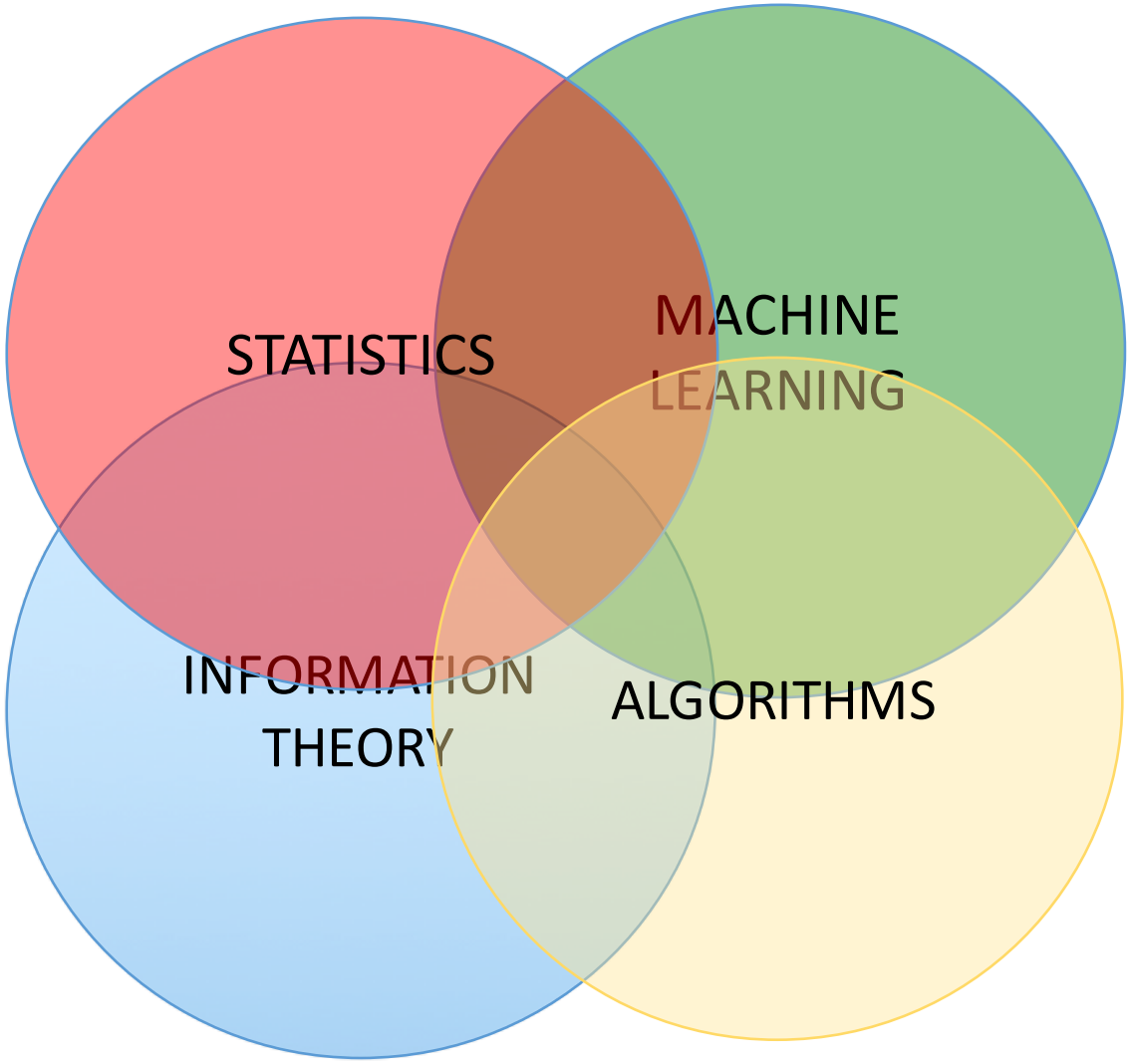
Theoretical Elements of Data Science

Jayadev Acharya, Alon Orlitsky, Ananda Theertha Suresh

ISIT Tutorial, 10 July 2016

Data Science



Information Theory

- Relevant models
- Clean formulations
- Theoretical limits
- Optimal algorithms
- Useful insights

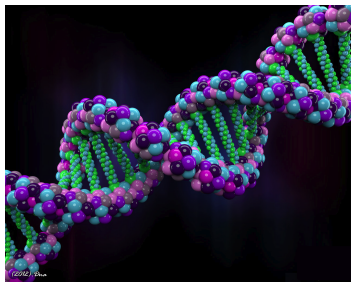
- Problems
  - New
  - Simple
  - Beautiful
  - Meaningful

Outline

- Distribution estimation
  - Min-max formulation
  - Competitive formulatation
- Property estimation
  - Entropy estimation
  - Estimating the unseen
- Property testing
  - Uniformity testing
  - Testing for class
- Motivation, formulation, ideas, simulations, problems

Old Problems, New Challenges

Old questions



Small domain

Many samples

Asymptotic analysis

Computation not crucial

Large domain

Not enough samples

New challenges

Part 1: Distribution estimation

Discrete Distributions

- Support set $\mathscr{X}$
- Distribution $p : \mathscr{X} \to \mathbb{R}_{\geq 0} \quad \sum_{x \in \mathscr{X}} p(x) = 1$
- Sequence of samples: $x^n = x_1, \ldots, x_n$
- i.i.d. distributions: $p(x^n) = \prod_{i=1}^{n} p(x_i)$

- Without loss of generality, $\mathscr{X} = \{1, \ldots, k\}$
- Distribution: $p = (p_1, \ldots, p_k)$
- Simplex in $\mathbb{R}^k$: $\Delta_k = \{(p_1, \ldots, p_k)\}$

Distribution Estimators

- Observe n samples x^n, estimate distribution
- Distribution estimator: $q : \{1, \ldots, k\}^n \to \Delta_k$
- $q(x^n)$ or q_{x^n}: estimate upon observing x^n
- For Δ_k: $q_{x^n} = (q_{x^n,1}, \ldots, q_{x^n,k})$

Empirical Estimator q^{emp}

- Estimated prob. of symbol: fraction of times it appeared
- $t_i(x^n)$: # times i appears in x^n
- $x^3 = 311$: $\quad t_1(311) = 2$, $t_2(311) = 0$, $t_3(311) = 1$
- $\sum_{i=1}^{k} t_i(x^n) = n$
- $q^{\text{emp}}_{x^n,i} \stackrel{\text{def}}{=} t_i(x^n)/n$
- $q^{\text{emp}}_{311,1} = 2/3, \quad q^{\text{emp}}_{121,2} = 0, \quad q^{\text{emp}}_{311,3} = 1/3$
- $\sum_{i=1}^{k} q^{\text{emp}}_{x^n,i} = \sum_{i=1}^{k} t_i(x^n)/n = n/n = 1$
- $q^{\text{emp}} \in \Delta_k$

Performance Evaluation

- $L(p, q)$: loss (risk) for estimating p by q
  - Euclidean distance, KL Divergence, ...
- Which x^n?
- Expected loss: $\mathop{\mathbb{E}}_{X^n \sim p} L(p, q_{X^n})$
- Which p?
- Worst: $\max_{p \in \Delta_k} \mathop{\mathbb{E}}_{X^n \sim p} L(p, q_{X^n})$

Min-max Loss (Risk)

$$r^L_{k,n} \stackrel{\text{def}}{=} \min_q \max_{p \in \Delta_k} \mathop{\mathbb{E}}_{X^n \sim p} L(p, q_{X^n})$$

- Objective: Find $r^L_{k,n}$ and q for important losses L

Keep it simple

- Simplest interesting distance
- ℓ_2, Euclidean distance: $\|p-q\|_2 = \sqrt{\sum_i (p_i - q_i)^2}$
- Simple? $\sqrt{\ }$ complex $\sqrt{-1}$
- Even for positive, complicated $\sqrt{2}$
- ℓ_2^2, mean square error:

$$\|p-q\|_2^2 = \sum_i (p_i - q_i)^2$$

- Engineering staple

ℓ_2^2 - *Empirical Estimator*

- $r_{k,n} = \min\limits_{q} \max\limits_{p \in \Delta_k} \mathop{\mathbb{E}}\limits_{X^n} \left\| p - q_{X^n} \right\|_2^2 = \min\limits_{q} \max\limits_{p \in \Delta_k} \mathop{\mathbb{E}}\limits_{X^n} \sum\limits_i (p_i - q_{X^n,i})^2$
- Empirical estimator
  - $q^{\text{emp}}_{X^n,i} = N_i/n$
  - T_i — # times i appears in X^n
  - $T_i \sim \text{Binom}(p_i, n)$
  - $\mathbb{E}T_i = np_i$
  - $\text{Var}(T_i) = \mathbb{E}(T_i - np_i)^2 = np_i(1-p_i)$

$$\mathbb{E} \sum_{i=1}^{k} \left(\frac{T_i}{n} - p_i \right)^2 = \mathbb{E} \sum_{i=1}^{k} \left(\frac{T_i - np_i}{n} \right)^2 = \sum_{i=1}^{k} \frac{\text{Var}(T_i)}{n^2}$$

$$= \sum_{i=1}^{k} \frac{p_i(1-p_i)}{n} = \frac{1 - \sum_{i=1}^{k} p_i^2}{n} \le \frac{1 - \frac{1}{k}}{n}$$

- Uniformly bounded over all distributions
- Optimal?

ℓ_2^2 – General

- $r_{k,n} = \min_q \max_{p \in \Delta_k} \mathop{\mathbb{E}}_{X^n} \|p - q_{X^n}\|_2^2 = \min_q \max_{p \in \Delta_k} \mathop{\mathbb{E}}_{X^n} \sum_i (p_i - q_{X^n,i})^2$
- Empirical frequency: $r_{k,n}^{q^{\text{emp}}} \leq \dfrac{1 - \frac{1}{k}}{n}$
- Add constant: $q_{X^n,i}^{+\sqrt{n}/k} \stackrel{\text{def}}{=} \dfrac{N_i + \sqrt{n}/k}{n + \sqrt{n}}$
- $r_{k,n}^{q^{+\sqrt{n}/k}} = \dfrac{1 - \frac{1}{k}}{(\sqrt{n}+1)^2} = \dfrac{1 - \frac{1}{k}}{n + 2\sqrt{n} + 1}$
- Matching lower bound: Dirichlet prior with parameter $\dfrac{\sqrt{n}}{k}$

Theorem

$$\forall k \geq 2, n \geq 1 \qquad r_{k,n} = \frac{1 - \frac{1}{k}}{(\sqrt{n}+1)^2}$$

Lower Bound for ℓ_2^2

- For any prior distribution π over Δ_k,

$$r_{k,n} = \min_q \max_{p \in \Delta_k} \mathbb{E}[||p - q||_2^2] \geq \min_q \mathbb{E}_{P \sim \pi} \mathbb{E}[||p - q||_2^2]$$

- Best estimator upon observing X^n

$$q_i^*(x^n) = \mathbb{E}[P_i | X^n = x^n]$$

- Use this q^* to obtain a lower bound
- Which prior?
- Suitable Dirichlet prior yields optimal bounds

- For simplicity, keep denoting $r_{k,n}$
- $r_{k,n} = \min_q \max_{p \in \Delta_k} \mathop{\mathbb{E}}_{X^n \sim p} \sqrt{\sum_i (p_i - q_{X^n,i})^2}$
- $\sqrt{\ }$ concave
- $\mathop{\mathbb{E}}_{X^n \sim p} \sqrt{\sum_i (p_i - q_{X^n,i})^2} \leq \sqrt{\mathop{\mathbb{E}}_{X^n \sim p} \sum_i (p_i - q_{X^n,i})^2}$
- $r_{k,n} \leq \dfrac{\sqrt{1 - \frac{1}{k}}}{\sqrt{n} + 1}$

The Trouble(s) with ℓ_2

- No probabilistic interpretation
- Poor for large alphabets k
  - $p = \Big(\underbrace{\frac{2}{k}, \ldots, \frac{2}{k}}_{k/2}, \underbrace{0, \ldots, 0}_{k/2}\Big), \quad q = \Big(\underbrace{0, \ldots, 0}_{k/2}, \underbrace{\frac{2}{k}, \ldots, \frac{2}{k}}_{k/2}\Big)$
  - Very different ... more than Spanish and Catalan
  - $\ell_2^2(p, q) = k \cdot \frac{4}{k^2} = \frac{4}{k} \to 0$
  - $\ell_2(p, q) = \frac{2}{\sqrt{k}} \to 0$
  - ℓ_2^2 and ℓ_2 say p, q essentially same
- Need distances that distinguish large distributions
  - $\ell_1(p, q) = 2$
  - $\chi^2(p||q) = D(p||q) = \infty$

ℓ_1 – *Definition and Motivation*

- $\|p - q\|_1 \stackrel{\text{def}}{=} \sum_{i=1}^{k} |p_i - q_i|$
- Probabilistic interpretations
- Errors in hypothesis testing, classification

Theorem (folklore)

To estimate p to within an ℓ_1 distance of ϵ, we need to have a number of samples $n = \Theta\left(\frac{k-1}{\epsilon^2}\right)$.

ℓ_1 Results

$$r_{k,n} = \min_q \max_{p \in \Delta_k} \mathbb{E} \, ||p - q_{X^n}||_1 = \min_q \max_{p \in \Delta_k} \mathbb{E} \sum_i |p_i - q_i|$$

Theorem (Kamath-O.-Pichapati-Suresh '15)

For fixed k, as $n \to \infty$, $\quad r_{k,n} = \sqrt{\frac{2(k-1)}{\pi n}} + O\left(\frac{1}{n^{3/4}}\right)$

Achieved by empirical estimator

Makes folklore theorem precise:

$$n \approx \frac{2(k-1)}{\pi \epsilon^2}$$

KL Divergence

$$D(p||q) = \sum_i p_i \log \frac{p_i}{q_i}$$

- Uses:
  - Additional # bits in compression
  - Prediction with log-loss
- For q^{emp}
  - If $T_i = 0$, $q_i^{\text{emp}} = 0 \implies \mathbb{E}[D(p||q)] = \infty$

KL Divergence Results

- As $n/k \to \infty$
  - Best add-constant estimator:: $q_i^{+0.509} \propto T_i + 0.509$

$$\max_{p \in \Delta_k} \mathbb{E}[D(p||q^{+0.509})] \sim 0.509 \cdot \frac{k-1}{n}$$

  - Braess Sauer '04: modification of add-constant estimator
    - $T_0 + 1/2$
    - $T_1 + 1$
    - $T_i + 3/4$ for $i > 1$

$$r^{\mathsf{KL}}_{k,n} = \frac{k-1}{2n}$$

- As $k/n \to \infty$
  - Paninski '04

$$r^{\mathsf{KL}}_{k,n} \sim \log \frac{k}{n}$$

$\Theta(k)$ samples are necessary to learn in KL

Open Problems

- What about $k \propto n$?
- Bounds by Paninski, us, Jiao et al.
- $$\lim_{k \to \infty} r_{k,ck}^{\ell_1} = ?$$
- $$\lim_{k \to \infty} r_{k,ck}^{KL} = ?$$

Competitive distribution estimation

Min-max Recap

- $\Delta_k \stackrel{\text{def}}{=}$ set of all distributions over $\{1, 2, \ldots, k\}$

$$q^* = \arg\min_q \max_{p \in \Delta_k} L(p, q)$$

- $T_x \stackrel{\text{def}}{=}$ # times x appears
- Empirical

$$q^{\mathsf{E}}(x) \propto T_x$$

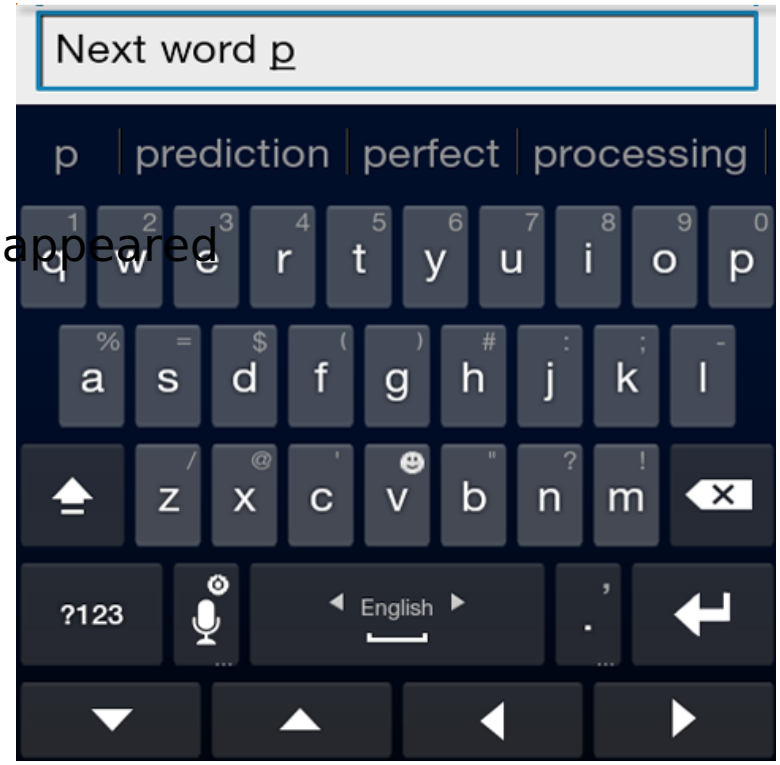
- Not optimal
- Near-optimal: Krichevsky-Trofimov

$$q^{\mathsf{KT}}(x) \propto T_x + \frac{1}{2}$$

- Information theory favorite

Large Domains

- Applications
  - Texting
    - Domain size: 10^6 words
    - Sample size: # times context appeared
  - Speech recognition
  - Natural language processing
  - Ecology
  - Immunology
  - Genomics
  - Biomics
  - ...
- sample size $\gg$ domain size
- Do old techniques still work?

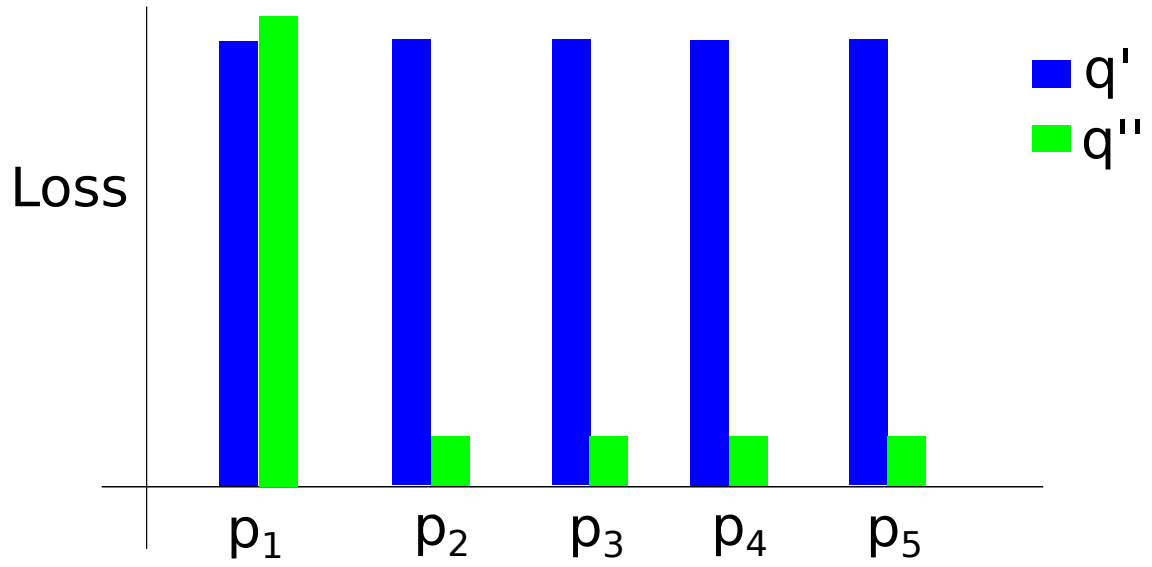


$+$ *or* $-$*?*

- Krichevsky-Trofimov: $q_i^{\mathsf{KT}} \propto t_i + 1/2 \quad \approx q_{k,n}$
- $r_{k,n} \sim \dfrac{k-1}{2n}$ for large n
- Sad min-max: Increases with k, as expected
- In practice:
- Absolute discounting: $q_i^{\mathsf{AD}} \propto t_i - \gamma \quad t_i \geq 1 \quad \gamma \approx 0.7$
- Bad min-max: Opposite of optimal KT

Worse Problem with Min-Max

- $q_{mM} = \arg\min_q \max_p L(p, q)$



- Worst loss of q' $<$ worst loss of q''
- Min-max estimator: q'
- Which would you choose? What if p_1 never happens?
- Exactly the case for applications!
- Loss of q_{mM} highest for uniform distributions, doesn't happen!

Worst Problem with Min-Max

- Min-max: best estimator for worst distribution
- Imagine: best vehicle for worst terrain




Goal: single vehicle, near-optimal for every terrain

Competitive Distribution Estimation

- Estimate each p as well as :
  - the best Human designed estimators

  - Super-human estimator designed with extra information

- As well as super-human $\Longrightarrow$ as well as any human
- Can compete with any super-human

Vehicles to Distributions

- Divide Δ_k to groups $\mathscr{P} = P_1, P_2, \ldots,$

$$r_n^{\mathscr{P}} = \min_q \max_i \left(\max_{p \in P_i} \mathsf{KL}(p, q) - \min_{q'} \max_{p' \in P_i} \mathsf{KL}(p', q') \right)$$

- Min-max formulation
  - Each group $=$ individual distribution
  - Compete with a genie that knows the distribution
  - Too pessimistic

- Competitive learning: genie that knows the group P_i
  - Possible genies: knows sparsity, entropy
  - Genie knows the probability multiset (r_n^σ)
  - $p(H) = 0.6, p(T) = 0.4$
    $\rightarrow \{p(H), p(T)\} = \{0.4, 0.6\}$

- Local min-max, adaptive, instance-by-instance

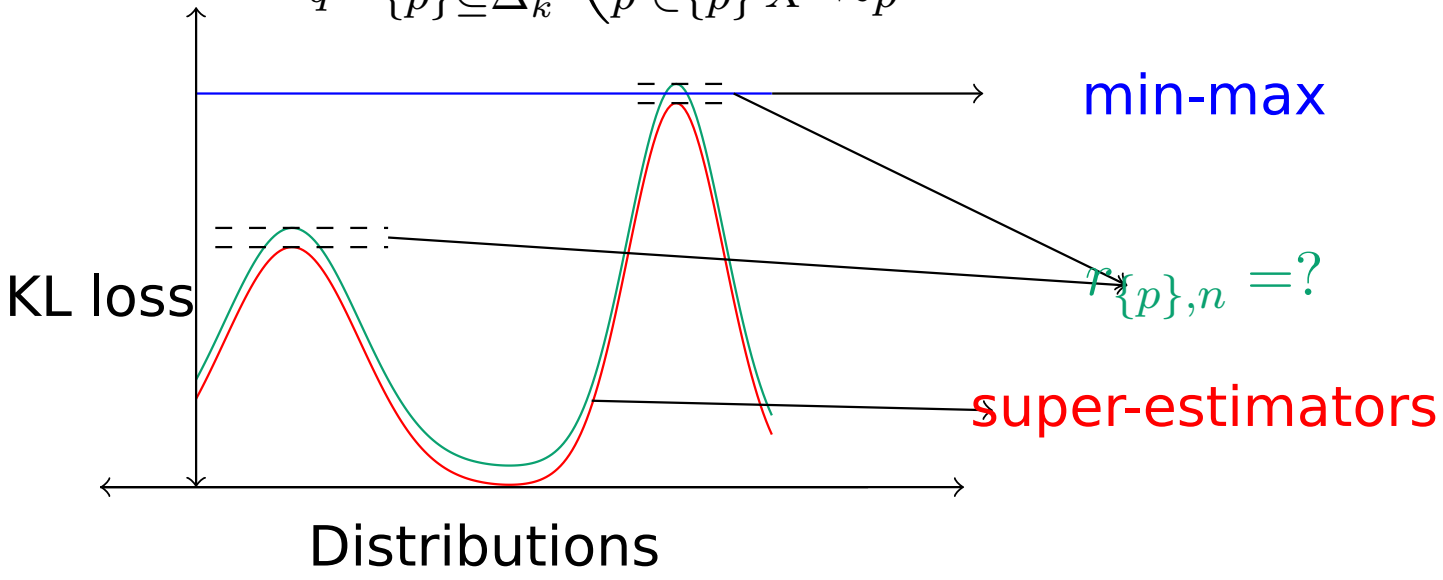
Multiset Genie Estimator

- Powerful genie estimator knows the probability multiset
  - $p_1 = 0.4,\ p_2 = 0.2,\ p_3 = 0.4 \implies \{p\} = \{0.4, 0.4, 0.2\}$
- Loss of optimal genie estimator

$$r_{\{p\},n} \stackrel{\text{def}}{=} \min_{q} \max_{p' \in \{p\}} \mathop{\mathbb{E}}_{X^n \sim p'} D(p' || q(X^n))$$

- Competitive regret

$$r_{\{\},n} \stackrel{\text{def}}{=} \min_{q} \max_{\{p\} \subseteq \Delta_k} \left(\max_{p' \in \{p\}} \mathop{\mathbb{E}}_{X^n \sim p'} D(p' || q(X^n)) - r_{\{p\},n} \right)$$



Good-Turing Estimator

- $\Phi(t) \stackrel{\text{def}}{=}$ # of symbols that appeared t times
  - $X^n = a\,b\,a\,b\,c\,f\,e\,e\,e \implies \Phi(1) = 2$
- Recall: $T_i =$ # of times symbol i appeared

$$q_i = \frac{T_i + 1}{n} \cdot \frac{\Phi(T_i + 1)}{\Phi(T_i)}$$

  - a appeared 2 times, $T_a = 2$

$$q_a = \frac{2+1}{9} \cdot \frac{\Phi(3)}{\Phi(2)} = \frac{2+1}{9} \cdot \frac{1}{2} = \frac{1}{6}$$

- Experiments + intuition: Good'53, Church and Gale'81
- Theory for total mass: McAllester Schapire'00, Drukh Mansour'04
- What about estimating distribution itself?
- Why is it better than optimal min-max?

Results

- Good-Turing + empirical:

$$r_n^\sigma \leq \frac{3 + o_n(1)}{n^{1/3}}$$

  - $o_n(1) \leq 3$ for moderate n
- Improved estimator:

$$r_n^\sigma \leq \mathscr{O}_n\left(\min\left(\frac{1}{n^{1/2}}, \frac{k}{n}\right)\right)$$

- Independent of the support size k
- Information-theoretic lower bound

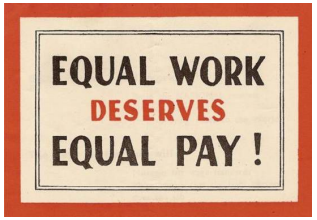
$$r_n^\sigma \geq \tilde{\Omega}_n\left(\min\left(\frac{1}{n^{2/3}}, \frac{k}{n}\right)\right)$$

What does Genie do?

$$\min_{q'} \max_{p' \in P_i} \mathsf{KL}(p', q')$$

- P_i contains all distributions with the same multiset
- Symmetry $\implies$?

$$x^5 = a\,b\,b\,a\,c \implies q'_a = q'_b$$



- $\mathscr{Q}^{\text{nat}}$: set of all such natural estimators

Knows multiset

Knows distribution, but is natural

Knows distribution

Distribution estimation → total probability estimation

- Genie knows the distribution and is natural: r_n^{nat}
- $x^5 = a\,b\,b\,a\,c, e \implies q_a = q_b = \dfrac{p_a + p_b}{2} = \dfrac{M(2)}{2}$
- $M(t)$: sum of probabilities of symbols appearing t times
- Properties of KL divergence:

$$q_i^*(X^n) = \frac{M(T_i)}{\Phi(T_i)}$$

- Estimate $\hat{M}$ and $q_i(X^n) = \dfrac{\hat{M}(T_i)}{\Phi(T_i)}$

$$r_n^{\text{nat}} = \min_{\hat{M}} \max_{p} \mathbb{E}\left[\sum_{t=0}^{n} M(t) \log \frac{M(t)}{\hat{M}(t)}\right]$$

Good-Turing estimators

- [OS'15]
  - r_n^{σ}: genie which knows the multiset
  - r_n^{nat}: genie knows distribution, but is natural

$$r_n^{\sigma} \leq r_n^{\mathsf{nat}} = \min_{\hat{M}} \max_{p} \mathbb{E}\left[\sum_{t=0}^{n} M(t) \log \frac{M(t)}{\hat{M}(t)}\right]$$

  - $M(t)$: total probability of symbols appearing t times
- Observation

$$\mathbb{E}[M(t)] = \frac{t+1}{n} \cdot \mathbb{E}[\Phi(t+1)]$$

- Good-Turing estimator:

$$\hat{M}(t) \approx \frac{t+1}{n}\Phi(t+1) \implies q_i = \frac{T_i+1}{n}\frac{\Phi(T_i+1)}{\Phi(T_i)}$$

- [OS'15]: Good-Turing + empirical satisfies

$$r_n^{\mathsf{nat}} = \mathscr{O}\left(\min\left(\frac{1}{n^{1/3}}, \frac{k}{n}\right)\right)$$

Combined probability: $M(t) = \sum_x p_x \mathbb{I}_t(x)$

Good Turing estimate:

$$G_t = \frac{t+1}{n} \cdot \Phi_{t+1} = \frac{t+1}{n} \cdot \sum_x \mathbb{I}_{t+1}(x)$$

$$\begin{aligned}
\mathbb{E}[G_t] &= \sum_x \frac{t+1}{n} \cdot \mathbb{E}[\mathbb{I}_{t+1}(x)] \\
&= \sum_x \frac{t+1}{n} e^{-\lambda_x} \cdot \frac{\lambda_x^{t+1}}{(t+1)!} \qquad (\lambda_x = np_x) \\
&= \sum_x \frac{\lambda_x}{n} \cdot e^{-\lambda_x} \cdot \frac{\lambda_x^t}{t!} \\
&= \sum_x p_x \cdot \mathbb{E}[\mathbb{I}_t(x)] \\
&= \mathbb{E}[M(t)]
\end{aligned}$$

0 bias!

Moderate variance

Simulations: Actual KL Divergence

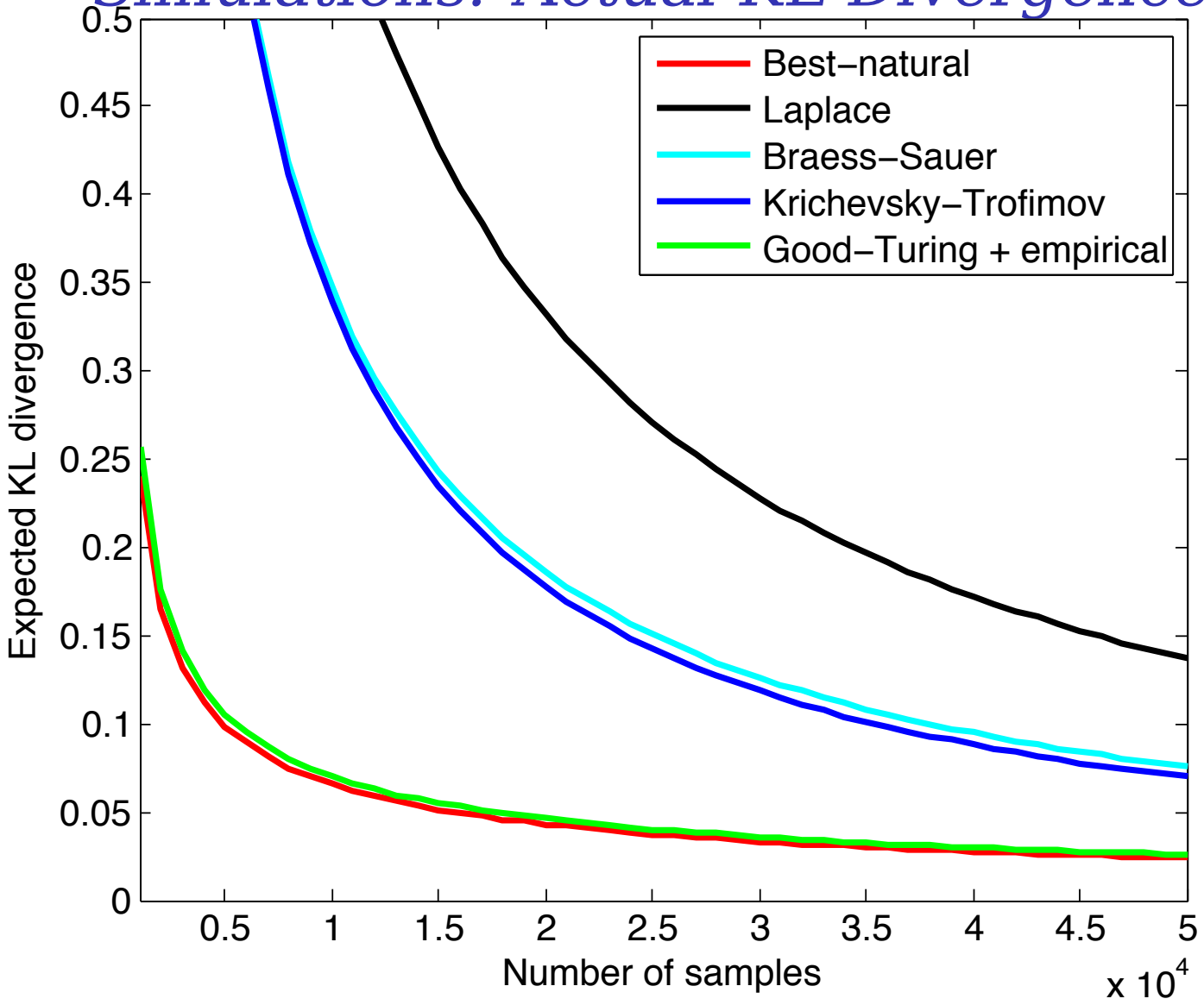
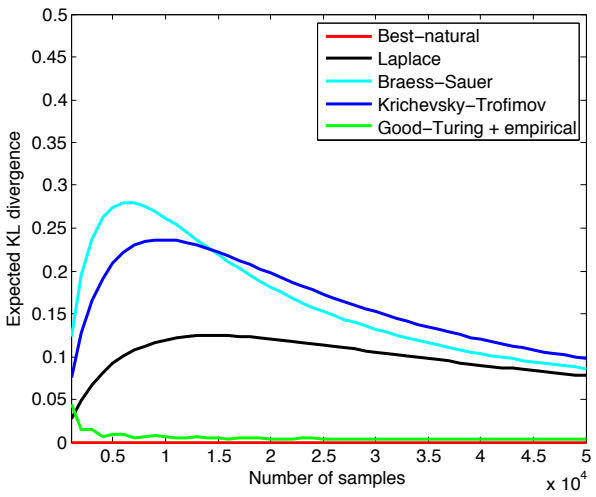
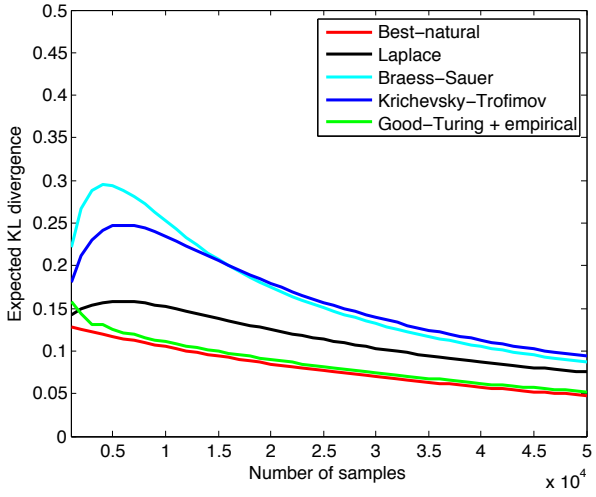


Figure: Zipf-1.5, Support 10000, # of samples from 1000 to 50000

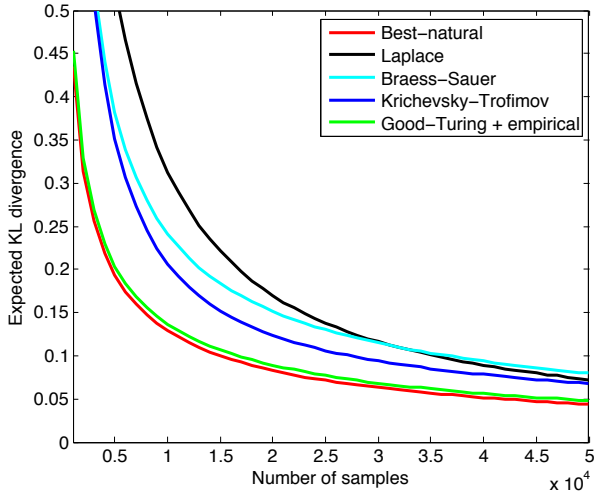
Simulations: actual KL divergence



(a) Uniform



(b) Step



(c) Zipf-1

Figure: Support 10000, # of samples ranging from 1000 to 50000

Optimal Missing Mass Estimation

- Estimating the missing mass is important in biology
  - Sample coverage
- Colwell et al, Chao et al
- How best can one estimate missing mass?

$$\min_{\hat{M}} \max_{p} \mathbb{E}_p \left[(M(0) - \hat{M}(0))^2 \right]$$

- Back of the envelope calculation: c/n.
- What is the right constant c?

Optimal Competitive Estimation

- What is the optimal rate of competitive estimation?

$$\tilde{\Omega}\left(\frac{1}{n^{2/3}}\right) \leq r_n^{\sigma} \leq \tilde{\mathscr{O}}\left(\frac{1}{n^{1/2}}\right)$$

- Other distance measures?
  - ℓ_2^2 distance: GT is near-optimal.
  - ℓ_1 distance: Valiant Valiant '15
  - Hellinger, Chi-squared?

Part 2: Property estimation

Property estimation

Estimate a functional of a distribution(s)

- Entropy
- Rényi entropy
- Support size
- Diversity indices
- Mutual information
- Distance measures: e.g., Jiao et al, Monday 5:10 PM

Estimate the distribution, output property value of estimated distribution

Why learn a k probabilities to deduce one number?

Is it easier than estimating the entire distribution?

Entropy estimation

Applications of entropy estimation

Amount of randomness in fingerprints

Amount of information is transmitted in optical nerves

Other applications:

- Genetic diversity Shenkin et al '91
- Quantifying neural activity Paninski'03, Nemenman et al '04
- Network anomaly detection Lall et al '06

Machine learning: better entropy estimators $\Longrightarrow$

- Better decision trees Nowozin et al '12
- Better graphical model estimation Jiao et al '14

Problem statement

p : discrete distribution over $\mathscr{X}$ with $\leq k$ elements

Given:

- Independent samples $X_1, X_2, \ldots, X_n$ from p
- k is unknown

Quantity of interest

$$f(p) = \sum_{x \in \mathscr{X}} f(p_x)$$

Design $\hat{f} : \mathscr{X}^n \to \mathbb{R}$,

$$\min_{\hat{f}} \max_{p \in \Delta_k} \mathbb{E}\left[(f(p) - \hat{f}(X^n))^2 \right]$$

Empirical:

$$f^e(X^n) = \sum_{x \in \mathscr{X}} f\left(\frac{T_x}{n}\right)$$

Better than empirical?

(Rényi) Entropy

Rényi entropy of order $\alpha > 0$,

$$H_\alpha(p) \stackrel{\text{def}}{=} \frac{1}{1-\alpha} \log \left(\sum_{x \in \mathscr{X}} p_x^\alpha \right)$$

Properties

- $H_\alpha(p) \geq 0$
- Continuous and non-increasing in α
- $\lim_{\alpha \to 1} H_\alpha(p) = H(p)$

Applications:

- Ecology (Simpson's index)
- Graph expansion
- Quality of random number generators

Empirical estimators for $H_2(p)$

$$H_2(p) = -\log\left(\sum_x p_x^2\right)$$

$u \stackrel{\text{def}}{=}$ uniform distribution over k elements

$$\sum_x u_x^2 = k \cdot 1/k^2 = 1/k \implies H_2(u) = \log k$$

Empirical estimator: since $\sum_x T_x^2 \geq \sum_x T_x = n$

$$H_2^e(X^n) = -\log \sum_x \left(\frac{T_x}{n}\right)^2 \leq \log n$$

For $n < k$,

$$\mathbb{E}[(H_2(p) - H_2^e(X^n))^2] \geq \log^2(k/n)$$

$n = \Omega(k)$ are necessary, high bias

Entropy to moments

$$M_2(p) \stackrel{\text{def}}{=} \sum_x p_x^2 \implies H_2(p) = -\log M_2(p)$$

$$\hat{H}_2(X^n) = -\log \hat{M}_2(X^n)$$

Since $\log(1+z) \leq z$

$$\mathbb{E}\left[(H_2(p) - \hat{H}_2(X^n))^2\right] = \mathbb{E}\left[\log^2 \frac{\hat{M}_2(X^n)}{M_2(p)}\right]$$
$$\leq \frac{\mathbb{E}[M_2(p) - \hat{M}_2(X^n))^2]}{M_2^2(p)}$$

$+$ve error in $H_2(p)$ is same as $\times$ve error in $M_2(p)$

Unbiased estimators for $M_2(p)$

$$\hat{M}_2^{\mathsf{un}}(X^n) = \frac{1}{n(n-1)} \cdot \sum_{x \in \mathscr{X}} T_x(T_x - 1)$$

Each $T_x \sim \mathrm{Bin}(\mathrm{n}, \mathrm{p_x})$, moments of Binomial distribution

$$\mathbb{E}[\hat{M}_2^{\mathsf{un}}(X^n)] = \frac{1}{n(n-1)} \cdot \mathbb{E}[T_x(T_x - 1)] = \sum_x p_x^2 = M_2(p)$$

Since $T_x(T_x - 1)$'s are pairwise negatively correlated

$$\begin{aligned} \mathsf{Var}(\hat{M}_2^{\mathsf{un}}(X^n)) &\leq \sum_x \frac{1}{n^2(n-1)^2} \cdot \mathsf{Var}(T_x(T_x - 1)) \\ &\leq \frac{4M_3(p)}{n} + \frac{2M_2(p)}{n(n-1)} \end{aligned}$$

MSE of proposed estimators for $H_2(p)$

Previous observations + convexity + Holder's inequality:

$$\begin{aligned}
\mathbb{E}[(H_2(p) - \hat{H}_2(X^n))^2] &\leq \frac{\mathbb{E}[(M_2(p) - \hat{M}_2^{\mathsf{un}}(X^n))^2]}{M_2^2(p)} \\
&\leq \frac{4M_3(p)}{nM_2^2(p)} + \frac{2M_2(p)}{n(n-1)M_2^2(p)} \\
&\leq \frac{4\sqrt{k}}{n} + \frac{2k}{n(n-1)}
\end{aligned}$$

Integral order Rényi entropies

Theorem

For any $p \in \Delta_k$ *and* $n < k$,

$$\mathbb{E}[(H_2(p) - \hat{H}_2(X^n))^2] \lesssim \frac{\sqrt{k}}{n}$$

For any $p \in \Delta_k$, For integral $\alpha > 1$, similar analysis yields:

Theorem

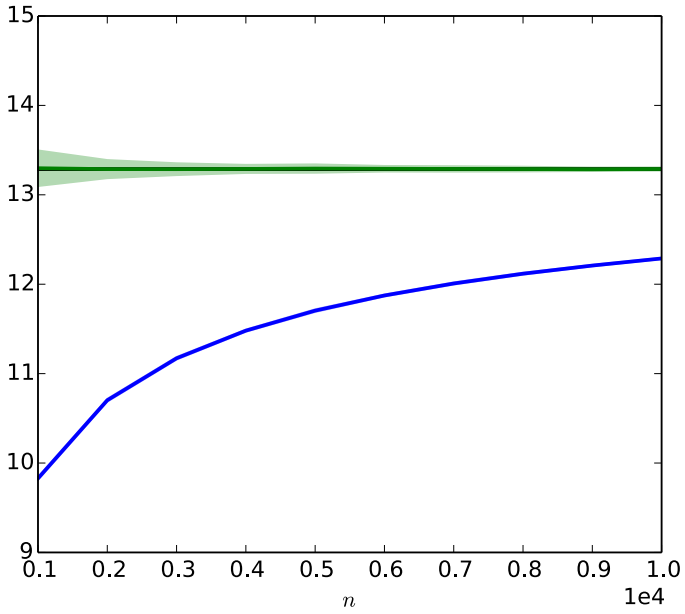
For $\alpha \in \mathbb{N} \setminus \{1\}$ *and* $n < k$,

$$\mathbb{E}[(H_\alpha(p) - \hat{H}_\alpha(X^n))^2] \lesssim \frac{k^{1-1/\alpha}}{n}$$

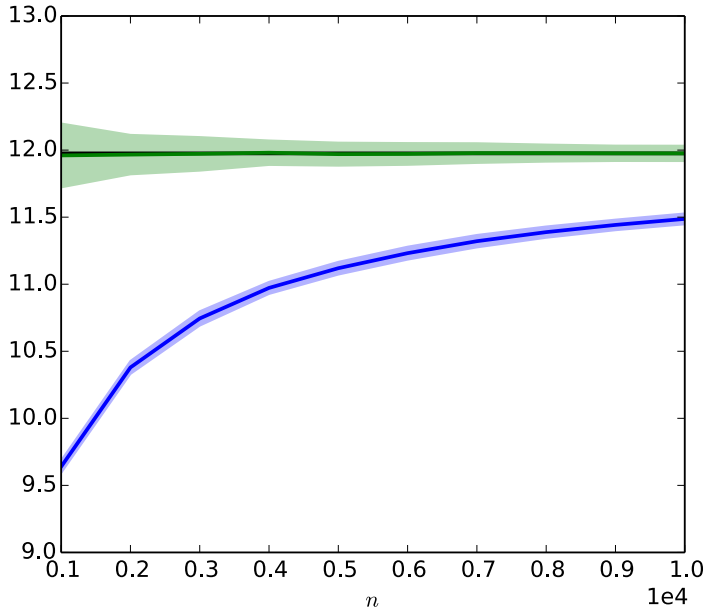
$\mathscr{O}(k^{1-1/\alpha})$ samples are sufficient to estimate $H_\alpha(p)$

Sublinear in k, much easier than distribution estimation

Experiments



Uniform



Zipf

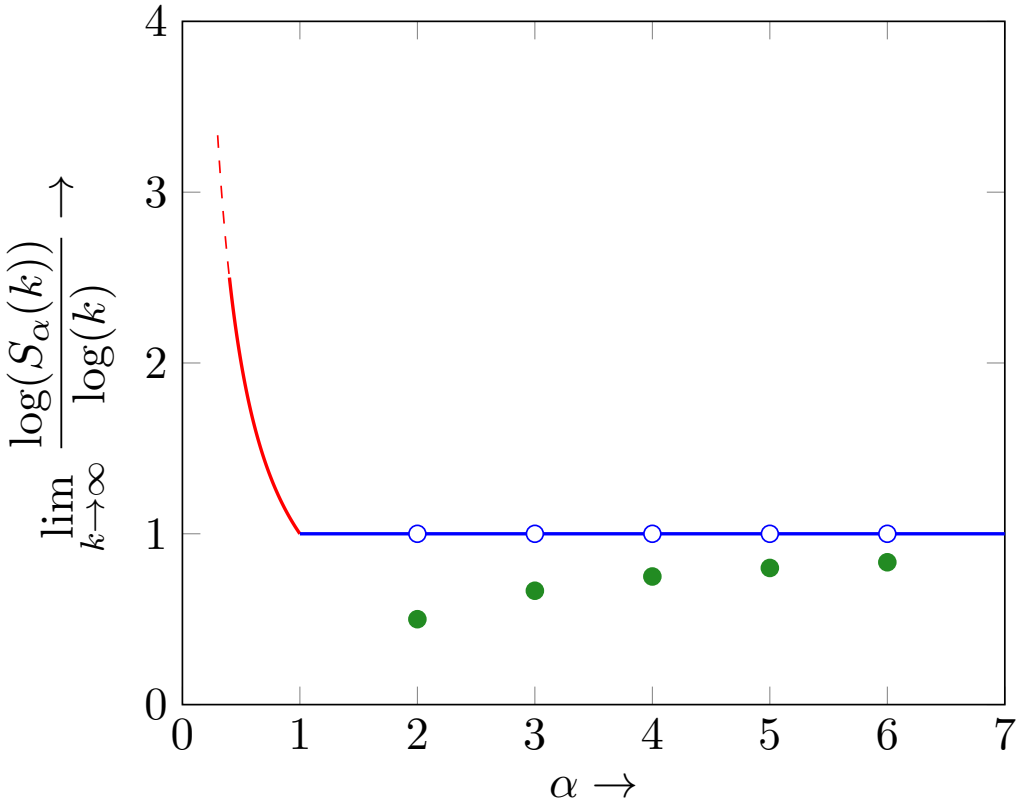
Estimate as a function of # of samples

$k = 10000$, unbiased, empirical

Non-integer values of α

$S_\alpha(k)$: complexity of estimating $H_\alpha(p)$ to ± 0.1

How do number of samples scale with k?



Integer α poly-sublinear Non-integer α nearly-linear

Non-integral Rényi entropy

- Sublinear estimators for non-integer values of α?
- Sublinear estimators for Shannon entropy $\alpha \to 1$?

$$H(p) \stackrel{def}{=} \sum_{x \in \mathscr{X}} p_x \log \frac{1}{p_x}$$

- Empirical estimator: suffers from high bias

$$H^e(p) \stackrel{def}{=} \sum_{x \in \mathscr{X}} \frac{T_x}{n} \log \frac{n}{T_x}$$

- Unbiased estimator: suffers from high variance

Tools from approximation theory

Rényi entropy for non-integer α, Shannon entropy, support

So far: unbiased estimators for $M_j(p) = \sum_x p_x^j$ for integers j

Approximate $f(p_x)$ by a series

$$f(y) \approx \tilde{f}(y) = \sum_j \tilde{f}_j y^j$$

$$\begin{aligned}
\sum_x f(p_x) &\approx \sum_x \tilde{f}(p_x) \\
&= \sum_x \sum_j \tilde{f}_j p_x^j \\
&= \sum_j \tilde{f}_j \sum_x p_x^j \\
&= \sum_j \tilde{f}_j M_j(p)
\end{aligned}$$

Estimators for general functions

- Unbiased estimator for $M_j(p)$ is $\hat{M}_j^{\mathsf{un}}(X^n)$
- Use unbiased estimator for $\tilde{f}$ instead of f

$$\hat{\tilde{f}} = \sum_j \tilde{f}_j \hat{M}_j^{\mathsf{un}}(X^n)$$

- Bias: error due to approximation $f - \tilde{f}$
- Variance: statistical error in estimation
- Usually:
  - Better approximation = large values of $\tilde{f}_j$
  - Large values of $\tilde{f}_j \implies$ large variance
- Correct bias-variance tradeoff $\implies$ good estimators

Results for Shannon entropy

$$H(p) = \sum_{x \in \mathscr{X}} p_x \log \frac{1}{p_x}$$

WY '14, JHVW '15:

$$\mathbb{E}[(H(p) - \hat{H}(X^n))^2] \leq \left(\frac{k}{n \log k}\right)^2 + \frac{\log^2 k}{n}$$

- # samples [Pan'04, VV'11, WY'14, JHVW'15]: # of samples for ε +ve accuracy w.h.p.:

$$\mathscr{O}\left(k \cdot \frac{1}{\varepsilon}\right) \to \Theta\left(\frac{k}{\log k} \cdot \frac{1}{\varepsilon}\right)$$

empirical $\to$ poly. approx

Open Problem : $\times$ve approximation

- $+$ve approximation of entropy: $\log k$ gain
- $\times$ve approximation: possibly much higher gains
- Batu et al '02: to get $1+\epsilon$ $\times$ve approximation w.h.p.:
  - $k^{1/(1+\epsilon)^2}/\epsilon^2 \log k$ samples are sufficient
  - $k^{1/2(1+\epsilon)^2}$ samples are necessary
  - polynomial improvement
- What is the normalized MSE for entropy?

$$\min_{\hat{H}} \max_{p \in \Delta_k : H(p) > h_0} \frac{\mathbb{E}[(H(p) - \hat{H}(X^n))^2]}{H^2(p)}$$

Unseen estimation

Applications of estimating the unseen

THE RELATION BETWEEN THE NUMBER OF SPECIES AND THE NUMBER OF INDIVIDUALS IN A RANDOM SAMPLE OF AN ANIMAL POPULATION

By R. A. FISHER (*Galton Laboratory*), A. STEVEN CORBET (*British Museum, Natural History*) and C. B. WILLIAMS (*Rothamsted Experimental Station*)

- Corbett collected butterflies in Malaya for 1 year

Frequency	1	2	3	4	5	6	7	..
Species	118	74	44	24	29	22	20	..

- # of seen species $= 118 + 74 + 44 + 24 + \ldots$

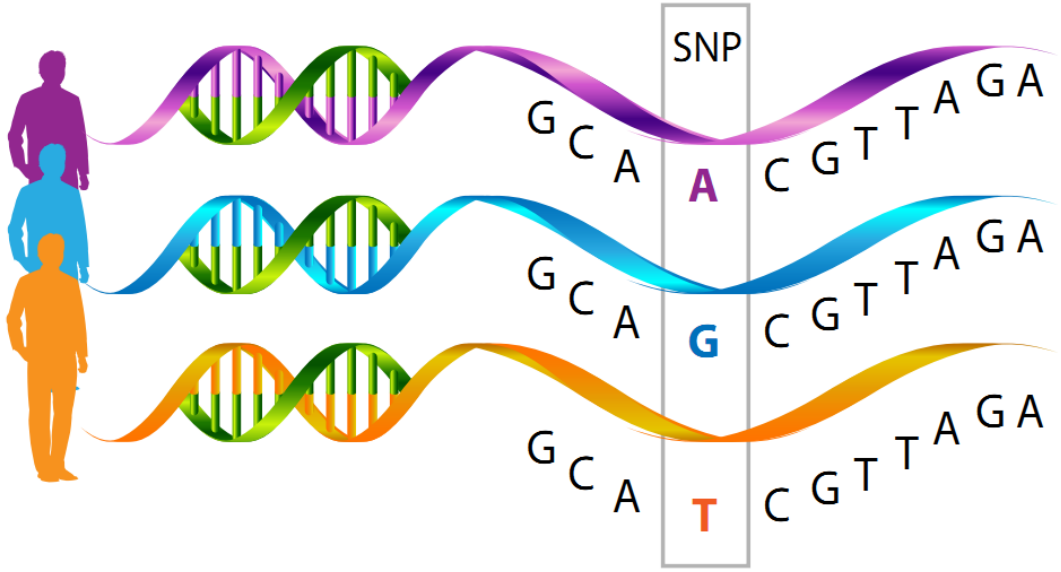
# of new species in the next year?

# of words in a book..

Estimating the number of unseen species: How many words did Shakespeare know?

By BRADLEY EFRON and RONALD THISTED

Applications of estimating the unseen



- Human genome project: collects SNP data
- Laza Lange Laird '09: ≈ 10 people (Chinese, Japanese, European)

Which one has more diversity?

- Use data to extrapolate to $100+$ people

Problem statement

- Distribution p over $\mathscr{X}$, $\sum_x p_x = 1$
  - $p_{\text{shall}} = 0.1$, $p_{\text{thou}} = 0.4, \ldots$.
- Observation: $X_1, X_2, \ldots, X_n \sim p$
- Estimate: $U \stackrel{\text{def}}{=}$ # of unseen symbols in new $n \cdot t$ samples
- $X^n =$ I S I T, $Y^{n \cdot t} =$ S P A I N
  - $n = 4$
  - $t = 5/4$
  - Three new symbols $\implies U = 3$

Higher the t, harder the problem

Function to be approximated

Goal: estimate # of symbols in additional $n \cdot t$ samples

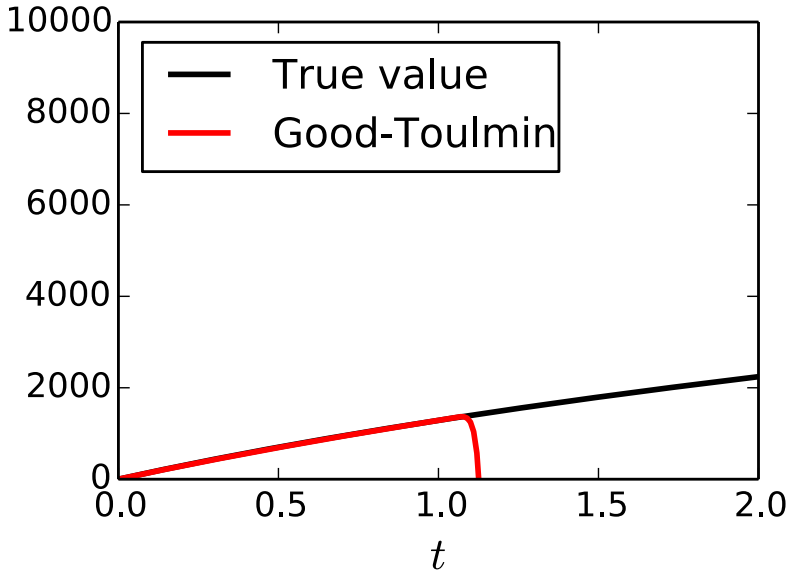
This talk: estimate the expected value

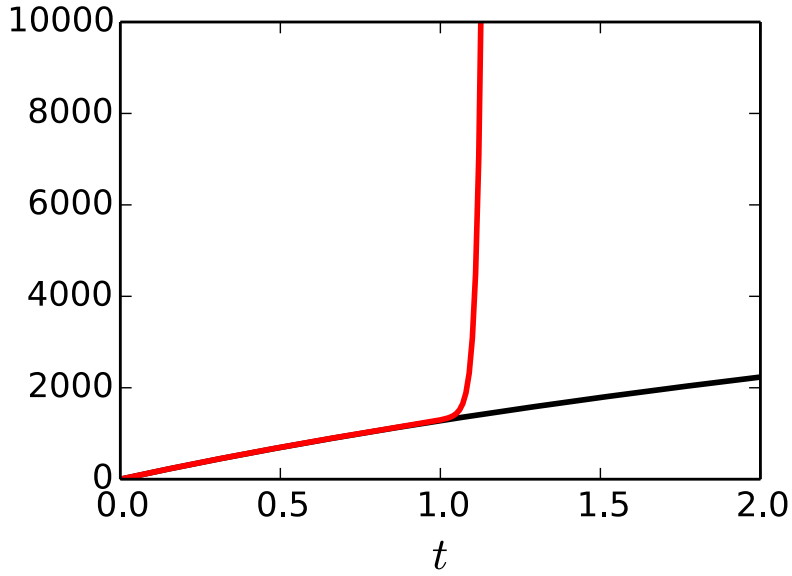
$$\begin{aligned} U &= \mathbb{E}\left[\sum_x \mathbb{I}_{T_x(X^n)=0}\mathbb{I}_{T_x(Y^{nt})>0}\right] \\ &= \sum_x \mathbb{E}\left[\mathbb{I}_{T_x(X^n)=0}\mathbb{I}_{T_x(Y^{nt})>0}\right] \\ &= \sum_x (1-p_x)^n \left(1-(1-p_x)^{nt}\right) \end{aligned}$$

Same as estimating $\sum_x f(p_x)$

Good-Toulmin estimator

Unbiased estimator Good Toulmin '56





Small variance for $t < 1$ Large variance for $t > 1$

Range of prediction $t \leq 1$

Other estimators: Chao, ACE, Jackknife, coverage based

No theoretical guarantees

Results

Orlitsky et al '15

- U^{SGT} : estimator by function approximation
- Observe: $0 \leq U \leq nt$

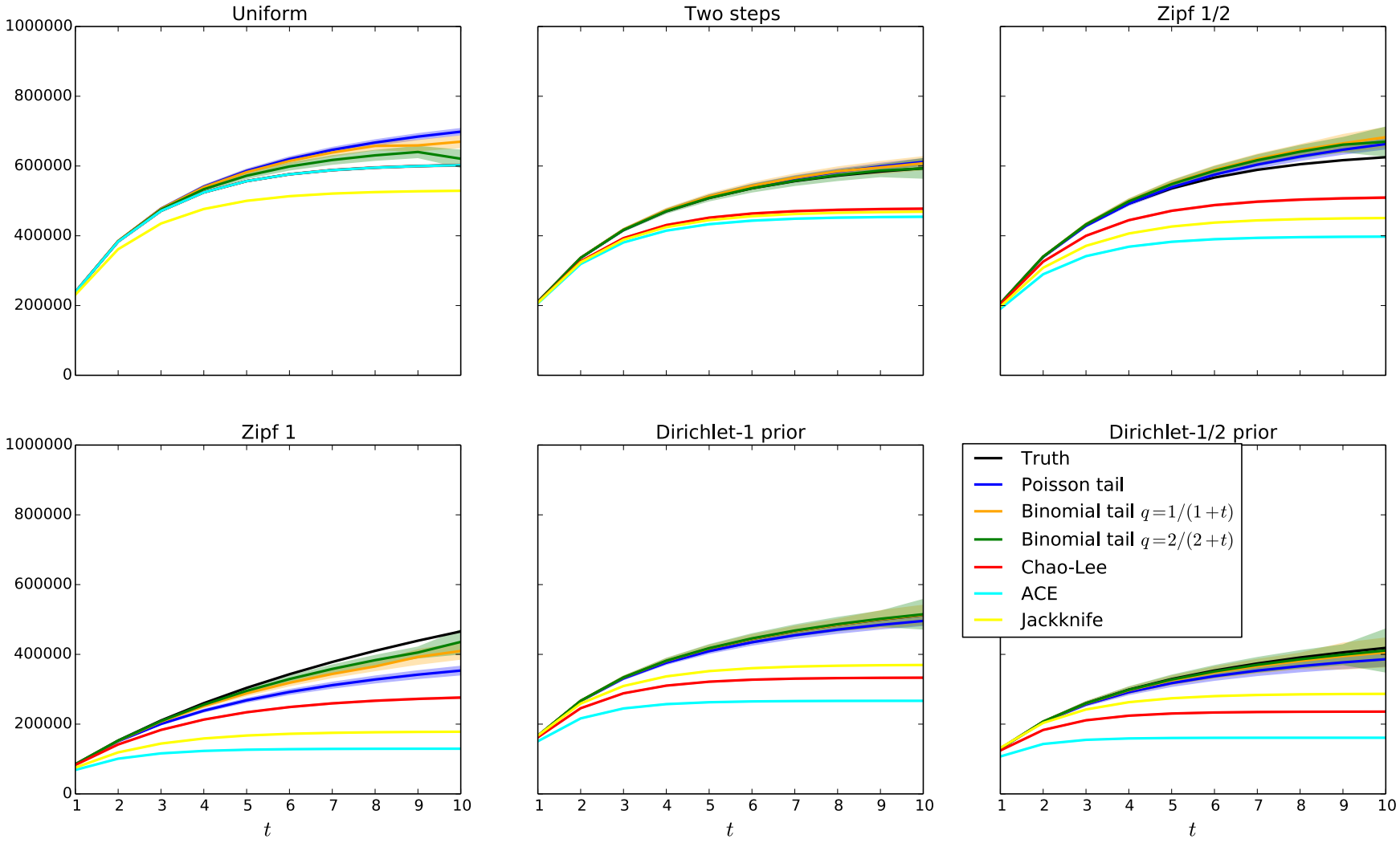
$$\frac{\mathbb{E}[(U - U^{\mathsf{SGT}})^2]}{(nt)^2} \lesssim \frac{1}{(nt)^{1/t}} = \cdot e^{-\frac{\log nt}{t}}$$

  Range of prediction: $t \approx \log n$

Contemporary: Zhou et al '15, Valiant Valiant '15

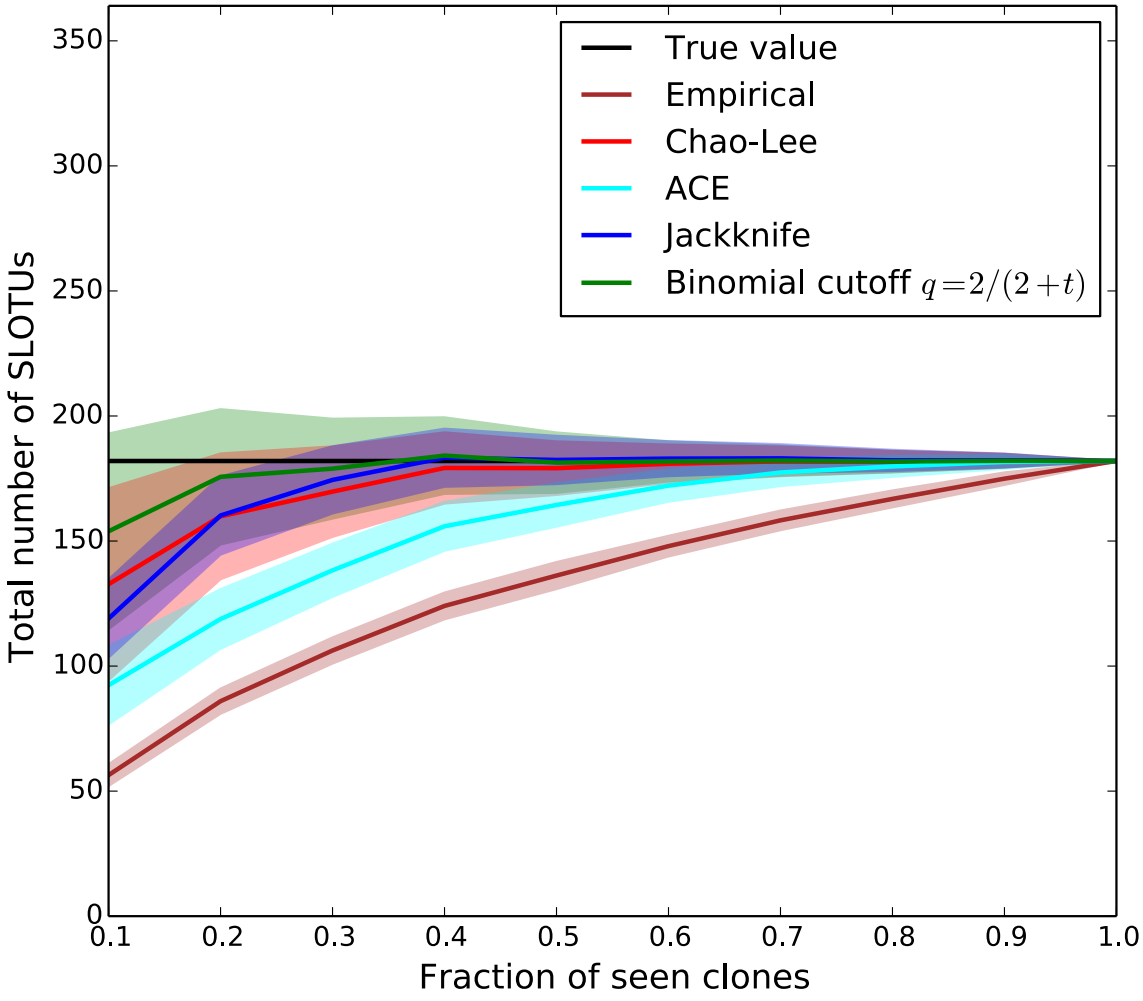
- Linear program
- Exponentially slower

Synthetic species discovery curves



$$k = 10^6,\ n = 5 \cdot 10^5$$

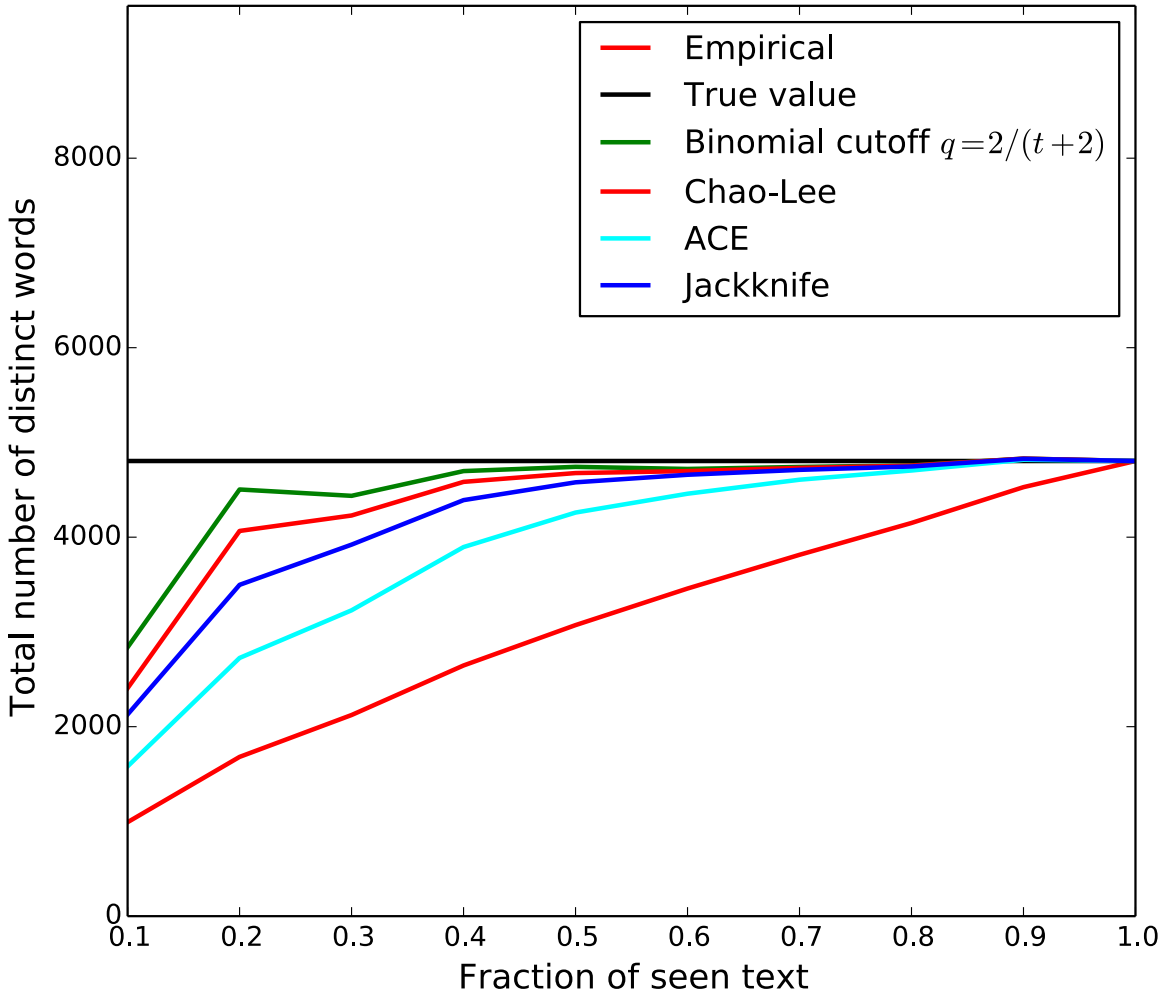
Bacteria on skin *Gao et al '07*



$$k = 182,\ n_{\text{total}} = 1221$$

$$\hat{k} = S_{\text{seen}} + \hat{U}$$

Hamlet



$$k = 4 \cdot 10^3,\ n_{\text{total}} = 3 \cdot 10^4$$

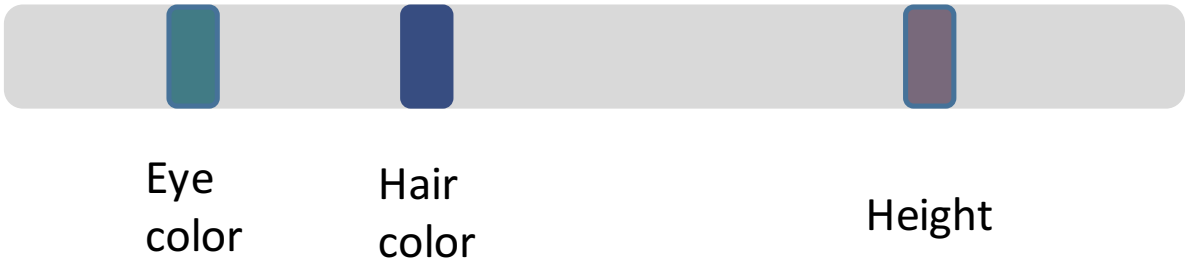
$$\hat{k} = S_{\text{seen}} + \hat{U}$$

Part 3: Property testing (Hypothesis testing)

Lottery

Is lottery a uniform distribution?

Linkage Disequilibrium



Each gene has different variations (alleles) in the population

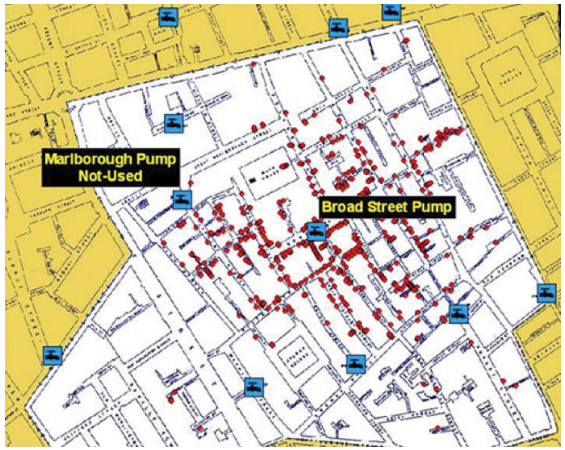
Are the alleles independent across loci?

Is data from the 1000 human genomes project sufficient to test such hypothesis?

Broad Street Cholera

August 31, 1854: Major outbreak of cholera in Soho

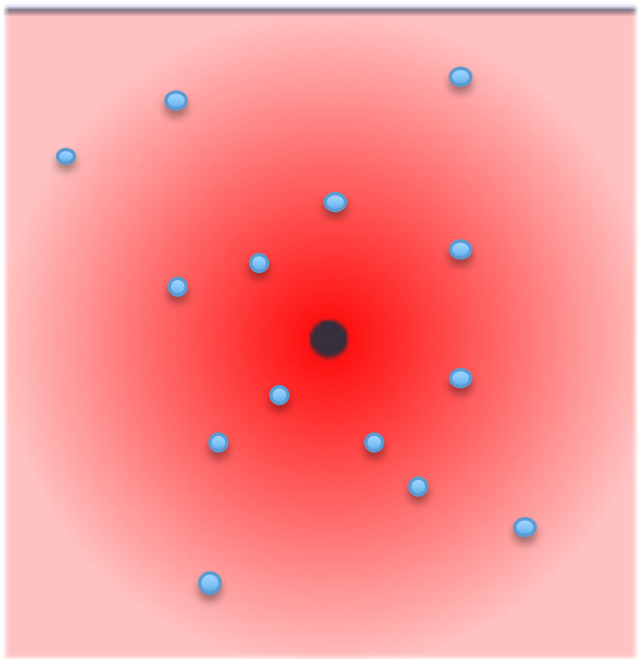
- in 10 days, $\sim$500 people die
- Dominant theory: polluted air
- Physician *John Snow* identifies Broad Street Pump as the source
- Founding event of science of epidemiology



Broad Street Cholera

A hypothesis testing formulation:

- Observe cholera cases in $[0, 1]^2$
- Model/hypothesis:
  - single source of disease
  - probability of infection reduces with distance



Are patients “consistent” with a unimodal distribution?

Problem statement

- $\mathscr{P}, \mathscr{Q}$: families of distributions
  - $\mathscr{P} \cap \mathscr{Q} = \emptyset$
- Two hypotheses:
  - $H_0 : p \in \mathscr{P}$ - Null hypothesis
  - $H_1 : p \in \mathscr{Q}$ - Alternate hypothesis
- Given independent samples from $p \in \mathscr{P} \cup \mathscr{Q}$
- Output $\hat{H} \in \{H_0, H_1\}$
- Error measures

$$\alpha = \Pr(\hat{H} = H_1 | H_0), \text{ and } \beta = \Pr(\hat{H} = H_0 | H_1)$$

False positive and missed detection

Extremely well studied problem

Composite hypothesis testing

- Neyman-Pearson lemma
- Kolmogorov-Smirnov test
- Pearson's Chi-squared test
- Generalized likelihood ratio test, etc

Focus

- Consistency (Error goes to zero as $n \to \infty$)
- Error exponents $(-\frac{1}{n}\log \alpha_n, -\frac{1}{n}\log \beta_n)$

Usually results hold for

$$n \gg k$$

# of draws to test if powerball is uniform $> k = 292$ million

Can we do better?

Testing large domain distributions

Formulation: Batu et al '00

- p: distribution generating independent samples
- $\mathscr{P}$: family of distributions over k
- Given samples, distinguish $p \in \mathscr{P}$ vs $d(p, \mathscr{P}) > \varepsilon$ with probability > 0.9
  - False positives and missed detection < 0.1

How many samples are necessary?

Focus:

- Small fixed error probability
- Computationally efficiency
- Typically # of samples $\ll k$

Hope: test if poweball is uniform with $\ll 292$ million draws

Start small: Fairness of a coin

b : bias of a Bernoulli random variable

Distinguish between

$$H_0 : b = 0.5 \text{ vs } H_1 : |b - 0.5| > \varepsilon$$

Test based on $T_1 \sim Bin(n, b)$

- $H_0 : b = 0.5$, $\mathbb{E}[T_1] = n/2$
- $H_1 : |b - 0.5| > \varepsilon$, $|\mathbb{E}[T_1] - n/2| > n\varepsilon$

$$|T_1 - n/2| \underset{H_1}{\overset{H_0}{\lessgtr}} n\varepsilon/2$$

- Under H_0: Chebychev's Inequality + Var$(T_1) = n/4$

$$\Pr(\hat{H} = H_1 | H_0) = \Pr(|T_1 - \mathbb{E}[T_1]| > \frac{n\varepsilon}{2}) \leq \frac{2}{n\varepsilon^2}$$

- For error to be smaller than 0.1,

$$n \gg \frac{1}{\varepsilon^2}$$

Same question, large k: uniformity

- u : Uniform distribution over k elements
- Given samples from p, decide

$$H_0 : p = u \text{ vs } |p - u|_1 > \varepsilon$$

samples necessary with k and ε?

Uniformity testing

- Only one proof: Estimating $\sum_x p_x^2$
- $H_0 : p = u, \sum_x u_x^2 = k \cdot \frac{1}{k^2} = \frac{1}{k}$
- $H_1 : ||p - u||_1 > \varepsilon$
- By Cauchy-Schwarz

$$k \cdot \left[\sum_x \left(p_x - \frac{1}{k}\right)^2\right] \geq \left(\sum_x \left|p_x - \frac{1}{k}\right|\right)^2 \geq \varepsilon^2$$

$$\implies \quad ||p - u||_2^2 \geq \frac{\varepsilon^2}{k}$$

$$\implies \quad \sum_x p_x^2 \geq \frac{1 + \varepsilon^2}{k}$$

$$H_0 : \sum_x p_x^2 = \frac{1}{k} \text{ vs } H_1 : \sum_x p_x^2 = \frac{1 + \varepsilon^2/2}{k}$$

An L_2 statistic test

$$H_0 : \sum_x p_x^2 = \frac{1}{k} \text{ vs } H_1 : \sum_x p_x^2 = \frac{1+\varepsilon^2/2}{k}$$

Test

$$T \stackrel{\text{def}}{=} \sum_x \frac{T_x(T_x-1)}{n(n-1)} \underset{H_1}{\overset{H_0}{\lessgtr}} \frac{1+\varepsilon^2/2}{k}$$

Analysis under H_0:

$$\mathbb{E}[T] = \sum_x p_x^2$$

$$\mathsf{Var}\,(T) \leq \frac{2M_2(p)}{n(n-1)} + \frac{4M_3(p)}{n} \lesssim \frac{1}{kn^2} + \frac{1}{k^2 n}$$

An L_2 based uniformity test

By Chebyshev's inequality

$$\begin{aligned}\Pr\left(T > \frac{1+\varepsilon^2/2}{k}\right) &= \Pr\left(T - \mathbb{E}[T] > \frac{\varepsilon^2}{2k}\right) \\ &< \frac{4k^2\mathsf{Var}(T)}{\varepsilon^4} \\ &\lesssim \frac{k}{n^2\varepsilon^4}\end{aligned}$$

Similar analysis under H_1

$\frac{\sqrt{k}}{\varepsilon^2}$ samples are sufficient for testing uniformity

Extensions: Testing identity

- Goodness of fit
- q : A known distribution over k elements
- Given samples from p, decide

$$H_0 : p = q \text{ vs } |p - q|_1 > \varepsilon$$

- BFFKRW'01, Paninski'08, VV'14, ADK'15, CDGR'15, DK'16

Theorem

Any distribution can be tested with $\Theta(\sqrt{k}/\varepsilon^2)$ samples.

Open problem: other measures

Sample complexity of testing uniformity (or identity) in f-divergences

- KL divergence: uniformity $\sqrt{k}/\varepsilon^2$, identity $k/\varepsilon^2 \log k$
- Hellinger distance?
- Chi-squared?
- ℓ_p?: results in Waggoner et al '15

Testing properties

So far:

- Test if the distribution is uniform
- Test if the distribution is a given distribution q

Testing Families

- Independence: Is distribution over $[k] \times [k]$ product of marginals?
- Monotonicity: Is the pdf monotone?
- Log-concavity: Is the pdf log-concave?

$\mathscr{P}$: family of distributions over k

$$H_0 : p \in \mathscr{P} \text{ vs } d(p, \mathscr{P}) > \varepsilon$$

A generic approach

Step 1:

- Use samples to learn a $q \in \mathscr{P}$ such that:
  - If $H_0 : p \in \mathscr{P}$, then $||p - q||_1 < \varepsilon/2$
  - If $H_1 : p \notin \mathscr{P}$, then $||p - q||_1 > \varepsilon$

Step 2: Test

$$H_0' :< ||p - q||_1 < \varepsilon/2 \text{ vs } H_1' : ||p - q||_1 > \varepsilon$$

# of samples:

Learning complexity of $\mathscr{P}$ + Testing complexity of a known q

Unfortunately, testing $|p - u|_1 < \varepsilon/2$ vs $|p - u|_1 > \varepsilon$ requires

$$\Omega(k/\log k) \text{ samples!}$$

A generic approach

Step 1:

- Use samples to learn a $q \in \mathscr{P}$ such that:
  - If $H_0 : p \in \mathscr{P}$, then $\chi^2(p, q) < \varepsilon^2/2$
  - If $H_1 : p \notin \mathscr{P}$, then $||p - q||_1 > \varepsilon$

Step 2: Test

$$H_0' : \chi^2(p, q) < \varepsilon^2/2 \text{ vs } H_1' : ||p - q||_1 > \varepsilon$$

# of samples:

Learning complexity of $\mathscr{P}$ + Testing complexity of a known q

- Learning complexity of $\mathscr{P}$: usually small (monotone, log-concave)
- Testing if complexity is known: $\lesssim \sqrt{k}/\varepsilon^2$

Example: Testing independence

Samples from p over $[k] \times [k]$

Test whether its a product distribution or ε away

- Domain size = k^2
- BKR'04, ADK'15, DK'16
- Step 1: learn the marginal distributions
- Step 2: Test if the underlying is close to product of marginals

Theorem (ADK'15, DK16)

Complexity of testing independence = $\Theta(k/\varepsilon^2)$.

Summary: topics covered

- Part 1: Distribution estimation
  - Min-max formulation
  - Competitive formulatation
- Part 2: Property estimation
  - Entropy estimation
  - Estimating the unseen
- Part 3: Property testing
  - Uniformity testing
  - Testing for familites

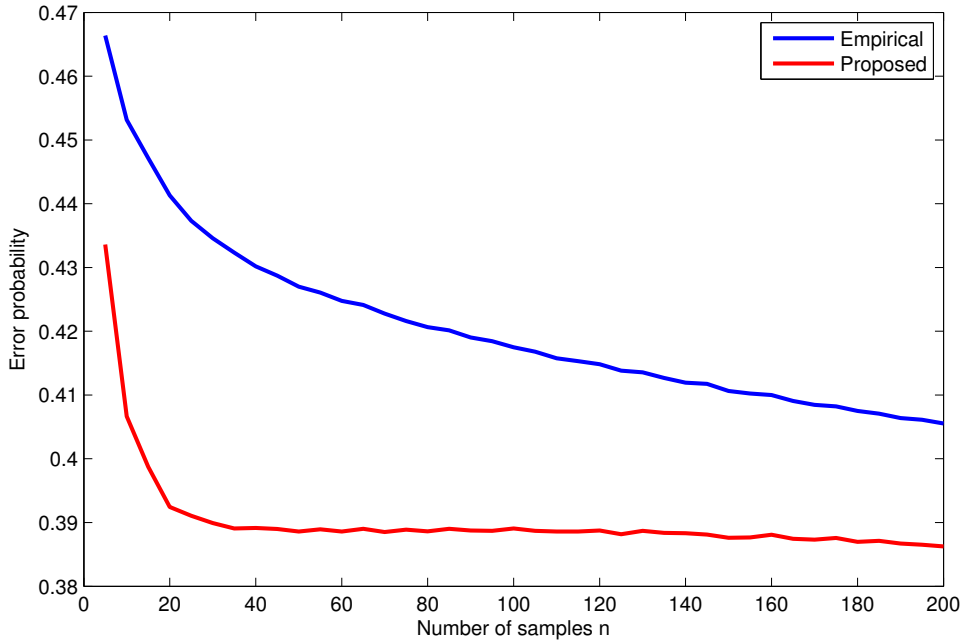
Summary: topics not covered

- Part 4: Closeness testing and classification
- Part 5: Competitive testing classification
- Part 6: Lower bounds

Competitive classification

Competitive classification

- $X^n \sim$ unknown p, $Y^n \sim$ unknown q
- $Z \sim$ p or q?
- $X^n = a\,b\,a\,b$, $Y^n = a\,b\,b\,b$, $Z = a$?
- Empirical classification:
  - Assign to distribution with higher occurrences
  - Suboptimal!



GT is competitively optimal irrespective of support k

Thank you