(x^n)$$

**Distribution maximizing the profile probability**

$$p_\Phi^{pml} = \arg\max_{p \in \mathcal{P}} p(\Phi)$$

# PML example

$X_1^3 = h, h, t$

$\Phi(h, h, t) = \{1,2\}$

$$p(\Phi = \{1,2\})$$

$$= p(s, s, d) + p(s, d, s) + p(d, s, s)$$

$$= 3 \cdot p(s, s, d)$$

$$= 3 \cdot \left(\sum_{x \neq y} p^2(x) p(y)\right)$$

Symmetric polynomial

# SML of $\{1,2\}$

$$p(\Phi = \{1,2\}) = 3\left(\sum_{x \neq y} p^2(x)p(y)\right)$$

$$p^{sml}(h) = \frac{2}{3}, p^{sml}(t) = \frac{1}{3}$$

$$p^{sml}(\{1,2\}) = 3\left(\left(\frac{2}{3}\right)^2\left(\frac{1}{3}\right) + \left(\frac{1}{3}\right)^2\left(\frac{2}{3}\right)\right) = \frac{18}{27} = \frac{2}{3}$$

# PML of $\{1,2\}$

$$p(\Phi = \{1,2\}) = 3\left(\sum_{x \neq y} p^2(x)p(y)\right)$$

$$\sum_{x \neq y} p^2(x)p(y) = \sum_{x} p^2(x)(1 - p(x))$$

$$= \sum_{x} p(x) \cdot p(x)(1 - p(x)) \leq \frac{1}{4}$$

$$p^{pml}(\{1\ ,2\}) = \frac{3}{4}$$

# PML of $\{1,1,2\}$

$\Phi(\alpha,\gamma,\beta,\gamma) = \{1,1,2\}$

$p^{pml}(\{1,1,2\}) = U[5]$

**PML can predict existence of unseen symbols**

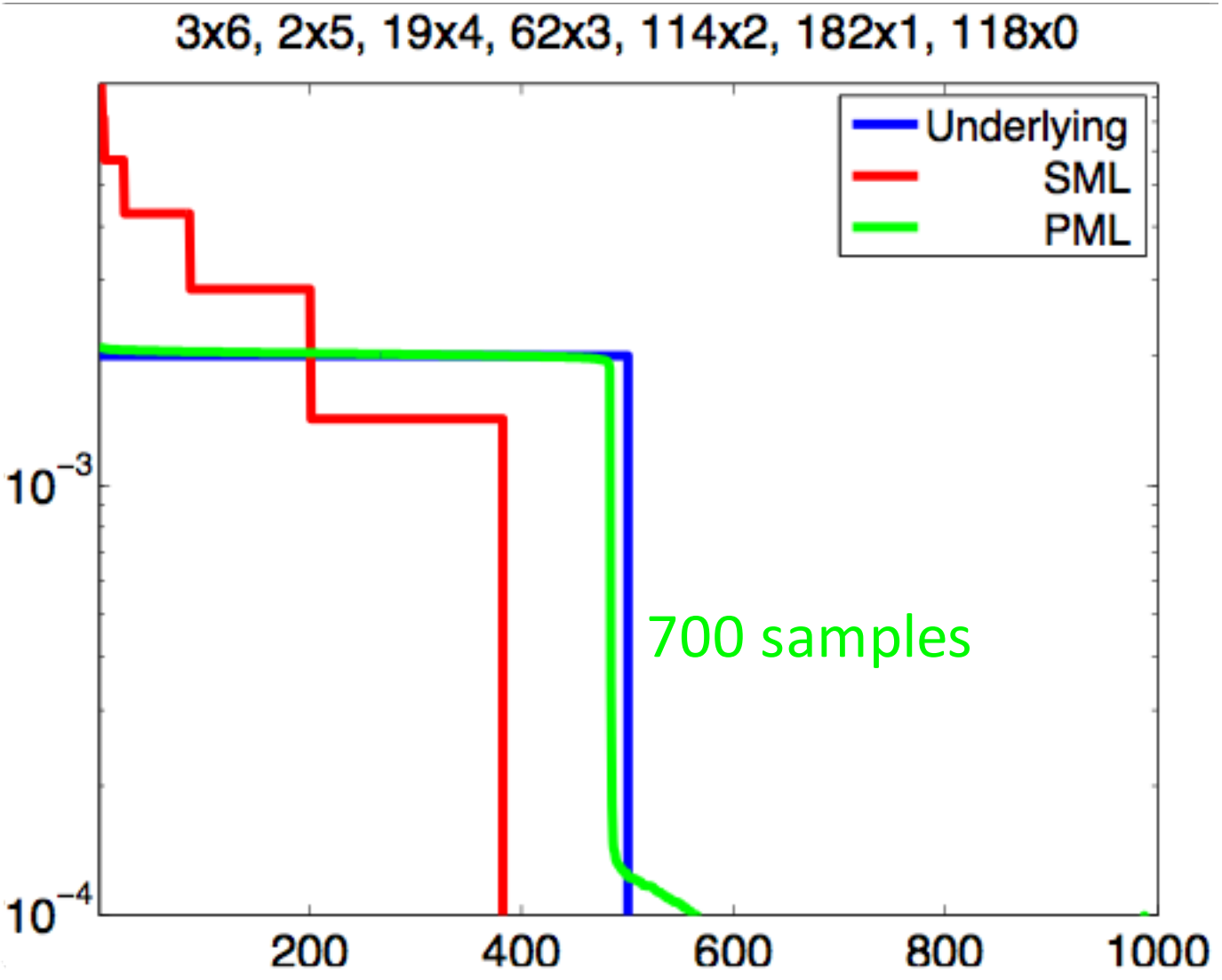
Maximize:

$$\sum_{x \neq y \neq z} p(x)^2 p(y) p(z)\,,$$

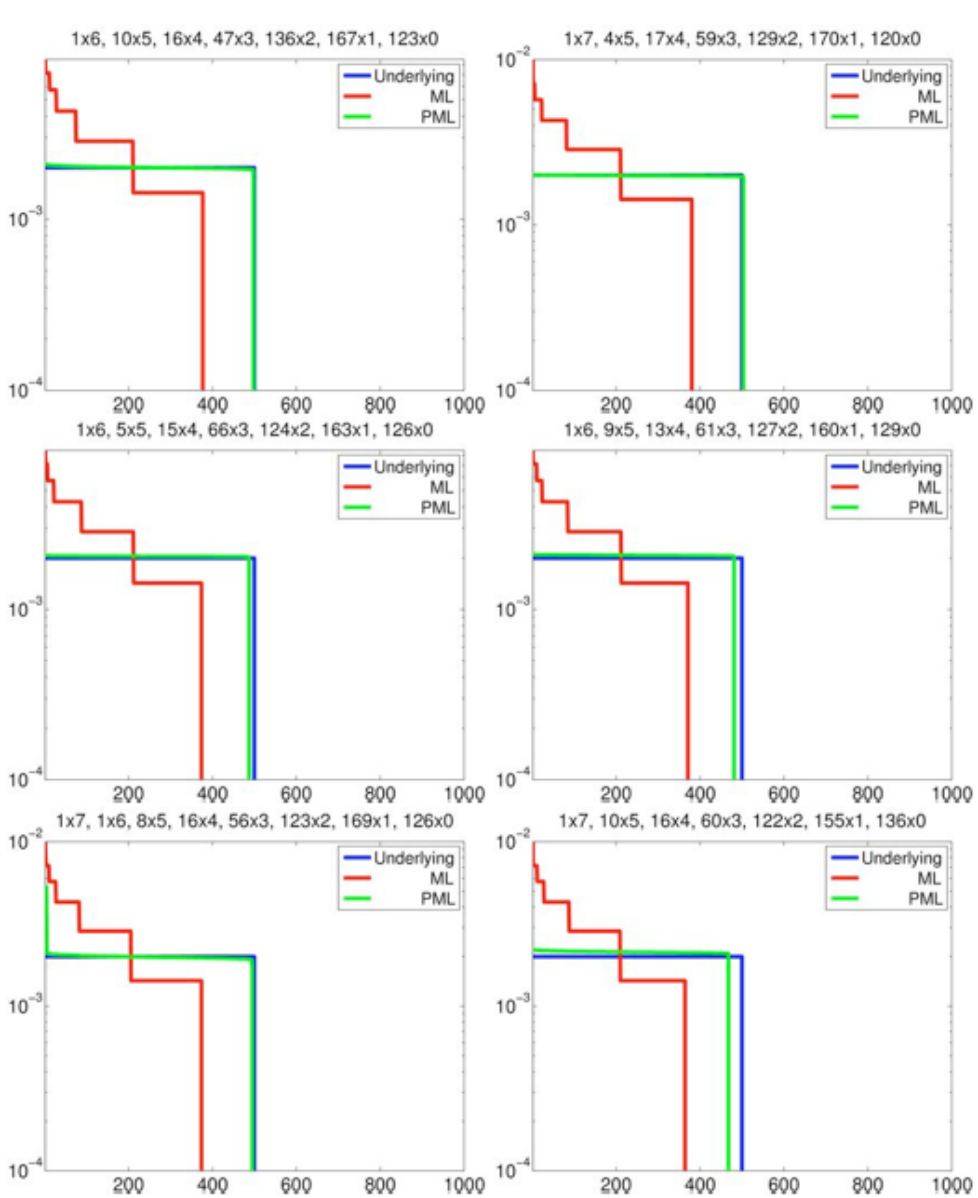
subject to:

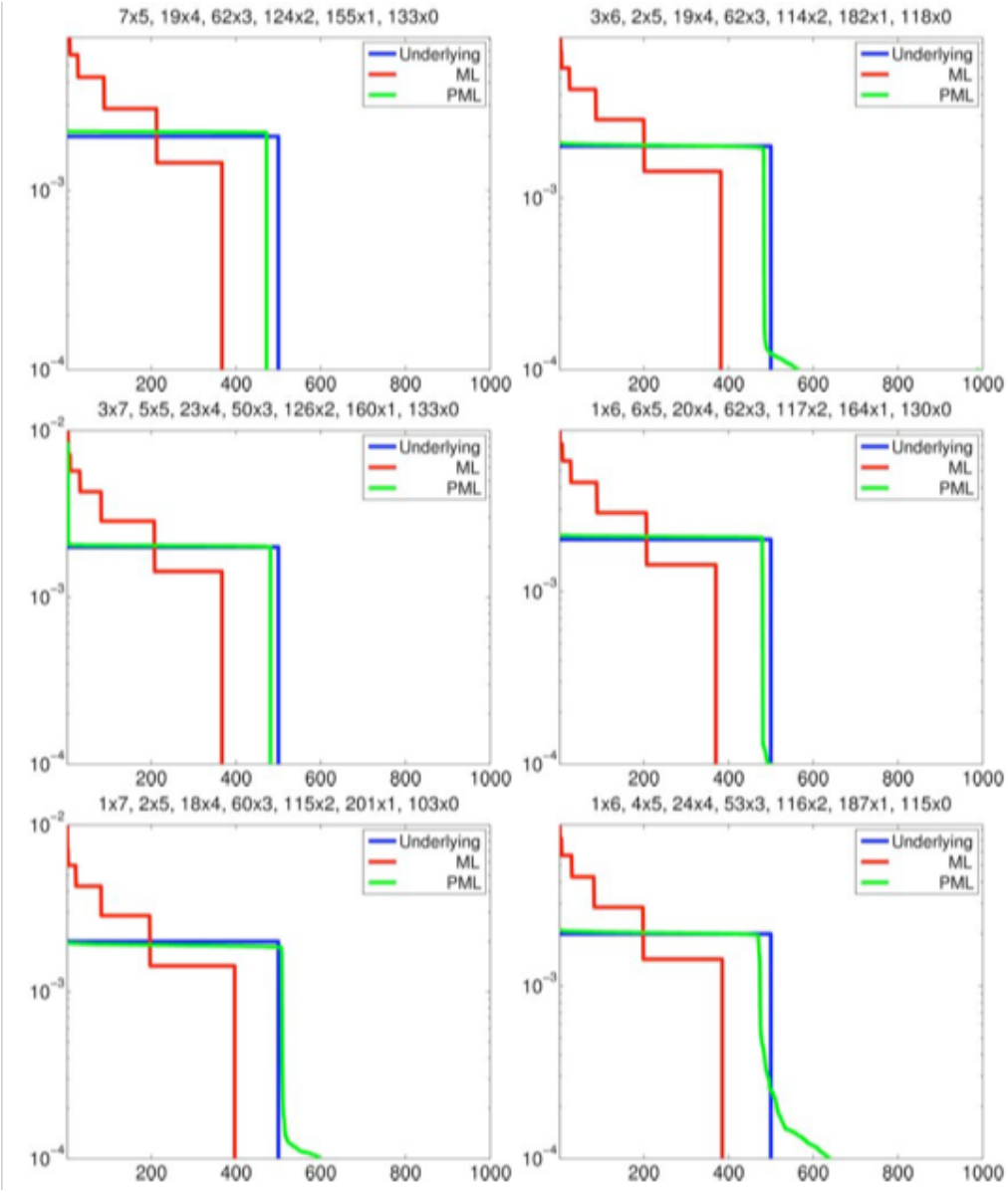
$$\sum_{x} p(x) = 1, p(x) \geq 0$$

# Uniform [500], 700 samples



# U[500], 700x, 12 experiments



# PML plug-in

To estimate symmetric $f(p)$:

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

Advantages:

- No tuning parameters
- Not function specific

Rooted in the maximum likelihood principle

# Ingredient 1: Goodness of ML

# General ML plugin estimation

$\mathcal{P}$: collection of distributions over an **abstract** domain $\mathcal{Z}$

$f: \mathcal{P} \to \mathbb{R}$ any property

Given $z \in \mathcal{Z}$ estimate $f$

ML estimator:

Determine $p_z^{\mathrm{ML}} \triangleq \arg\max_{p \in \mathcal{P}} p(z)$

Output $f(p_z^{\mathrm{ML}})$

**How good is MLE?**

# Competitiveness of ML plugin

**Theorem**: Suppose $\hat{f}: \mathcal{Z} \to \mathbb{R}$ is such that $\forall\, p \in \mathcal{P}$,

$$\Pr_{Z \sim p}\left(\left|\hat{f}(Z) - f(p)\right| > \varepsilon\right) < \delta,$$

then MLE plugin error bounded by

$$\Pr_{Z \sim p}\left(\left|f\left(p_Z^{\mathrm{ML}}\right) - f(p)\right| > 2 \cdot \varepsilon\right) < \delta \cdot |\mathcal{Z}|.$$

**Competitive** with the best $\hat{f}$

# Competitiveness of MLE plugin - proof

Consider any $p \in \mathcal{P}$

$\mathcal{Z}_{\geq\delta} \triangleq \{z \in \mathcal{Z}: p(z) \geq \delta\}$

- $z \in \mathcal{Z}_{\geq\delta}$:
  - $\left|\hat{f}(z) - f(p)\right| \leq \varepsilon$ (by condition in Theorem)
  - $p_z^{\mathrm{ML}}(z) \geq p(z) \geq \delta$, hence $\left|\hat{f}(z) - f(p_z^{\mathrm{ML}})\right| \leq \varepsilon$
  - Triangle inequality: $\left|f(p_z^{\mathrm{ML}}) - f(p)\right| \leq 2\varepsilon$

- $z \in \mathcal{Z}_{<\delta} \triangleq \{z: p(z) < \delta\}$
  - $\Pr\left(\left|f\left(p_z^{\mathrm{MLE}}\right) - f(p)\right| > 2\varepsilon\right) \leq \Pr(\mathcal{Z}_{<\delta}) < \delta \cdot |\mathcal{Z}|$

# Ingredient 2: **Error probabilities**

# PML performance bound

**Theorem**: If $n = S(f, \mathcal{P}, \varepsilon, \delta)$, then

$$S^{pml}(f, \mathcal{P}, 2 \cdot \varepsilon, |\Phi^n| \cdot \delta) \leq n$$

$|\Phi^n|$: number of profiles of length $n$

Profile of length n: partition of n

{3}, {1,2}, {1,1,1} → 3, 2+1, 1+1+1

$|\Phi^n| =$ partition # of $n$

Hardy-Ramanujan: $|\Phi^n| < e^{3\sqrt{n}}$

# PML performance: Try 1

**Theorem**: $n = S(f, \mathcal{P}, \varepsilon, 1/3) \Rightarrow S^{pml}(f, \mathcal{P}, 2\varepsilon, 1/3) \leq O(n^2)$.

**Proof**:

- Boost error probability:
  - Take $n \cdot \ell$ independent samples, and divide them in $\ell$ parts
  - Estimate $f$ for each of the samples
  - Take the median of the estimates $\Longrightarrow \delta < \exp(-\ell)$
- $\left|\Phi^{n\ell}\right| < \exp(3(n\ell)^{\cdot 5})$

Error probability of PML most $\exp(-\ell + 3(n\ell)^{\cdot 5})$

When $\ell > n$, the error dominates # profiles

# PML performance: Try 2

Recall

$$S(H, \Delta_k, \varepsilon, 1/3\ ) = \Theta\left(\frac{k}{\varepsilon \cdot \log k}\right)$$

With twice the samples error drops **exponentially**

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

- **Modified** estimators with small bounded differences
- Stronger guarantees from McDiarmid's inequality

Most technical part of the paper

Similar results for other properties

# Combining everything

Fast error for properties we study:

If $\boldsymbol{n = S(f, \mathcal{P}, \varepsilon, 1/3)}$, then $\boldsymbol{S\left(f, \mathcal{P}, \varepsilon, e^{-4\sqrt{n}}\right) \leq 4n}$

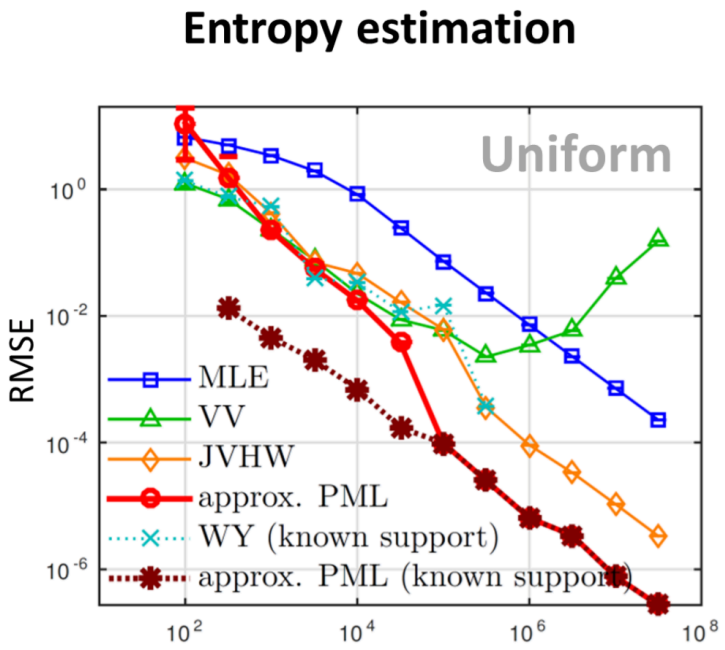
ML plug-in result:

If $\boldsymbol{S\left(f, \mathcal{P}, \varepsilon, e^{-4\sqrt{n}}\right) \leq 4n}$, then $\boldsymbol{S^{pml}\left(f, \mathcal{P}, 2\varepsilon, e^{-\sqrt{n}}\right) \leq 4n}$

Combining, we are done!

# Computing PML distribution

- EM algorithm [Orlitsky Pan Sajama Santhanam Viswanathan Zhang '04,'13]
- Approximate PML via Bethe Permanents [Vontobel'14]
- Extensions of Markov Chains [Vatedka Vontobel'16]
- Approximation via relaxation [Jiao Pavlichin Weissman '17]:



# Chapter 5: **Directions**

# Approximate PML

- Perhaps finding exact PML is hard
- Can show that approximating PML enough

**Question:**

Compute $\exp(-n^{\beta})$ approximate PML for any $\beta < 1$

Even this is optimal (for large $k$)

# Higher dimensions

- Estimate KL divergence between discrete distributions given samples (under assumptions of course)

[Bu Zou Liang Veeravalli ‘16, Han Jiao Weissman ’16]

[Acharya ’18]: PML is optimal for KL divergence estimation

- Higher order partitions

# Independent Proof Techniques

MLE good when something else is good

- **ML performance independent of other results?**

# Other

Is PML optimal for **every** symmetric property?

Can we do something for continuous distributions?

# Summary

- Symmetric property estimation
- PML plug-in approach
  - Universal, simple to state
  - Independent of particular properties

# In Fisher's words …

Of course nobody has been able to prove that MLE is best under all circumstances. MLE 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

Thank You!