Learning PBD Powers

Kontonis Vasilis

November 13, 2017

Corelab, NTUA

Contents

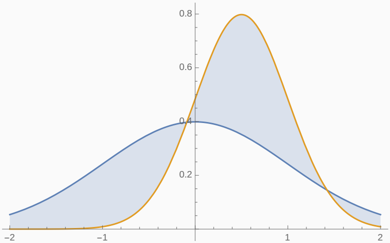
Introduction

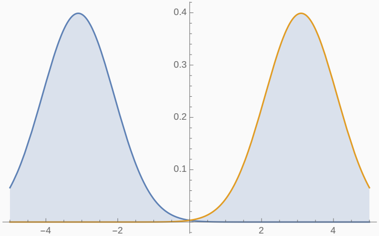
Distribution Learning

- Draw samples from an unknown distribution D.
- Output an approximation of the density function of D.

Distances of Distributions

$$d_{\mathrm{tv}}\left(Q, P\right) = \frac{1}{2}\int |p(x) - q(x)|\,\mathrm{d}x = \sup_{A\in\mathcal{A}} |P(A) - Q(A)|.$$





Dvoretzky Kiefer Wolfowitz Inequality

Kolmogorov Distance

$$d_{\mathrm{kol}}\left(Q, P\right) = \sup_{x \in \mathbb{R}} \left|F_Q(x) - F_P(x)\right|$$

Empirical CDF and DKW Inequality

$$\hat{F}_N(x) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{[X_i \leqslant x]}, \quad x \in \mathbb{R}$$

For every $\varepsilon > 0$

$$\mathbb{P}\left[d_{\mathrm{kol}}\left(F_P, \hat{F}_N\right) > \varepsilon\right] \leqslant 2\mathrm{e}^{-2N\varepsilon^2}$$

- With $N = O\left(1/\varepsilon^2\right)$ samples we approximate the unknown CDF.
- **Question:** Is there anything left to do?

PBDs

Poisson Binomial Distributions

- X_i's are 0/1 Bernoulli with $\mathbf{E}[X_i] = p_i$.
- $X = \sum_{i=1}^{n} X_i$ is a n-PBD with probability vector $\boldsymbol{p} = (p_1, \dots p_n)$.
- $\mathbf{E}[X] = \sum p_i$, $\mathbf{V}[X] = \sum p_i(1 - p_i)$.
- If p_i's are small then X is close in TVD to a $\mathrm{Pois}(\sum p_i)$.
- If $\mathbf{V}[X]$ is large, then X is close in TVD to a (Discretized) Normal.

Learning PBDs

Simply running Birge's Algorithm [1] is *not* good enough, $O\left(\log n/\varepsilon^3\right)$ samples are needed.

- Learn Sparse:
  - Truncate the support: Draw $O\left(1/\varepsilon^2\right)$ samples, sort them and let a, b be such that $X(a \leqslant i \leqslant b) = 1 - \varepsilon$.
  - If $b - a > 1/\varepsilon^3$ then output fail, else run Birge's unimodal algorithm on $X_{[a,b]}$.
- Learn Heavy:
  - Estimate the variance $\hat{\sigma}^2$ and the mean $\hat{\mu}$ of X using $O\left(1/\varepsilon^2\right)$ samples.
  - The discretized normal $\mathrm{DN}\left(\hat{\mu}, \hat{\sigma}^2\right)$ is ε-close, so just output $\hat{\mu}$ and $\hat{\sigma}^2$.
- Choose between Sparse and Heavy.

Tough Curriculum

- set of m items, e.g. set of students.
- a set of n items, e.g. set of courses.
- Each students has passed course i with probability p_i independently from other students.

Question: What is the distribution of the number of different courses that a set of k students will have passed?
Answer: For course i not to be in the set we need to exclude from all students. This happens with probability $(1 - p_i)^k$. Therefore i is included in the union of courses with probability $1 - (1 - p_i)^k$.

PBD Powers

- Let P be a n-PBD defined by $\boldsymbol{p}$.
- P^i is the i-th PBD power of P defined by $\boldsymbol{p}^i$.

Question: Given samples from a subset of the powers of P can we learn the other powers? Can we do better than learning each one of them separately?

Binomial Powers

Approximating Binomials

Binomial TVD [3]

We want to approximate all the distributions $B(n, p^i), i \in \mathbb{N}$.

$$|p - q| \leqslant \varepsilon \sqrt{\frac{p(1-p)}{n}} = \mathrm{err}\,(n, p, \varepsilon) \implies d_{\mathrm{tv}}\,(B(n,p), B(n,q)) \leqslant \varepsilon$$

Approximating p

- Estimator:
$$\hat{p} = \frac{\sum_{i=1}^{N} X_i}{Nn}$$
- Sample Complexity:
Choosing $N = O\left(\ln(1/\delta)/\varepsilon^2\right)$ from Chernoff's bound we have:
$$\mathbb{P}\left[|\hat{p} - p| > \mathrm{err}\,(n, p, \varepsilon)\right] < \delta.$$

Real Powers of p

Assume that $p \approx 1 - \frac{1}{n}$ or $p \approx \frac{1}{n}$.

$$p = 0.\underbrace{99\ldots9}_{\#\log n} \quad \underbrace{458382}_{\text{“constant” part}}$$

$$p = 0.\underbrace{00\ldots0}_{\#\log n} \quad \underbrace{235711}_{\text{“constant” part}}$$

- Sampling from the first power reveals the first part of p, since $\sqrt{p(1-p)/n} \approx 1/n$.
- Is this good enough to approximate all binomial powers ?
- $0.9995^{1000} \approx 0.6064, 0.9997^{1000} \approx 0.7407$

A blast from the past

$$|x - y| = \frac{|x^{2^i} - y^{2^i}|}{\prod_{j=0}^{i-1}(x^{2^j} + y^{2^j})}$$

Finding the Sweet Spot

By Mean Value Theorem applied to mapping $x \mapsto x^l$ we obtain

$$p^l - \hat{q}_1^l \leqslant l p^{l-1}(p - \hat{q}_1), l \in (1, +\infty)$$

and

$$\hat{q}_2^l - p^l \leqslant l p^{l-1}(\hat{q}_2 - p),\ \ l \in (0, 1)$$

We need to find a function $u(p)$ such that for all $l > 0$:

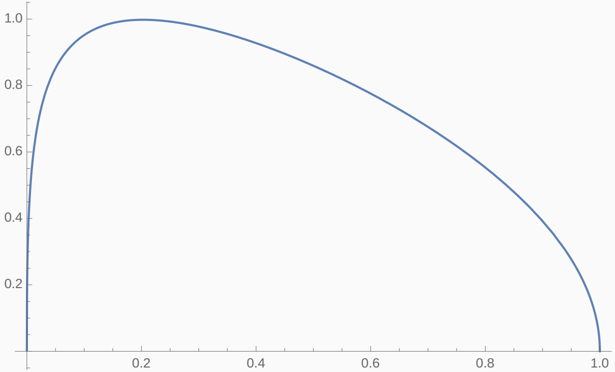
$$\begin{aligned} u(p) l p^{l-1} \operatorname{err}(n, p, \varepsilon) &\leqslant \operatorname{err}\left(n, p^l, \varepsilon\right) \qquad (1) \\ u(p) l p^{l-1} \sqrt{\frac{p(1-p)}{n}} &\leqslant \sqrt{\frac{p^l(1-p^l)}{n}} \\ u^2(p) &\leqslant \frac{p}{1-p} \frac{p^{-l} - 1}{l^2} \end{aligned}$$

Finding the Sweet Spot

$f(l) = \frac{p^{-l}-1}{l^2}$ is convex attaining its minimum at $\bar{l} = -\frac{C}{\ln p}$, $f(\bar{l}) = C \ln^2(1/p)$.

Now we can choose:

$$u(p) = D\,\sqrt{\frac{p}{1-p}}\,\ln(1/p),\ D \approx 1.24$$



Getting in Range

Magic Power: $a = -\frac{1}{\ln p}$.
Question: Can we guess the "magic" power using samples from the first one for all values of p?
Answer: No, if p is very close to 1 then we cannot hope to learn the number of 9's in its decimal representation.

We approximate a with $\hat{a} = -\frac{1}{\ln \hat{p}}$.

- $p^a = \hat{p}^{\hat{a}} = 1/\mathrm{e}$.
- If $|p - \hat{p}| \leqslant \mathrm{err}\,(n, p, \varepsilon)$ then $\frac{1}{\mathrm{e}^2} \leqslant p^{\hat{a}} \leqslant \frac{1}{\mathrm{e}^{3/2}}$.
- Works for $p \in [\varepsilon^2/n, 1 - \varepsilon^2/n]$.

Question: What if p is closer to 1 or 0?
Answer: For $p \in [\varepsilon^2/n^d, 1 - \varepsilon^2/n^d]$ we need $O(\log(d)/\varepsilon^2)$ samples.

Learning Binomial Powers

Algorithm 1 Binomial Powers

Input : $O(\ln(1/\delta)^2/\varepsilon^2)$ samples from the powers of $B(n, p)$.
Output : $\hat{a}$, $\hat{q}_1, \hat{q}_2$.

1: Draw $O(\ln(1/\delta)/\varepsilon^2)$ samples from $B(n, p)$ to obtain the approximation $\hat{p}$.
2: Let $\hat{a} \leftarrow -1/\ln(\hat{p})$.
3: Draw $O\left(\ln(1/\delta)^2/\left(\varepsilon^2\psi(p^{\hat{a}})^2\right)\right)$ samples from $B(n, p^{\hat{a}})$ to get estimations $\hat{q}_1$, $\hat{q}_2$ of p, $\hat{q}_1 \leqslant p \leqslant \hat{q}_2$,
4: **return** $\hat{a}$, $\hat{q}_1$, $\hat{q}_2$

Question: How do we obtain $\hat{q}_1, \hat{q}_2$?

Learning the Parameters of a PBD

Learning the Powers vs Learning the Parameters

Question: PBD powers $\Leftrightarrow$ Parameter Estimation ?

- Assuming that all p_i's are well separated we can learn them by sampling from the powers of a PBD.
- Is there a PBDs where learning its **powers** is easy but learning its **parameters** is hard?

Learning the Parameters

- $P(x) = \prod(x - p_i) = x^n + c_{n-1}x^{n-1} + \ldots + c_0$.
- $\mu_j = \mathbf{E}P_j = \sum p_i^j$.

Newton Identities

$$\begin{pmatrix} 1 & & & & \\ \mu_1 & 2 & & & \\ \mu_2 & \mu_1 & 3 & & \\ \vdots & \vdots & \ddots & \ddots & \\ \mu_{n-1} & \mu_{n-2} & \ldots & \mu_1 & n \end{pmatrix} \begin{pmatrix} c_{n-1} \\ c_{n-2} \\ c_{n-3} \\ \vdots \\ c_0 \end{pmatrix} = \begin{pmatrix} -\mu_1 \\ -\mu_2 \\ -\mu_3 \\ \vdots \\ -\mu_n \end{pmatrix} \Leftrightarrow \boldsymbol{Ac} = \boldsymbol{b}$$

We know the μ_i's only approximately from sampling the PBD powers.

Question: How to measure the impact of noise to the solution of the system?

$$\|\boldsymbol{c} - \hat{\boldsymbol{c}}\|_\infty \leqslant u\ O\left(n^{3/2}2^n\right)$$

Finding the Roots

Pan's Algorithm

- $P(x) = \sum_{i=0}^{n} c_i x^i = c_n \prod_{i=1}^{n} (x - p_i)$, $c_n \neq 0$.
- $|p_j| \leqslant 1$ for all j.
- Computes roots such that $|\hat{p}_j - p_j| < \varepsilon$ for $j = 1, \ldots, n$.
- Precision needed:

$$\|\boldsymbol{c} - \hat{\boldsymbol{c}}\|_\infty = 2^{O(-n \max(\log(1/\varepsilon), \log(n))}. \tag{2}$$

- Overall we need $2^{O(n \max(\log(1/\varepsilon), \log(n)))}$ samples.

Le Cam's Inequality

Question: How do we show a sampling complexity Lower Bound ?

Hypothesis Testing

- We know that samples can come from either P_1 or P_2.
- How many samples do we need in order to decide whether they come from P or Q?
- Ψ is a testing function, $\Psi : \mathcal{X} \to \{1, 2\}$
- Let V be the random variable of the choice of the unknown distribution.

Le Cam's Inequality

$$\inf_{\Psi} \mathbb{P}\left[\Psi(X^N) \neq V\right] = 1 - d_{\mathrm{tv}}\left(P_1^N, P_2^N\right)$$

Minimax Risk, Sample Complexity

- $\mathfrak{P}$ is a family of distributions.
- $P \in \mathfrak{P}$ is a distribution.
- Θ is the space of the parameter we want to estimate.
- $\theta : \mathfrak{P} \to \Theta$, $\theta(P)$ is the parameter of P we want to estimate.
- $\hat{\theta} : \mathcal{X}^N \to \Theta$ is the estimator.
- $X^N = (X_1, \ldots, X_N) \sim P^N$ is the sample vector, N is the number of samples.
- ρ is a semimetric on Θ.

$$\mathfrak{M}_N\left(\theta(\mathfrak{P}), \rho\right) \coloneqq \inf_{\hat{\theta}} \sup_{P \in \mathfrak{P}} \mathbf{E}_{P^N}\left[\rho\left(\hat{\theta}(X^N),\ \theta(P)\right)\right].$$

$$n(\varepsilon, \theta(\mathfrak{P})) = \inf\{N : \mathfrak{M}_N(\theta(\mathfrak{P}), \rho) \leqslant \varepsilon\}$$

Generalization of Minimax Risk

- $\mathfrak{P}$ is a family of *sequences* of distributions.
- $\mathcal{P} \in \mathfrak{P}$ is a *sequence* of distributions.
- Θ is the space of the parameter we want to estimate.
- $\theta : \mathfrak{P} \to \Theta$, $\theta(P)$ is the parameter of P we want to estimate.
- $\hat{\theta} : \mathcal{X}^m \to \Theta$ is the estimator.
- $X^m = (X_1, \ldots, X_{m_1} \ldots, X_{m_k}) \sim P^m$ is the sample vector, N is the number of samples.
- ρ is a semimetric on Θ.

Definition

$$\mathfrak{M}_N\left(\theta(\mathfrak{P}), \rho\right) \coloneqq \inf_{\hat{\theta}} \inf_{|m|=N} \sup_{\mathcal{P}\in\mathfrak{P}} \mathbf{E}_{P^m}\left[\rho\left(\hat{\theta}(X^m),\ \theta(\mathcal{P})\right)\right]. \qquad (3)$$

From Estimation to Testing

Canonical Hypothesis Testing

- "nature" chooses V uniformly from $\mathcal{V}$.
- Conditioned on $V = v$, we draw the sample X^m from the N-fold product distribution P_v^m.

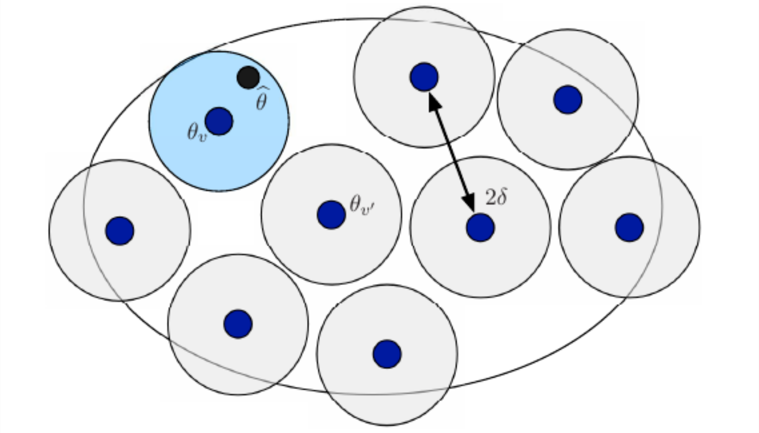
Given X^m our goal is to determine V.

Lower Bound
Let $\mathfrak{F}_{\mathcal{V}} \subseteq \mathfrak{P}$ be a family of sequences of distributions indexed by $v \in \mathcal{V}$ such that $\rho\left(\theta(\mathcal{P}_v, \mathcal{P}_u)\right) \geqslant 2\delta$ for all $\mathcal{P}_v$, $\mathcal{P}_u \in \mathfrak{F}_{\mathcal{V}}$, where, $v \neq u \in \mathcal{V}$ and $\delta > 0$. Then

$$\mathfrak{M}_N\left(\theta(\mathfrak{P}), \rho\right) \geqslant \delta \inf_{m=|N|} \inf_{\Psi} \nu^m\left(\Psi(X^m) \neq V\right).$$

Testing Function:

$\Psi(X^m) \coloneqq \operatorname{argmin}_{v \in \mathcal{V}}\{\rho(\hat{\theta}, \theta_v)\}$, $\rho(\hat{\theta}, \theta_v) \leqslant \delta \Leftrightarrow \Psi(\hat{\theta}) = v$.



Le Cam's Method

Main Idea: Find **two sequences** of distributions that are close in **Total Variation** but their **parameters** are far.

- $\mathcal{P}, \mathcal{Q} \in \mathfrak{P}$ and $\delta > 0$.
- $\rho\left(\theta(\mathcal{P}), \theta(\mathcal{Q})\right) \geqslant 2\delta$ then

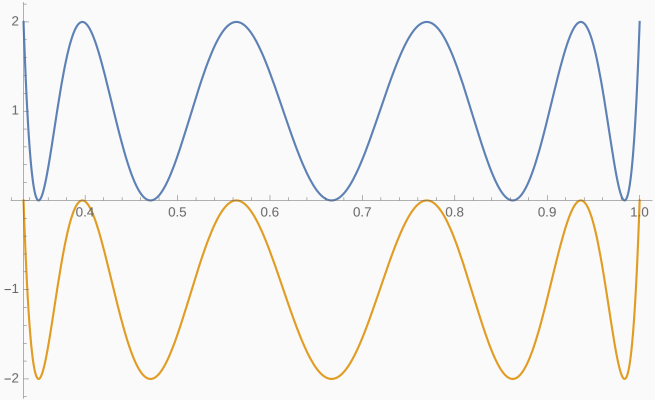
After N observations (samples) the minimax risk has lower bound

$$\mathfrak{M}_N\left(\theta(\mathfrak{P}), \rho\right) \geqslant \frac{\delta}{2}\left(1 - \sqrt{2}\sqrt{1 - \left(1 - d_{\mathrm{tv}}\left(\mathcal{P}, \mathcal{Q}\right)\right)^N}\right).$$

The Lower Bound

Construction of [2]:

- $n = \Theta(\log(N/\varepsilon))$
- $p_j \coloneqq (1 + \cos(\frac{2\pi j}{n}))/8$, $q_j \coloneqq (1 + \cos(\frac{2\pi j + \pi}{n}))/8$, $j \in [n]$.
- $j = n/4 + O(1)$, we have that $|p_i - q_j| = \Omega(1/\log(N/\varepsilon))$



The Lower Bound

- p_i's are the roots of $T_n(8x-1)-1$.
- q_j's are the roots of $T_n(8x-1)+1$.
- for all $l \in \{1, 2, \ldots, n-1\}$, $\sum_{i=1}^{n} p_i^l = \sum_{i=1}^{n} q_i^l$
- for $l \geqslant n$, $3^l(\sum_{i=1}^{n}(p_i^l - q_i^l)) \leqslant n(3/4)^n = \log(N/\varepsilon)(3/4)^{\log(N/\varepsilon)}$.

All PBD powers of $\boldsymbol{p}, \boldsymbol{q}$ are very close in TVD.

$$d_{\mathrm{tv}}\left(P_i, Q_i\right) \leqslant c/N$$

. Therefore the number of samples samples N should be $\Omega(2^{1/\varepsilon})$.

Questions?

References i

L. Birgé.
Estimating a Density under Order Restrictions: Nonasymptotic Minimax Risk.
The Annals of Statistics, 15(3):995–1012, Sept. 1987.

I. Diakonikolas, D. Kane, and A.Stewart.
Properly learning poisson binomial distributions in almost polynomial time.
In *Proceedings of the 29th Conference on Learning Theory, (COLT'16)*, pages 850–878, 2016.

B. Roos.
Improvements in the Poisson approximation of mixed Poisson distributions.
Journal of Statistical Planning and Inference, 113:467–483, 2003.