Chebyshev polynomials, moment matching and optimal estimation of the unseen

Yihong Wu

Department of ECE
University of Illinois at Urbana-Champaign

yihongwu@illinois.edu

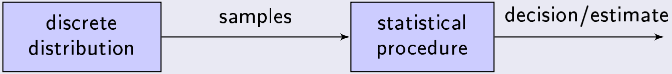
Joint work with Pengkun Yang (Illinois)

Mar 17, 2014

Problem setup

Task

Given samples from a discrete distribution, how to make statistical inference on certain property of the distribution?



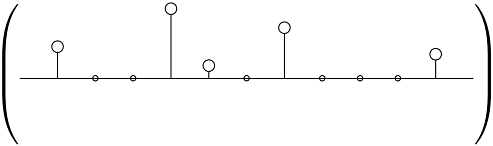
Estimating the unseen

- Support size:

$$S(P) = \sum_i \mathbf{1}_{\{p_i > 0\}}$$

- Example:

$$S$$



$$= 5$$

- $\Leftrightarrow$ estimating the number of unseens (SEEN + UNSEEN = $S(P)$)

Classical results

Classical results

- maybe the Egyptians have studied it...

Classical results

- maybe the Egyptians have studied it...
- Ecology:

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*)

Classical results

- maybe the Egyptians have studied it...
- Ecology:

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*)

- Linguistics, numismatics, etc:

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

BY BRADLEY EFRON AND RONALD THISTED

Department of Statistics, Stanford University, California

Classical results

- maybe the Egyptians have studied it...
- Ecology:

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*)

- Linguistics, numismatics, etc:

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

BY BRADLEY EFRON AND RONALD THISTED

Department of Statistics, Stanford University, California

- Will not discuss probability estimation [Good-Turing, Orlitsky et al., ...]

Mathematical formulation

- Data: $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P$

Mathematical formulation

- Data: $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P$
- Estimate: $\hat{S} = \hat{S}(X_1, \ldots, X_n)$ close to $S(P)$ in prob or expectation

Mathematical formulation

- Data: $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P$
- Estimate: $\hat{S} = \hat{S}(X_1, \ldots, X_n)$ close to $S(P)$ in prob or expectation
- Goal: find minimal sample size & fast algorithms

Mathematical formulation

- Data: $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P$
- Estimate: $\hat{S} = \hat{S}(X_1, \ldots, X_n)$ close to $S(P)$ in prob or expectation
- Goal: find minimal sample size & fast algorithms
- Need to assume minimum non-zero mass

Sample complexity

Space of distributions

$$\mathcal{D}_k \triangleq \{\text{prob distributions whose non-zero mass is at least } 1/k\}$$

Sample complexity

Space of distributions

$$\mathcal{D}_k \triangleq \{\text{prob distributions whose non-zero mass is at least } 1/k\}$$

Sample complexity

$$n^*(k,\epsilon) \triangleq \min\{n : \exists \hat{S}, \text{s.t. } \mathbb{P}[|\hat{S} - S(P)| \le \epsilon k] \ge 0.5, \forall P \in \mathcal{D}_k\}$$

Sample complexity

Space of distributions

$$\mathcal{D}_k \triangleq \{\text{prob distributions whose non-zero mass is at least } 1/k\}$$

Sample complexity

$$n^*(k,\epsilon) \triangleq \min\{n : \exists \hat{S}, \text{s.t. } \mathbb{P}[|\hat{S} - S(P)| \le \epsilon k] \ge 0.5, \forall P \in \mathcal{D}_k\}$$

Remarks

- Upgrade the confidence: $n \to n \log \frac{1}{\delta} \Rightarrow 0.5 \to 1-\delta$ (subsample + median + Hoeffding)

Sample complexity

Space of distributions

$$\mathcal{D}_k \triangleq \{\text{prob distributions whose non-zero mass is at least } 1/k\}$$

Sample complexity

$$n^*(k,\epsilon) \triangleq \min\{n : \exists \hat{S}, \text{s.t. } \mathbb{P}[|\hat{S} - S(P)| \le \epsilon k] \ge 0.5, \forall P \in \mathcal{D}_k\}$$

Remarks

- Upgrade the confidence: $n \to n \log \frac{1}{\delta} \Rightarrow 0.5 \to 1-\delta$ (subsample + median + Hoeffding)
- Zero error ($\epsilon = 0$): $n^*(k,0) \asymp k \log k$ (coupon collector)

Naive approach: plug-in

- WYSIWYE:

$$\hat{S}_{\text{seen}} = \text{ number of seen symbols}$$

Naive approach: plug-in

- WYSIWYE:

$$\hat{S}_{\mathsf{seen}} = \text{ number of seen symbols}$$

- underestimate:

$$\hat{S}_{\mathsf{seen}} \leq S(P), \quad P\text{-a.s.}$$

- severely underbiased in the sublinear-sampling regime: $n \ll k$

Do we have to estimate the distribution itself?

From a statistical perspective

- high-dimensional problem
  - estimating P provably requires $n = \Theta(k)$ samples
  - empirical distribution is optimal up to constants
- functional estimation
  - scalar functional (support size) $\overset{?}{\Rightarrow} n = o(k)$ suffices
  - plug-in is frequently suboptimal

Sufficient statistics

- Histogram:

$$N_j = \sum_i \mathbf{1}_{\{X_i = j\}} : \text{ \# of occurences of } j^{\text{th}} \text{ symbol}$$

Sufficient statistics

- Histogram:

$$N_j = \sum_i \mathbf{1}_{\{X_i = j\}} : \text{ \# of occurences of } j^{\text{th}} \text{ symbol}$$

- Histogram^2/fingerprints/profiles:

$$h_i = \sum_j \mathbf{1}_{\{N_j = i\}} : \text{ \# of symbols that occured exactly } i \text{ times}$$

Sufficient statistics

- Histogram:

$$N_j = \sum_i \mathbf{1}_{\{X_i = j\}} : \text{ \# of occurences of } j^{\text{th}} \text{ symbol}$$

- Histogram^2/fingerprints/profiles:

$$h_i = \sum_j \mathbf{1}_{\{N_j = i\}} : \text{ \# of symbols that occured exactly } i \text{ times}$$

- h_0: # of unseens

Linear estimators

Estimators that are linear in the fingerprints:

$$\hat{S} = \sum_i f(N_i) = \sum_{j \geq 1} f(j) h_j$$

Linear estimators

Estimators that are linear in the fingerprints:

$$\hat{S} = \sum_i f(N_i) = \sum_{j \geq 1} f(j) h_j$$

Classical procedures:

- Plug-in:

$$\hat{S}_{\mathsf{seen}} = h_1 + h_2 + h_3 + \dots$$

Linear estimators

Estimators that are linear in the fingerprints:

$$\hat{S} = \sum_i f(N_i) = \sum_{j \geq 1} f(j) h_j$$

Classical procedures:

- Plug-in:

$$\hat{S}_{\mathsf{seen}} = h_1 + h_2 + h_3 + \ldots$$

- Good-Toulmin '56: empirical Bayes

$$\hat{S}_{\mathsf{GT}} = t h_1 - t^2 h_2 + t^3 h_3 - t^4 h_4 + \ldots$$

Linear estimators

Estimators that are linear in the fingerprints:

$$\hat{S} = \sum_i f(N_i) = \sum_{j \geq 1} f(j) h_j$$

Classical procedures:

- Plug-in:

$$\hat{S}_{\mathsf{seen}} = h_1 + h_2 + h_3 + \ldots$$

- Good-Toulmin '56: empirical Bayes

$$\hat{S}_{\mathsf{GT}} = t h_1 - t^2 h_2 + t^3 h_3 - t^4 h_4 + \ldots$$

- Efron-Thisted '76: Bayesian

$$\hat{S}_{\mathsf{ET}} = \sum_{j=1}^{J} (-1)^{j+1} t^j b_j h_j$$

where $b_j = \mathbb{P}[\mathrm{Binomial}(J, 1/(t+1)) \geq j]$

State of the art

- $\hat{S}_{\mathsf{seen}}$: $n^*(k,\epsilon) \leq k \log \frac{1}{\epsilon}$

State of the art

- $\hat{S}_{\mathsf{seen}}$: $n^*(k,\epsilon) \leq k \log \frac{1}{\epsilon}$
- Valiant '08, Raskhodnikova et al. '09, Valiant-Valiant '11-'13: sublinear is possible.
  - Upper bound: $n^*(k,\epsilon) \lesssim \frac{k}{\log k} \frac{1}{\epsilon^2}$ by LP [Efron-Thisted '76]
  - Lower bound: $n^*(k,\epsilon) \gtrsim \frac{k}{\log k}$

State of the art

- $\hat{S}_{\text{seen}}$: $n^*(k,\epsilon) \le k \log \frac{1}{\epsilon}$
- Valiant '08, Raskhodnikova et al. '09, Valiant-Valiant '11-'13: sublinear is possible.
  - Upper bound: $n^*(k,\epsilon) \lesssim \frac{k}{\log k} \frac{1}{\epsilon^2}$ by LP [Efron-Thisted '76]
  - Lower bound: $n^*(k,\epsilon) \gtrsim \frac{k}{\log k}$

Theorem (W.-Yang '14)

$$n^*(k,\epsilon) \asymp \frac{k}{\log k} \log^2 \frac{1}{\epsilon}$$

Minimax risk

Theorem (W.-Yang '14)

$$\inf_{\hat{S}} \sup_{P \in \mathcal{D}_k} \mathbb{E}[(\hat{S} - S(P))^2] \asymp k^2 \exp\left(-\sqrt{\frac{n \log k}{k}} \vee \frac{n}{k}\right)$$

Remainder of this talk

Objectives

- a principled way to obtain rate-optimal linear estimator
- a natural lower bound to establish optimality via duality

Best polynomial approximation

Best polynomial approximation

- $\mathcal{P}_L = \{$polynomials of degree at most $L\}$.
- $I = [a, b]$: a finite interval.
- Optimal approximation error

$$E_L(f, I) \triangleq \inf_{p \in \mathcal{P}_L} \sup_{x \in I} |f(x) - p(x)|$$

Best polynomial approximation

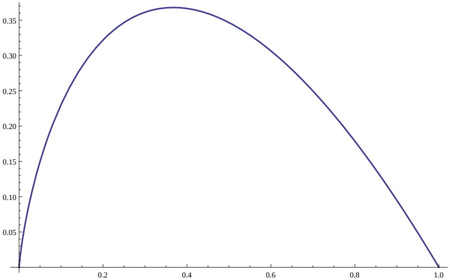
- $\mathcal{P}_L = \{$polynomials of degree at most $L\}$.
- $I = [a, b]$: a finite interval.
- Optimal approximation error

$$E_L(f, I) \triangleq \inf_{p \in \mathcal{P}_L} \sup_{x \in I} |f(x) - p(x)|$$

- Stone-Weierstrass theorem: f continuous $\Rightarrow E_L(f, I) \xrightarrow{L \to \infty} 0$
- Speed of convergence related to modulus of continuity.
- Finite-dim convex optimization/Infinite-dim LP
- Many fast algorithms (e.g., Remez)

Example

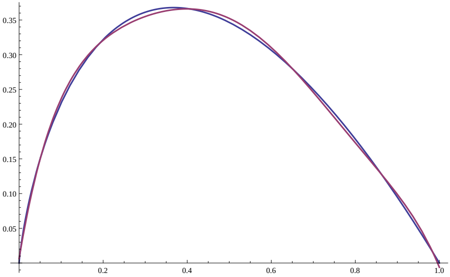
deg-6 approximation



Chebyshev alternation theorem

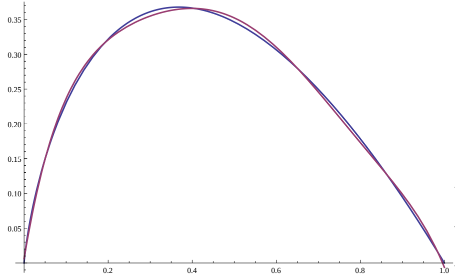
Example

deg-6 approximation

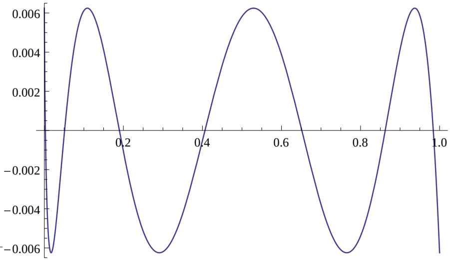


Chebyshev alternation theorem

deg-6 approximation



Chebyshev alternation theorem



Moment matching

$$\begin{aligned}\mathcal{E}_L(f,I) \triangleq \sup\ & \mathbb{E}\left[f(U)\right] - \mathbb{E}\left[f(U')\right] \\ \text{s.t. } & \mathbb{E}\left[U^j\right] = \mathbb{E}\left[U'^j\right], \quad j = 1,\ldots,L \\ & U, U' \in I\end{aligned}$$

Moment matching

$$
\begin{aligned}
\mathcal{E}_L(f, I) \triangleq \sup \; & \int f \mathrm{d}\mu - \int f \mathrm{d}\mu' \\
\text{s.t.} \; & \int f \mathrm{d}\mu = \int f \mathrm{d}\mu', \quad j = 1, \ldots, L, \\
& \mu, \mu' \text{ supported on } I
\end{aligned}
$$

Moment matching

$$
\begin{aligned}
\mathcal{E}_L(f, I) \triangleq \sup\ & \mathbb{E}\left[f(U)\right] - \mathbb{E}\left[f(U')\right] \\
\text{s.t. } & \mathbb{E}\left[U^j\right] = \mathbb{E}\left[U'^j\right], \quad j = 1, \ldots, L, \\
& U, U' \in I
\end{aligned}
$$

Moment matching

$$\begin{aligned}\mathcal{E}_L(f, I) \triangleq \sup\ & \mathbb{E}\left[f(U)\right] - \mathbb{E}\left[f(U')\right] \\ \text{s.t. } & \mathbb{E}\left[U^j\right] = \mathbb{E}\left[U'^j\right], \quad j = 1, \ldots, L, \quad \lambda_j \in \mathbb{R} \\ & U, U' \in I\end{aligned}$$

Infinite-dim linear programming. Dual:

$$\inf_{\lambda_1^L} \sup_{U, U' \in I} \mathbb{E}\left[f(U)\right] - \mathbb{E}\left[f(U')\right] + \sum_j \lambda_j (\mathbb{E}\left[U^j\right] - \mathbb{E}\left[U'^j\right])$$

Moment matching

$$\mathcal{E}_L(f,I) \triangleq \sup \mathbb{E}\left[f(U)\right] - \mathbb{E}\left[f(U')\right]$$
$$\text{s.t. } \mathbb{E}\left[U^j\right] = \mathbb{E}\left[U'^j\right], \quad j=1,\ldots,L, \quad \lambda_j \in \mathbb{R}$$
$$U, U' \in I$$

Infinite-dim linear programming. Dual:

$$\inf_{\lambda_1^L} \sup_{U,U' \in I} \mathbb{E}\left[f(U)\right] - \mathbb{E}\left[f(U')\right] + \sum_j \lambda_j (\mathbb{E}\left[U^j\right] - \mathbb{E}\left[U'^j\right])$$

Moment matching

$$\mathcal{E}_L(f, I) \triangleq \sup \mathbb{E}\left[f(U)\right] - \mathbb{E}\left[f(U')\right]$$
$$\text{s.t. } \mathbb{E}\left[U^j\right] = \mathbb{E}\left[U'^j\right], \quad j = 1, \ldots, L, \quad {\color{red}\lambda_j \in \mathbb{R}}$$
$$U, U' \in I$$

Infinite-dim linear programming. Dual:

$$\inf_{\lambda_1^L} \sup_{U,U' \in I} \mathbb{E}\left[f(U)\right] - \mathbb{E}\left[f(U')\right] + \sum_j \lambda_j (\mathbb{E}\left[U^j\right] - \mathbb{E}\left[U'^j\right])$$
$$= \inf_{\lambda_1^L} \sup_{U \in I} \mathbb{E}\Big[f(U) - \sum_j \lambda_j U^j\Big] - \inf_{U' \in I} \mathbb{E}\Big[f(U') - \sum_j \lambda_j U'^j\Big]$$

Moment matching

$$\begin{aligned}\mathcal{E}_L(f,I) \triangleq \sup\ & \mathbb{E}\left[f(U)\right] - \mathbb{E}\left[f(U')\right] \\ \text{s.t. } & \mathbb{E}\left[U^j\right] = \mathbb{E}\left[U'^j\right], \quad j = 1,\ldots,L, \quad \lambda_j \in \mathbb{R} \\ & U, U' \in I\end{aligned}$$

Infinite-dim linear programming. Dual:

$$\begin{aligned}& \inf_{\lambda_1^L} \sup_{U,U'\in I} \mathbb{E}\left[f(U)\right] - \mathbb{E}\left[f(U')\right] + \sum_j \lambda_j (\mathbb{E}\left[U^j\right] - \mathbb{E}\left[U'^j\right]) \\ = & \inf_{\lambda_1^L} \sup_{U\in I} \mathbb{E}\Big[f(U) - \sum_j \lambda_j U^j\Big] - \inf_{U'\in I} \mathbb{E}\Big[f(U') - \sum_j \lambda_j U'^j\Big] \\ = & \inf_{\lambda_0^L} \Big(\sup_{u\in I} f(u) - \sum_j \lambda_j u^j\Big) - \Big(\inf_{u\in I} f(u') - \sum_j \lambda_j u^j\Big)\end{aligned}$$

Moment matching

$$\begin{aligned}\mathcal{E}_L(f,I) \triangleq \sup\ & \mathbb{E}\left[f(U)\right] - \mathbb{E}\left[f(U')\right] \\ \text{s.t. } & \mathbb{E}\left[U^j\right] = \mathbb{E}\left[U'^j\right], \quad j = 1,\ldots,L, \quad {\color{red}\lambda_j \in \mathbb{R}} \\ & U, U' \in I\end{aligned}$$

Infinite-dim linear programming. Dual:

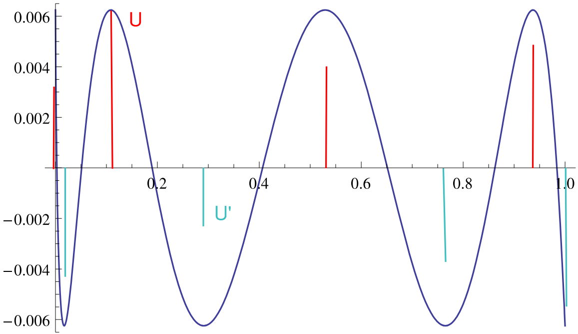
$$\begin{aligned}& \inf_{\lambda_1^L} \sup_{U,U' \in I} \mathbb{E}\left[f(U)\right] - \mathbb{E}\left[f(U')\right] + \sum_j \lambda_j (\mathbb{E}\left[U^j\right] - \mathbb{E}\left[U'^j\right]) \\ = & \inf_{\lambda_1^L} \sup_{U \in I} \mathbb{E}\left[f(U) - \sum_j \lambda_j U^j\right] - \inf_{U' \in I} \mathbb{E}\left[f(U') - \sum_j \lambda_j U'^j\right] \\ = & \inf_{\lambda_0^L} \left(\sup_{u \in I} f(u) - \sum_j \lambda_j u^j\right) - \left(\inf_{u \in I} f(u') - \sum_j \lambda_j u^j\right) \\ = & \ 2 \inf_{p \in \mathcal{P}_L} \sup_{u \in I} |f(u) - p(u)|\end{aligned}$$

Moment matching $\Leftrightarrow$ best polynomial approximation

$$\boxed{\mathcal{E}_L(f, I) = 2E_L(f, I)}$$

Moment matching $\Leftrightarrow$ best polynomial approximation

$$\boxed{\mathcal{E}_L(f, I) = 2E_L(f, I)}$$



Optimal estimator

Poissonization

- Poisson sampling model
  - draw sample size $n' \sim \mathrm{Poi}(n)$
  - draw n' i.i.d. samples from P.
- Histograms are independent: $N_i \overset{\text{ind}}{\sim} \mathrm{Poi}(np_i)$
- sample complexity/minimax risks remain unchanged within constant factors

MSE

Recall

$$\text{MSE} = \text{bias}^2 + \text{variance}$$

Main problem of $\hat{S}_{\text{seen}}$: huge bias.

Unbiased estimators?

Unbiased estimators?

Unbiased estimator for $f(P)$ from n samples:

- Independent sampling: $f(P)$ is polynomial of degree $\leq n$
- Poissonized sampling: $f(P)$ is real analytic.

Unbiased estimators?

Unbiased estimator for $f(P)$ from n samples:

- Independent sampling: $f(P)$ is polynomial of degree $\leq n$
- Poissonized sampling: $f(P)$ is real analytic.

Example

- Flip a coin with bias p for n times and estimate $f(p)$
- Sufficient stat: $Y \sim \mathrm{Binomial}(n, p)$.
- Unbiased estimator exists $\Leftrightarrow$ $f(p)$ is a polynomial of degree $\leq n$

$$\mathbb{E}[\hat{f}(Y)] = \sum_{k=0}^{n} \hat{f}(k) \binom{n}{k} p^k (1-p)^k.$$

No unbiased estimator

$$S(P) = \sum_i \mathbf{1}_{\{p_i > 0\}}$$

No unbiased estimator

$$S(P) = \sum_i \mathbf{1}_{\{p_i > 0\}}$$

- Approximate $\mathbf{1}_{\{x>0\}}$ by $q(x) = \sum_{m=0}^{L} a_m x^m$
- Find an unbiased estimator for the proxy

$$\tilde{S}(P) = \sum_i q(p_i)$$

- $|\text{bias}| \leq$ uniform approx error

No unbiased estimator

$$S(P) = \sum_i \mathbf{1}_{\{p_i > 0\}}$$

- Approximate $\mathbf{1}_{\{x>0\}}$ by $q(x) = \sum_{m=0}^{L} a_m x^m$
- Find an unbiased estimator for the proxy

$$\tilde{S}(P) = \sum_i q(p_i)$$

- $|\text{bias}| \leq$ uniform approx error
- But the function is discontinuous...

Linear estimators

Consider estimators that are linear in the fingerprints:

$$\hat{S} = \sum_i f(N_i) = \sum_{j \geq 1} f(j) h_j$$

Guidelines:

- $f(0) = 0$

Linear estimators

Consider estimators that are linear in the fingerprints:

$$\hat{S} = \sum_i f(N_i) = \sum_{j \geq 1} f(j) h_j$$

Guidelines:

- $f(0) = 0$
- $f(j) = 1$ for sufficiently large $j > L$

Linear estimators

Consider estimators that are linear in the fingerprints:

$$\hat{S} = \sum_i f(N_i) = \sum_{j \geq 1} f(j) h_j$$

Guidelines:

- $f(0) = 0$
- $f(j) = 1$ for sufficiently large $j > L$
- How to choose $f(1), \ldots, f(L)$?

Bias

Choose

- $L = c_0 \log k$.
- $\hat{S} = \sum_{j \geq 1} f(N_i), \quad N_i \sim \mathrm{Poi}(np_i)$

Bias:

$$\mathbb{E}[\hat{S} - S] = \sum \mathbb{E}[(f(N_i) - 1)\mathbf{1}_{\{N_i \leq L\}}]\mathbf{1}_{\{p_i > 1/k\}}$$

Choose

- $L = c_0 \log k$.
- $\hat{S} = \sum_{j \geq 1} f(N_i), \quad N_i \sim \mathrm{Poi}(np_i)$

Bias:

$$\begin{aligned} \mathbb{E}[\hat{S} - S] &= \sum \mathbb{E}[(f(N_i) - 1)\mathbf{1}_{\{N_i \leq L\}}]\mathbf{1}_{\{p_i > 1/k\}} \\ &\approx \sum \underbrace{\mathbb{E}[(f(N_i) - 1)\mathbf{1}_{\{N_i \leq L\}}]}\mathbf{1}_{\{2L/n > p_i > 1/k\}} \end{aligned}$$

Bias

Choose

- $L = c_0 \log k$.
- $\hat{S} = \sum_{j \geq 1} f(N_i), \quad N_i \sim \text{Poi}(np_i)$

Bias:

$$\begin{aligned}
\mathbb{E}[\hat{S} - S] &= \sum \mathbb{E}[(f(N_i) - 1)\mathbf{1}_{\{N_i \leq L\}}]\mathbf{1}_{\{p_i > 1/k\}} \\
&\approx \sum \underbrace{\mathbb{E}[(f(N_i) - 1)\mathbf{1}_{\{N_i \leq L\}}]}_{e^{-np_i} \times \text{ poly of deg } L!} \mathbf{1}_{\{2L/n > p_i > 1/k\}}
\end{aligned}$$

Bias

- Observe

$$\mathbb{E}[(f(N)-1)\mathbf{1}_{\{N\leq L\}}] = e^{-\lambda} \underbrace{\sum_{j\geq 0} \frac{f(j)-1}{j!}\lambda^j}_{q(\lambda)}$$

- Then

$$|\text{bias}| \leq k \sup_{n/k\leq\lambda\leq c\log k} |q(\lambda)|$$

- Choose the best deg-L polynomial q s.t. $q(0)=-1$

Bias

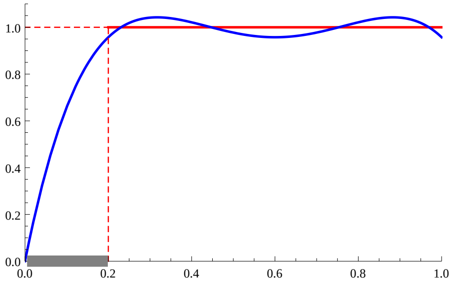
- Observe

$$\mathbb{E}[(f(N)-1)\mathbf{1}_{\{N\le L\}}] = \underbrace{e^{-\lambda}\sum_{j\ge 0}\frac{f(j)-1}{j!}\lambda^j}_{q(\lambda)}$$

- Then

$$|\text{bias}| \le k \sup_{n/k\le\lambda\le c\log k} |q(\lambda)|$$

- Choose the best deg-L polynomial q s.t. $q(0)=-1$
- Solution: Chebyshev polynomial



best approximation to one by polynomial passing through origin is Chebyshev polynomial

$$p_L(x) = 1 - \frac{\cos L \arccos x}{\cos L \arccos a}$$

Final estimator

- Chebyshev polynomial: $r \triangleq c_1 \log k$ and $l \triangleq \frac{n}{k}$,

$$-\frac{\cos L \arccos(\frac{2}{r-l}x - \frac{r+l}{r-l})}{\cos L \arccos(-\frac{r+l}{r-l})} \triangleq \sum_{j=0}^{L} a_m x^m.$$

Final estimator

- Chebyshev polynomial: $r \triangleq c_1 \log k$ and $l \triangleq \frac{n}{k}$,

$$-\frac{\cos L \arccos(\frac{2}{r-l}x - \frac{r+l}{r-l})}{\cos L \arccos(-\frac{r+l}{r-l})} \triangleq \sum_{j=0}^{L} a_m x^m.$$

- Choose

$$f(j) = \begin{cases} 0 & j = 0 \\ 1 + a_j j! & j = 1, \ldots, L \\ 1 & j > L. \end{cases}$$

Final estimator

- Chebyshev polynomial: $r \triangleq c_1 \log k$ and $l \triangleq \frac{n}{k}$,

$$-\frac{\cos L \arccos(\frac{2}{r-l}x - \frac{r+l}{r-l})}{\cos L \arccos(-\frac{r+l}{r-l})} \triangleq \sum_{j=0}^{L} a_m x^m.$$

- Choose

$$f(j) = \begin{cases} 0 & j = 0 \\ 1 + a_j j! & j = 1, \dots, L \\ 1 & j > L. \end{cases}$$

- Linear estimator (precomputable coefficients): no sample splitting!!

$$\hat{S} = \sum_{j=1}^{L} f(j) h_j + \sum_{j>L} h_j$$

Final estimator

- Chebyshev polynomial: $r \triangleq c_1 \log k$ and $l \triangleq \frac{n}{k}$,

$$-\frac{\cos L \arccos(\frac{2}{r-l}x - \frac{r+l}{r-l})}{\cos L \arccos(-\frac{r+l}{r-l})} \triangleq \sum_{j=0}^{L} a_m x^m.$$

- Choose

$$f(j) = \begin{cases} 0 & j = 0 \\ 1 + a_j j! & j = 1, \dots, L \\ 1 & j > L. \end{cases}$$

- Linear estimator (precomputable coefficients): no sample splitting!!

$$\hat{S} = \sum_{j=1}^{L} f(j) h_j + \sum_{j>L} h_j$$

- Significantly faster than LP [Efron-Thisted '76, Valiant-Valiant '11]

Analysis

1. bias $\leq$ approximation error of Chebyshev polynomial:

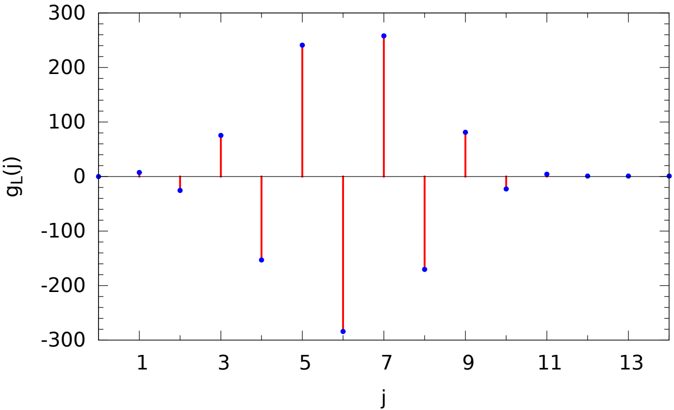
$$\frac{1}{|\cos M \arccos(-\frac{r+l}{r-l})|} \asymp \exp\left(-c\sqrt{\frac{n\log k}{k}}\right),$$

2. variance $\approx$ poly(k).

Optimal estimator

Plot of coefficients ($k = 10^6$ and $n = 2 \times 10^5$):

$$\hat{S} = \sum_{j \geq 1} f(j) h_j$$



Why oscillatory and alternating?

$$\hat{S} = \sum_{j \geq 1} f(j) h_j$$

The same oscillation also happens in:

- Good-Toulmin '56: empirical Bayes

$$\hat{S}_{\mathsf{GT}} = th_1 - t^2 h_2 + t^3 h_3 - t^4 h_4 + \ldots$$

- Efron-Thistle '76: Bayesian

$$\hat{S}_{\mathsf{ET}} = \sum_{j=1}^{J} (-1)^{j+1} t^j b_j h_j$$

I HAVE NO EXPLANATION!

Impossibility results

Minimax lower bound

$$n^*(k,\epsilon) \gtrsim \frac{k}{\log k} \log^2 \frac{1}{\epsilon}$$

Total variation

- $\mathsf{TV}(P_0, P_1) = \frac{1}{2}\int |\mathrm{d}P_0 - \mathrm{d}P_1|$
- optimal error probability for testing P_0 vs P_1

$$1 - \mathsf{TV}(P_0, P_1) = \min_{\psi} P_0[\psi = 1] + P_1[\psi = 0]$$

Poisson mixtures

Given $U \sim \mu$,

$$\mathbb{E}[\mathrm{Poi}(U)] = \int_{\mathbb{R}_+} \mathrm{Poi}(\lambda)\mu(\mathrm{d}\lambda)$$

Randomization

Two-prior argument (composite HT):

- draw random distribution P $\xrightarrow{\text{Poisson}} N_i \overset{\text{ind}}{\sim} \mathrm{Poi}(n\mathsf{p}_i)$
- draw random distribution P′ $\xrightarrow{\text{Poisson}} N_i' \overset{\text{ind}}{\sim} \mathrm{Poi}(n\mathsf{p}_i')$

Randomization

Two-prior argument (composite HT):

- draw random distribution $\mathsf{P} \xrightarrow{\text{Poisson}} N_i \overset{\text{ind}}{\sim} \mathrm{Poi}(n\mathsf{p}_i)$
- draw random distribution $\mathsf{P}' \xrightarrow{\text{Poisson}} N_i' \overset{\text{ind}}{\sim} \mathrm{Poi}(n\mathsf{p}_i')$

Le Cam's lemma applies if

- $S(\mathsf{P})$ and $S(\mathsf{P}')$ differ with high probability
- Distributions of N and N' are indistinguishable (k-dim Poisson mixtures)

Randomization

Two-prior argument (composite HT):

- draw random distribution $\mathsf{P} \xrightarrow{\text{Poisson}} N_i \overset{\text{ind}}{\sim} \mathrm{Poi}(n\mathsf{p}_i)$
- draw random distribution $\mathsf{P}' \xrightarrow{\text{Poisson}} N_i' \overset{\text{ind}}{\sim} \mathrm{Poi}(n\mathsf{p}_i')$

Le Cam's lemma applies if

- $S(\mathsf{P})$ and $S(\mathsf{P}')$ differ with high probability
- Distributions of N and N' are indistinguishable (k-dim Poisson mixtures)

Main hurdle: difficult to work with distributions on high-dimensional probability simplex.

Key construction: reduction to one dimension

- Given U, U' with unit mean:

$$\mathsf{P} = \frac{1}{k}(\underbrace{U_1, \ldots, U_k}_{\overset{\text{i.i.d.}}{\sim} U}), \quad \mathsf{P}' = \frac{1}{k}(\underbrace{U'_1, \ldots, U'_k}_{\overset{\text{i.i.d.}}{\sim} U'})$$

- By LLN,
  - P and P' are not, but close to, probability distributions.

Key construction: reduction to one dimension

- Given U, U' with unit mean:

$$\mathsf{P} = \frac{1}{k}(\underbrace{U_1, \dots, U_k}_{\overset{\text{i.i.d.}}{\sim} U}), \quad \mathsf{P}' = \frac{1}{k}(\underbrace{U_1', \dots, U_k'}_{\overset{\text{i.i.d.}}{\sim} U'})$$

- By LLN,
  - P and P' are not, but close to, probability distributions.
  - support size concentrates on the mean:

$$\mathbb{E}\left[S(\mathsf{P})\right] - \mathbb{E}\left[S(\mathsf{P}')\right] = k(\mathbb{P}\left\{U > 0\right\} - \mathbb{P}\left\{U' > 0\right\})$$

Key construction: reduction to one dimension

- Given U, U' with unit mean:

$$\mathsf{P} = \frac{1}{k}(\underbrace{U_1, \ldots, U_k}_{\overset{\text{i.i.d.}}{\sim} U}), \quad \mathsf{P}' = \frac{1}{k}(\underbrace{U'_1, \ldots, U'_k}_{\overset{\text{i.i.d.}}{\sim} U'})$$

- By LLN,
  - P and P' are not, but close to, probability distributions.
  - support size concentrates on the mean:

$$\mathbb{E}\left[S(\mathsf{P})\right] - \mathbb{E}\left[S(\mathsf{P}')\right] = k(\mathbb{P}\left\{U > 0\right\} - \mathbb{P}\left\{U' > 0\right\})$$

- Sufficient statistic are iid:

$$N_i \overset{\text{i.i.d.}}{\sim} \mathbb{E}[\mathrm{Poi}(nU/k)], \quad N'_i \overset{\text{i.i.d.}}{\sim} \mathbb{E}[\mathrm{Poi}(nU'/k)].$$

- Suffice to show $\mathsf{TV}(\underbrace{\mathbb{E}[\mathrm{Poi}(nU/k)], \mathbb{E}[\mathrm{Poi}(nU'/k)]}_{\text{one-dimensional Poisson mixtures}}) = o(1/k)$.

Moment matching $\Rightarrow$ statistically close Poisson mixtures

Lemma

- $U, U' \in [0, \frac{k \log k}{n}]$
- $\mathbb{E}\left[U^j\right] = \mathbb{E}\left[U'^j\right], \; j = 1, \ldots, L = C \log k$
- *Then*
$$\mathsf{TV}(\mathbb{E}\left[\mathrm{Poi}\left(nU/k\right)\right], \mathbb{E}\left[\mathrm{Poi}\left(nU'/k\right)\right]) = o(1/k)$$

Optimize the lower bound

Let $\lambda = k \log k / n$.

Choose the best U, U':

$$\begin{aligned} \sup\ & \mathbb{P}\left\{U = 0\right\} - \mathbb{P}\left\{U' = 0\right\} \\ \text{s.t. } & \mathbb{E}\left[U\right] = \mathbb{E}\left[U'\right] = 1 \\ & \mathbb{E}\left[U^j\right] = \mathbb{E}\left[U'^j\right], \quad j \in [L] \\ & U, U' \in \{0\} \cup [1, \lambda] \end{aligned}$$

Optimize the lower bound

Let $\lambda = k \log k/n$.

Choose the best U, U':

$$\begin{aligned} \sup\ & \mathbb{P}\{U = 0\} - \mathbb{P}\left\{U' = 0\right\} \\ \text{s.t. } & \mathbb{E}[U] = \mathbb{E}\left[U'\right] = 1 \\ & \mathbb{E}\left[U^j\right] = \mathbb{E}\left[U'^j\right], \quad j \in [L] \\ & U, U' \in \{0\} \cup [1, \lambda] \end{aligned}$$

Optimize the lower bound

Let $\lambda = k \log k/n$.

Choose the best U, U':

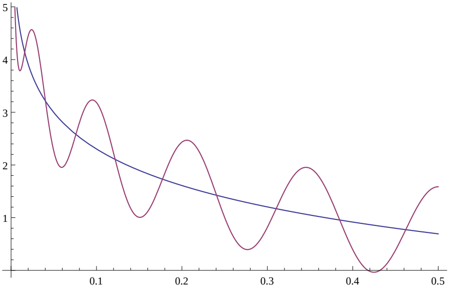
$$
\begin{aligned}
& \sup \ \mathbb{P}\{U = 0\} - \mathbb{P}\left\{U' = 0\right\} \\
& \quad \text{s.t. } \mathbb{E}[U] = \mathbb{E}\left[U'\right] = 1 \\
& \qquad \mathbb{E}\left[U^j\right] = \mathbb{E}\left[U'^j\right], \quad j \in [L] \\
& \qquad U, U' \in \{0\} \cup [1, \lambda] \\
= & \sup \ \mathbb{E}[1/X] - \mathbb{E}\left[1/X'\right] \\
& \quad \text{s.t. } \mathbb{E}\left[X^j\right] = \mathbb{E}\left[X'^j\right], \quad j \in [L] \\
& \qquad X, X' \in [1, \lambda],
\end{aligned}
$$

$$
P_U(\mathrm{d}u) = \left(1 - \mathbb{E}\left[\tfrac{1}{X}\right]\right) \delta_0(\mathrm{d}u) + \tfrac{1}{u} P_X(\mathrm{d}u)
$$

Let $\lambda = k \log k / n$.

Choose the best U, U':

$$
\begin{aligned}
& \sup \ \mathbb{P}\{U = 0\} - \mathbb{P}\left\{U' = 0\right\} \\
& \text{s.t. } \mathbb{E}[U] = \mathbb{E}\left[U'\right] = 1 \\
& \qquad \mathbb{E}\left[U^j\right] = \mathbb{E}\left[U'^j\right], \quad j \in [L] \\
& \qquad U, U' \in \{0\} \cup [1, \lambda] \\
= & \sup \ \mathbb{E}[1/X] - \mathbb{E}\left[1/X'\right] \\
& \text{s.t. } \mathbb{E}\left[X^j\right] = \mathbb{E}\left[X'^j\right], \quad j \in [L] \\
& \qquad X, X' \in [1, \lambda], \\
= & \quad 2E_L(1/x, [1, \lambda]) \gtrsim e^{-c\sqrt{\frac{n \log k}{k}}}
\end{aligned}
$$



Related work in statistics

Our inspiration: earlier work on Gaussian models

- Ibragimov-Nemirovskii-Khas'minskii '87: smooth functions
- Lepski-Nemirovski-Spokoiny '99: L_q norm of Gaussian regression function
- Cai-Low '11: L_1 norm of normal mean

Comparison

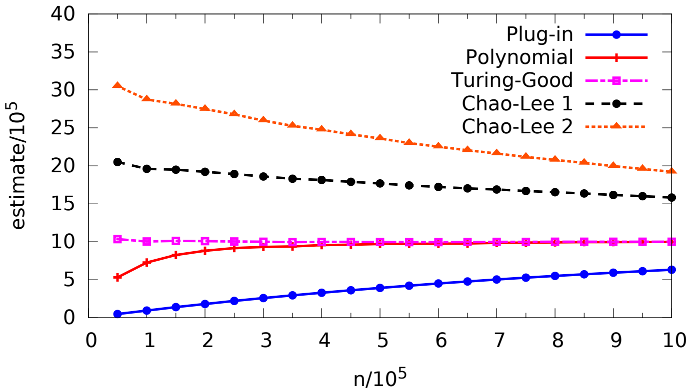
Lower bound in [Valiant-Valiant '11]

- Deal with fingerprints – high-dim distribution with dependent components
- Approximate distribution by quantized Gaussian
- Bound distance between mean and covariance matrices

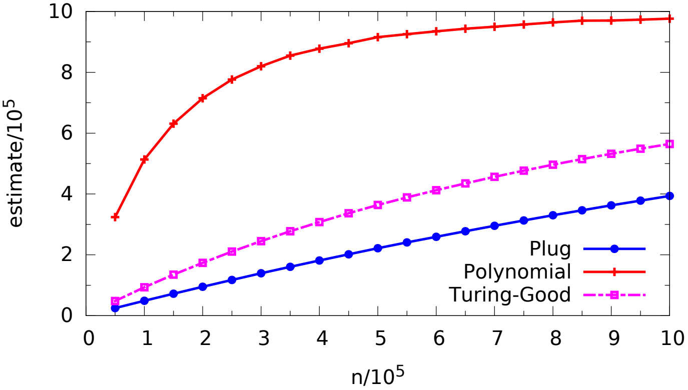
Lower bound here: reduce to one dimension

Experiments

Uniform over 1 million elements

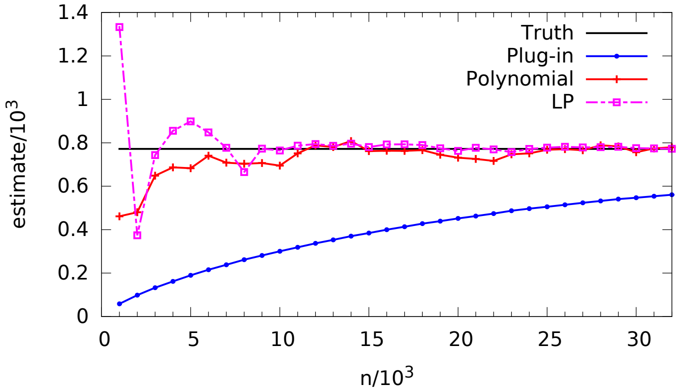


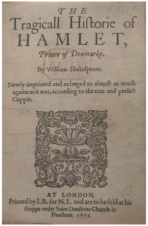
Uniform mixed with point mass



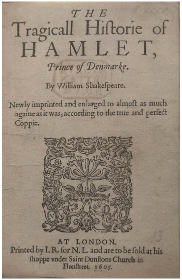
How many words did Shakespeare know?

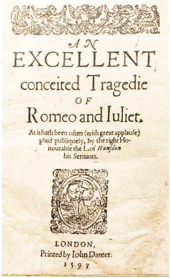
- Hamlet: total words 32000, total distinct words ~ 7700,
- deg-10 Chebyshev polynomial
- sampling with replacement
- compare with LP [Efron-Thisted '76, Valiant-Valiant '13]

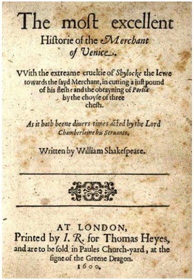




How many words did Shakespeare know?







Feed the entire Shakespearean canon into the estimator:

- $\hat{S} = 68944 \sim 73257$
- Efron-Thisted '76: 66534

Species problem

Formulation

Given an urn containing k balls, estimate the number of distinct colors S by sampling (e.g. with replacement).

- Special case of support size estimation: $p_i \in \{0, \frac{1}{k}, \frac{2}{k}, \ldots\}$.
- Same sample complexity as DISTINCT-ELEMENT problem in TCS.

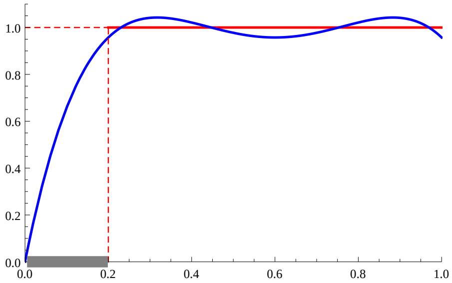
Species problem

Formulation

Given an urn containing k balls, estimate the number of distinct colors S by sampling (e.g. with replacement).

- Special case of support size estimation: $p_i \in \{0, \frac{1}{k}, \frac{2}{k}, \ldots\}$.
- Same sample complexity as DISTINCT-ELEMENT problem in TCS.
- Use Chebyshev: $\frac{k}{\log k}$ samples can achieve achieve $0.1k$
- Converse: $\frac{k}{\log k}$ samples are necessary to achieve $0.1k$ [Valiant '12]

Can we do better?

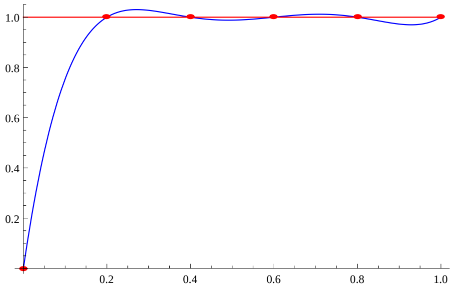


Use Lagrange interpolation polynomial to achieve zero bias

- Uniform approximation: $\epsilon \lesssim \exp(-c\sqrt{\log k})$
- Interpolation: $\epsilon \lesssim \exp(-c \log k)$.

$$q_L(x) = 1 - \frac{\prod_{j=1}^{L}(j-x)}{L!}$$

Can we do better?



Use Lagrange interpolation polynomial to achieve zero bias

- Uniform approximation: $\epsilon \lesssim \exp(-c\sqrt{\log k})$
- Interpolation: $\epsilon \lesssim \exp(-c \log k)$.

$$q_L(x) = 1 - \frac{\prod_{j=1}^{L}(j-x)}{L!}$$

More generally...

$$\text{minimax risk } \gtrsim k^2 \exp\left(-c\frac{n\log k}{k}\right)$$

- Tight when $n = 0.1k$
- Compare to general support size:

$$\text{minimax risk } \asymp k^2 \exp\left(-c\sqrt{\frac{n\log k}{k}}\right)$$

Estimating entropy

$$H(P) = \sum p_i \log \frac{1}{p_i}$$

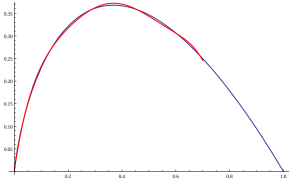
Theorem (W.-Yang '14)

Sample complexity to estimate within ϵ bits: $n \asymp \max\left\{ \frac{k}{\epsilon \log k}, \frac{\log^2 k}{\epsilon^2} \right\}$

(upper bound also in Jiao et al. '14)

Strategy

- degree: $L \sim \log k$
- small masses: polynomial approximation
- large masses: plug-in with bias correction
- coeff's bounded by Chebyshev



Estimating Rényi entropy

- Estimating $H_\alpha(P) = \frac{1}{1-\alpha}\log\sum p_i^\alpha$ [Jiao et al. '14, Acharya et al. '14]

Concluding remarks

To estimate

$$F(P) = \sum f(p_i)$$

Sample complexity is roughly governed by the following convex optimization problem (over logarithmic variables):

$$\begin{aligned} \mathcal{F}(\lambda) \triangleq \sup \quad & \mathbb{E}\left[f(U)\right] - \mathbb{E}\left[f(U')\right] \\ s.t. \quad & \mathbb{E}\left[U^j\right] = \mathbb{E}\left[U'^j\right] \;\; j = 1, \ldots, \log k, \\ & \mathbb{E}\left[U\right] \leq 1/k, \\ & U, U' \in [0, \log k/n], \end{aligned}$$

- Lower bound: primal program (inapproximability result)
- Upper bound: dual program (approximability result)

Concluding remarks

- Many open problems and directions
  - Confidence intervals
  - Adaptive estimation
  - How to go beyond iid sampling
  - How to incorporate structures

References

- W. & P. Yang (2014). *Minimax rates of entropy estimation on large alphabets via best polynomial approximation*. arXiv:1407.0381
- W. & P. Yang (2015). *Chebyshev polynomials, moment matching, and optimal estimation of the unseen*. arXiv:1503.xxxx

Bias

Choose

- $M = c \log k$.
- $\hat{S} = \sum_{j \geq 1} f(N_i), \quad N_i \sim \mathrm{Poi}(np_i)$

Bias:

$$\mathbb{E}[\hat{S} - S] = \sum \mathbb{E}[f(N_i)] - \mathbf{1}_{\{p_i > 0\}}$$

Bias

Choose

- $M = c \log k$.
- $\hat{S} = \sum_{j \geq 1} f(N_i), \quad N_i \sim \mathrm{Poi}(np_i)$

Bias:

$$\mathbb{E}[\hat{S} - S] = \sum \mathbb{E}[f(N_i)] - \mathbf{1}_{\{p_i > 0\}}$$

$$\overset{f(0)=0}{=} \sum \mathbb{E}[(f(N_i) - 1)]\mathbf{1}_{\{p_i > 0\}}$$

Bias

Choose

- $M = c \log k$.
- $\hat{S} = \sum_{j \geq 1} f(N_i), \quad N_i \sim \mathrm{Poi}(np_i)$

Bias:

$$
\begin{aligned}
\mathbb{E}[\hat{S} - S] &= \sum \mathbb{E}[f(N_i)] - \mathbf{1}_{\{p_i > 0\}} \\
&\overset{f(0)=0}{=} \sum \mathbb{E}[(f(N_i) - 1)] \mathbf{1}_{\{p_i > 0\}} \\
&= \sum \mathbb{E}[(f(N_i) - 1)] \mathbf{1}_{\{p_i > 1/k\}}
\end{aligned}
$$

Bias

Choose

- $M = c \log k$.
- $\hat{S} = \sum_{j \geq 1} f(N_i), \quad N_i \sim \mathrm{Poi}(np_i)$

Bias:

$$\begin{aligned}
\mathbb{E}[\hat{S} - S] &= \sum \mathbb{E}[f(N_i)] - \mathbf{1}_{\{p_i > 0\}} \\
&\overset{f(0)=0}{=} \sum \mathbb{E}[(f(N_i) - 1)]\mathbf{1}_{\{p_i > 0\}} \\
&= \sum \mathbb{E}[(f(N_i) - 1)]\mathbf{1}_{\{p_i > 1/k\}} \\
&= \sum \mathbb{E}[(f(N_i) - 1)\mathbf{1}_{\{N_i \leq L\}}]\mathbf{1}_{\{p_i > 1/k\}}
\end{aligned}$$

Choose

- $M = c \log k$.
- $\hat{S} = \sum_{j \geq 1} f(N_i), \quad N_i \sim \mathrm{Poi}(np_i)$

Bias:

$$
\begin{aligned}
\mathbb{E}[\hat{S} - S] &= \sum \mathbb{E}[f(N_i)] - \mathbf{1}_{\{p_i > 0\}} \\
&\overset{f(0)=0}{=} \sum \mathbb{E}[(f(N_i) - 1)] \mathbf{1}_{\{p_i > 0\}} \\
&= \sum \mathbb{E}[(f(N_i) - 1)] \mathbf{1}_{\{p_i > 1/k\}} \\
&= \sum \mathbb{E}[(f(N_i) - 1) \mathbf{1}_{\{N_i \leq L\}}] \mathbf{1}_{\{p_i > 1/k\}} \\
&\overset{whp}{=} \sum \underbrace{\mathbb{E}[(f(N_i) - 1) \mathbf{1}_{\{N_i \leq L\}}]} \mathbf{1}_{\{L/2n > p_i > 1/k\}}
\end{aligned}
$$

Choose

- $M = c \log k$.
- $\hat{S} = \sum_{j \geq 1} f(N_i), \quad N_i \sim \mathrm{Poi}(np_i)$

Bias:

$$
\begin{aligned}
\mathbb{E}[\hat{S} - S] &= \sum \mathbb{E}[f(N_i)] - \mathbf{1}_{\{p_i > 0\}} \\
&\overset{f(0)=0}{=} \sum \mathbb{E}[(f(N_i) - 1)]\mathbf{1}_{\{p_i > 0\}} \\
&= \sum \mathbb{E}[(f(N_i) - 1)]\mathbf{1}_{\{p_i > 1/k\}} \\
&= \sum \mathbb{E}[(f(N_i) - 1)\mathbf{1}_{\{N_i \leq L\}}]\mathbf{1}_{\{p_i > 1/k\}} \\
&\overset{whp}{=} \sum \underbrace{\mathbb{E}[(f(N_i) - 1)\mathbf{1}_{\{N_i \leq L\}}]} \mathbf{1}_{\{L/2n > p_i > 1/k\}}
\end{aligned}
$$

Observe: $g(\lambda) \triangleq \mathbb{E}[(f(N) - 1)\mathbf{1}_{\{N \leq L\}}] = e^{-\lambda} \times$ poly of deg L