Poisson's Binomial Distribution (PBD)

$$PBD(p_1 \cdots p_n) : X = \sum_{i=1}^{n} X_i$$

X_i independent, $\{0,1\}$ r.v.'s

$E[X_i] = p_i$ not necessarily identically distributed

examples 1) all p_i's $= 1/2$ $X \sim$ Binomial distribution

2) $p_1 = 1/2$ $p_2 = 1$ $p_3 = p_4 = \cdots = p_n = 0$

$\Pr[X=0] = 0$

$\Pr[X=1] = 1/2$

$\Pr[X=2] = 1/2$

$\Pr[X=3,4,\ldots n] = 0$

$X \sim 1 + \$$

PBD vs Poisson$\left(\sum_{i=1}^{n} p_i\right)$: $\leq 2 \sum_{i=1}^{n} p_i^2$ [LeCam] (1)

$\leq 2 \sum \frac{p_i^2}{\sum p_i}$ (2)

so Poisson approximation is pretty good, but not arbitrarily good (ie. you can't say you want ϵ accuracy & get that close)

Translated Poisson Distribution:

$$TP(\mu, \sigma^2) : Y = \lfloor \mu - \sigma^2 \rfloor + Z$$

$Z \sim$ Poisson $(\sigma^2 + \underbrace{\{\mu - \sigma^2\}}_{\text{fractional part of } \mu - \sigma^2})$

PBD vs TPD:

Thm $d_{TV}(PBD(p_1 \cdots p_n), TP(\mu, \sigma^2)) \leq \dfrac{\sqrt{\sum p_i^3(1-p_i)} + 2}{\sum p_i(1-p_i)}$

still not there

Structure Thm:

<u>Thm</u> PBD "looks like" (to within ε L_1 error) either:

(i) ($\frac{1}{\varepsilon}$-sparse) support of PBD is almost all (as fctn of ε) on $\underbrace{\text{interval of length}}$ $O(1/\varepsilon^3)$

i.e. all but $O(\frac{1}{\varepsilon^3})$ variables have p_i close to 0 or 1 + can be viewed as "fixed"

so we have PBD on $C(\frac{1}{\varepsilon^3})$ variables that can "move"

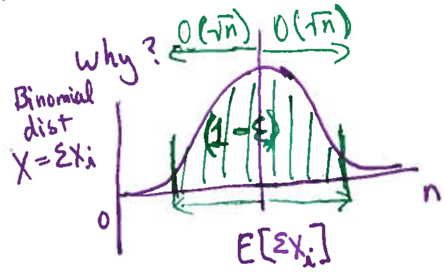
$\Rightarrow$ tiny effective support size, so can <u>learn</u> each probability of elements in support.

(ii) ($\frac{1}{\varepsilon}$-heavy **Binomial**) PBD looks like a binomial on $\underbrace{\text{large}}_{> poly(\frac{1}{\varepsilon})}$ number of iid vars.

Use of structure Thm:

learning: Thm $\Rightarrow$ small cover

testing: Thm $\Rightarrow$ effective support of distribution is $O(n^{1/2})$ $\Rightarrow$ $O(n^{1/4})$ samples needed



maximized in case 2. But Binomial puts almost all of its weight on $\sqrt{n}$ places in the middle.

More detailed structure: for $X = \sum X_i$, let $k \leftarrow O(1/\varepsilon)$

<u>Thm</u> $\exists\ Y_1 \cdots Y_n$ s.t.

1. $\| \sum X_i - \sum Y_i \|_1 \le O(1/k)$

2. One of following holds:

(i) (k-sparse) $\exists\ \ell \le k^3$ s.t. $\forall\ i \le \ell$

so $0 \le \sum Y_i \le k^3$ $\left\{ \begin{array}{l} E[Y_i] \in \{ \frac{1}{k^2}, \frac{2}{k^2}, \ldots, \frac{k^2-1}{k^2} \} \\ \& \ \forall\ i > \ell \quad E[Y_i] \in \{0,1\} \end{array} \right.$

cover size: $(k^3+1)(k^2)^{k^3} \cdot (n+1)$

↑ choices of ℓ ↑ choices of $E[Y_i]$ for $i \le \ell$ ↖ choices # i s.t. $E[Y_i] = 0$

OR

(2) ((n,k)-Binomial form) $\exists\ \ell \in [n]$ & $q \in \{ \frac{1}{n}, \frac{2}{n}, \ldots, \frac{n}{n} \}$ ← actually, $q \ge \frac{1}{k}$, $q \le \frac{k-1}{k}$

s.t. $\forall\ i \le \ell \quad E[Y_i] = q$

& $\forall\ i > \ell \quad E[Y_i] = 0$

⇒ cover size for this part $\le n^2$

also $\underbrace{\ell q}_{E[\sum Y_i]} \ge k^2$ & $\ell q(1-q) \ge k^2 - k - 1$

} note $\ell \ge k^2$; if q small $\sim \frac{1}{k}$ then $\ell \ge k^3$

Cover = union of (1) + (2) covers

Proof Outline

let $k = O(1/\varepsilon)$

Step 1: eliminate vars with expectation in $(0, \frac{1}{k})$ or $(\frac{k-1}{k}, 1)$ w/o much change

for all i s.t. $p_i \in \{0, \frac{1}{k}\}$

take their sum + figure out how many (p_i')'s with prob $\frac{1}{k}$ would have similar sum

ie. $r \cdot \frac{1}{k} \approx \sum p_i$

Set 1^{st} r such i to $\frac{1}{k}$ + rest to 0

use $d_{TV}\left(\sum_{i \text{ small}} X_i, \text{Poisson}\left(\sum_{i \text{ small}} p_i\right)\right) \le \frac{\sum_{i \text{ small}} p_i^2}{\sum_{i \text{ small}} p_i} \le \frac{\frac{1}{k}\sum p_i}{\sum p_i} = \frac{1}{k}$

$$d_{TV}\Big(\text{poisson}\big(\underbrace{\textstyle\sum_{i \text{ small}} p_i}_{\lambda_1}\big), \text{poisson}\big(\underbrace{\textstyle\sum_{i \text{ small}} p_i'}_{\lambda_2}\big)\Big) \le \frac{1}{2}\left(e^{|\lambda_1 - \lambda_2|} - e^{-|\lambda_1 - \lambda_2|}\right)$$

$$= \frac{1}{2}\left(e^{\frac{1}{k}} - e^{-1/k}\right)$$

$$\le \frac{1.5}{k}$$

$d_{TV}\left(\text{poisson}(\sum p_i'), \sum_{i \text{ small}} X_i'\right) \le \frac{1}{k}$

$\Delta \neq \Rightarrow \text{dist} \le \frac{3.5}{k}$

get total $\frac{7}{k}$ dist when do heavy elts.

Step 2:

k-sparse case:

weaker proof – [Use $d_{TV}(\Sigma X_i, \Sigma Y_i) \le \sum_i d_{TV}(X_i, Y_i)$ when X_i, Y_i indep]

if round each p_i to nearest multiple of $\frac{1}{k^4}$

get $d_{TV}(\Sigma X_i, \Sigma X_i') \le k^3 \cdot \frac{1}{k^4} = \frac{1}{k}$

they do something smarter!

idea something like step 1 (& relate to Binomial)

use a different bound on similarity to Binomial

use different grouping – k groups

each contributes $O(\frac{1}{k^2})$ error

total $O(\frac{1}{k})$ error

if not k-sparse:

approx by Binomial distribution

$B(m', q)$

vars fixed to 1

$$m' \equiv \frac{(\Sigma p_i' + t)^2}{(\Sigma p_i'^2 + t)} \qquad q \equiv \frac{l^*}{n}$$

l^* chosen s.t. $\frac{\Sigma p_i' + t}{m'} \in \left[\frac{l^*-1}{n}, \frac{l^*}{n}\right]$

Can show via bound on similarity to translated Poisson Dist that approx is good.

$\square$

Further improvements:

Can weed more out of cover by using

Roos's Thm $\Rightarrow$ if $\Sigma p_i^t = \Sigma q_i^t \quad \forall\, t = 1 \ldots c(\log 1/\varepsilon)$

$\Rightarrow \|p - q\|_1 \le \varepsilon$

$\Leftarrow$ not quite as stated here unless all $p_i, q_i \le 1/2$ (otherwise need to separate i s.t. $p_i, q_i \le 1/2$ from i s.t. $p_i, q_i > 1/2$)