Updated 11/4/2024

Part 3: Practice
L16: Method of Types (1/2)

Javed Aslam, Wolfgang Gatterbauer

cs7840 Foundations and Applications of Information Theory (fa24)

https://northeastern-datalab.github.io/cs7840/fa24/

10/30/2024

Last time	Today	Next time
• Decision Trees • MDL	• Method of Types	• Finish MoT • Applications to statistics – Large Deviation Theory

Review: AEP & Typical Sets

Def: The typical set $A_\epsilon^{(n)}$ with respect to $p(x)$ is the set of all sequences $(x_1 x_2 \ldots x_n) \in \mathcal{X}^n$ such that

$$2^{-n(H(X)+\epsilon)} \le p(x_1, x_2 \ldots x_n) \le 2^{-n(H(X)-\epsilon)}$$

empirical probability of sequence

Upshot:

- $2^{-n(H(X)+\varepsilon)} \leq p(x_1, x_2 \ldots x_n) \leq 2^{-n(H(X)-\varepsilon)}$

a typical sequence has empirical probability $\sim 2^{-nH(X)}$

- $|A_\varepsilon^{(n)}| \leq 2^{n(H(X)+\varepsilon)}$
- $|A_\varepsilon^{(n)}| \geq (1-\varepsilon)\, 2^{n(H(X)-\varepsilon)}$

there are $\sim 2^{nH(X)}$ typical sequences

- $Pr[A_\varepsilon^{(n)}] > 1-\varepsilon$ for n sufficiently large

typical sequences contain almost all probability

AEP, Weak Typicality, Strong Typicality & Method of Types

Example: $\vec{p} = (1/2, 1/4, 1/8, 1/8)$ $\quad H(\vec{p}) = 1\,3/4$

(Weak) Typical set: $A_\varepsilon^{(n)}$ s.t. $2^{-n(H(X)+\varepsilon)} \leq p(x_1 \ldots x_n) \leq 2^{-n(H(X)-\varepsilon)}$

$n=16$: This should be "typical"

$$(8,4,2,2) \Rightarrow \text{Prob} = (1/2)^8 \cdot (1/4)^4 \cdot (1/8)^2 \cdot (1/8)^2$$

$$= \frac{1}{2^8} \cdot \frac{1}{2^8} \cdot \frac{1}{2^6} \cdot 1/2^6 = \frac{1}{2^{28}}$$

$$= \frac{1}{2^{16 \cdot 1\,3/4}}$$

- What about $(4,12,0,0)$?

$$\Pr = (1/2)^4 \cdot (1/4)^{12} = \frac{1}{2^4} \cdot \frac{1}{2^{24}} = 1/2^{28}$$

$\leftarrow$ (weakly) typical, but our intuition is that it is "unusual"

$(4, 12, 0, 0)$ is (weakly) "typical" but not expected

$\forall a \in \mathcal{X}$

Strong typicality: $A_\varepsilon^{*(n)} = \left\{ \vec{x} \in \mathcal{X}^n \;\middle|\; \begin{array}{ll} \left|\frac{1}{n} N(a|\vec{x}) - P(a)\right| < \frac{\varepsilon}{|\mathcal{X}|} & \text{if } P(a) > 0 \\ N(a|\vec{x}) = 0 & \text{if } P(a) = 0 \end{array} \right.$

Sample result: $Q^n(T(P)) \approx 2^{-n D(P \| Q)}$

MoT setup:

- Let $x_1 x_2 \dots x_n$ be a sequence of length n drawn from a dist Q over an alphabet $\mathcal{X} = \{a_1, a_2, \dots a_{|\mathcal{X}|}\}$

Def: The <u>type</u> $P_{\vec{x}}$ of a sequence $\vec{x}$ is the <u>empirical distribution</u> associated w/ $\vec{x}$, i.e. $P_{\vec{x}}(a) = \frac{N(a|\vec{x})}{n} \quad \forall a \in \mathcal{X}$

where $N(a|\vec{x})$ is the # a's in $\vec{x}$.

Def: Let $\mathcal{P}_n$ denote the set of types w/ denominator n

Examples: 6 sided die, $n=5$, $\vec{x} = 13464$

$$P_{\vec{x}} = \left(\frac{1}{5}, \frac{0}{5}, \frac{1}{5}, \frac{2}{5}, \frac{0}{5}, \frac{1}{5}\right)$$

$$\mathcal{P}_5 = \left\{ \left(\frac{0}{5}, \frac{0}{5}, \frac{0}{5}, \frac{0}{5}, \frac{0}{5}, \frac{5}{5}\right), \left(\frac{0}{5}, \frac{0}{5}, \frac{0}{5}, \frac{0}{5}, \frac{1}{5}, \frac{4}{5}\right), \dots, \left(\frac{5}{5}, \frac{0}{5}, \frac{0}{5}, \frac{0}{5}, \frac{0}{5}, \frac{0}{5}\right) \right\}$$

$$|\mathcal{P}_n| \geq \binom{n+|\mathcal{X}|-1}{|\mathcal{X}|-1}$$

balls-in-bins argument

e.g. 6-sided die w/ $n=5$

$\Rightarrow$ we are throwing 5 balls into 6 bins

Thm: $|\mathcal{P}_n| \leq (n+1)^{|\mathcal{X}|}$

Pf: · a type, for each symbol, has a <u>variable numerator</u> and a <u>fixed denominator</u>

· numerator can only take on $n+1$ values $\{0,1,\ldots,n\}$

· $|\mathcal{X}|$ symbols can each take on these (at most) $n+1$ values

$\Rightarrow$ $(n+1)^{|\mathcal{X}|}$ (ignores constraints & dependencies, so upper bound)

Def: For any $P \in \mathcal{P}_n$, the <u>type class</u> of P is the set of sequences w/ type P

$$T(P) = \{\vec{x} \in \mathcal{X}^n : P_{\vec{x}} = P\}$$

Thm: If $X_1 X_2 \ldots X_n$ are drawn iid according to Q,

then the probability of $\vec{x}$ depends only on its type and is

$$Q^n(\vec{x}) = 2^{-n\left(H(P_{\vec{x}}) + D(P_{\vec{x}} \| Q)\right)}$$

Pf: $$Q^n(\vec{x}) = \prod_{i=1}^{n} Q(x_i)$$

$$= \prod_{a \in \mathcal{X}} Q(a)^{N(a|\vec{x})}$$

$$= \prod_{a \in \mathcal{X}} Q(a)^{n \cdot P_{\vec{x}}(a)}$$

$$= \prod_{a \in \mathcal{X}} 2^{n P_{\vec{x}}(a) \lg Q(a)}$$

$$= \prod_{a \in \mathcal{X}} 2^{n\left(P_{\vec{x}}(a) \lg Q(a) - P_{\vec{x}}(a) \lg P_{\vec{x}}(a) + P_{\vec{x}}(a) \lg P_{\vec{x}}(a)\right)}$$

$$= \prod_{a \in \mathcal{X}} 2^{n\left(P_{\vec{x}}(a) \lg Q(a) - P_{\vec{x}}(a) \lg P_{\vec{x}}(a) + P_{\vec{x}}(a) \lg P_{\vec{x}}(a)\right)}$$

$$= 2^{n \sum_{a \in \mathcal{X}} \left(-P_{x}(a) \lg \frac{P_{x}(a)}{Q(a)} + P_{x}(a) \lg P_{x}(a)\right)}$$

$$= 2^{n\left(-D(P_{\vec{x}} \| Q) - H(P_{x})\right)} \quad \checkmark$$

Updated 11/4/2024

Part 3: Practice
L17: Method of Types (2/2)
[Sanov's theorem, Large Deviation Theory]

Javed Aslam, Wolfgang Gatterbauer

cs7840 Foundations and Applications of Information Theory (fa24)

https://northeastern-datalab.github.io/cs7840/fa24/

11/4/2024

Last time

- Method of Types

Today

- Continue MoT
- Sanov's Theorem
- Applications

Next time

- Finish applications of MoT
  - if needed
- Back to Wolfgang

Q1: $Q^n(\vec{x})$? – prob. of seq. in type class

Q2: $|T(P)|$? – size of type class

Q3: $Q^n(T(P))$? – prob. of type class

Q1: $Q^n(\vec{x})$? - probability of sequence in type class

A1: Last time...

$$Q^n(\vec{x}) = 2^{-n\left(H(P_{\vec{x}}) + D(P_{\vec{x}} \| Q)\right)}$$

Corollary: If $\vec{x}$ in type class of Q, then

$$Q^n(\vec{x}) = 2^{-n\left(H(Q) + D(Q \| Q)\right)}$$
$$= 2^{-n H(Q)}$$

$\Rightarrow$ Generalizes AEP results to atypical sequences

Q2: $|T(P)|$? — size of type class

A2: For any type $P \in \mathcal{P}_n$,

$$\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{nH(P)} \le \frac{1}{|\mathcal{P}_n|} 2^{nH(P)} \le |T(P)| \le 2^{nH(P)}$$

Pf: (upper bound)

$$\begin{aligned} 1 &\ge P^n(T(P)) \\ &= \sum_{\vec{x} \in T(P)} P^n(\vec{x}) \\ &= \sum_{\vec{x} \in T(P)} 2^{-nH(P)} \\ &= |T(P)| \cdot 2^{-nH(P)} \end{aligned}$$

Corollary: If $\vec{x}$ in type class of Q, then

$$\begin{aligned} Q^n(\vec{x}) &= 2^{-n(H(Q) + D(Q \| Q))} \\ &= 2^{-nH(Q)} \end{aligned}$$

$$\Rightarrow |T(P)| \le 2^{nH(P)} \checkmark$$

Pf: (lower bound) $\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{nH(P)} \leq \frac{1}{|\mathcal{P}_n|} 2^{nH(P)} \leq |T(P)|$

Claim: $P^n(T(P)) \geq P^n(T(P')) \quad \forall \; P' \in \mathcal{P}_n$

$\Rightarrow$ $T(P)$ is the most probable type class under P;
unsurprising but technical result – see text

Then... $1 = \sum_{Q \in \mathcal{P}_n} P^n(T(Q)) \leq \sum_{Q \in \mathcal{P}_n} \max_Q P^n(T(Q)) = \sum_{Q \in \mathcal{P}_n} P^n(T(P))$

$$= |\mathcal{P}_n| \cdot P^n(T(P)) = |\mathcal{P}_n| \cdot \sum_{\vec{x} \in T(P)} P^n(\vec{x})$$

$$= |\mathcal{P}_n| \cdot \sum_{\vec{x} \in T(P)} 2^{-nH(P)} = |\mathcal{P}_n| \cdot |T(P)| \cdot 2^{-nH(P)}$$

$$\Rightarrow |T(P)| \geq \frac{1}{|\mathcal{P}_n|} \cdot 2^{nH(P)} \geq \frac{1}{(n+1)^{|\mathcal{X}|}} \cdot 2^{nH(P)}$$

since $|\mathcal{P}_n| \leq (n+1)^{|\mathcal{X}|}$

Q3: $Q^n(T(P))$ – probability of a type class

A3: $$\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{-nD(P\|Q)} \le \frac{1}{|\mathcal{P}_n|} \cdot 2^{-nD(P\|Q)} \le Q^n(T(P)) \le 2^{-nD(P\|Q)}$$

Pf: Combine Q1 & Q2

Q1: $$Q^n(\vec{x}) = 2^{-n(H(P_{\vec{x}}) + D(P_{\vec{x}}\|Q))}$$

Q2: $$\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{nH(P)} \le \frac{1}{|\mathcal{P}_n|} 2^{nH(P)} \le |T(P)| \le 2^{nH(P)}$$

$\Rightarrow$ Generalizes AEP results to atypical sets

Large Deviation Theory: Bounding the probability of "rare" events

- Let E be a subset of probability mass functions, defining a "rare" event

  e.g. Roll a fair 6-sided die n times – rare event of interest may be having mean die roll being at least 4
  $\Rightarrow$ E would be all distributions over 6 die faces s.t. mean ≥ 4, such as $(0,0,0,1,0,0)$ or $(0,0,\frac{1}{2},0,\frac{1}{2},0)$ ...

- We wish to bound

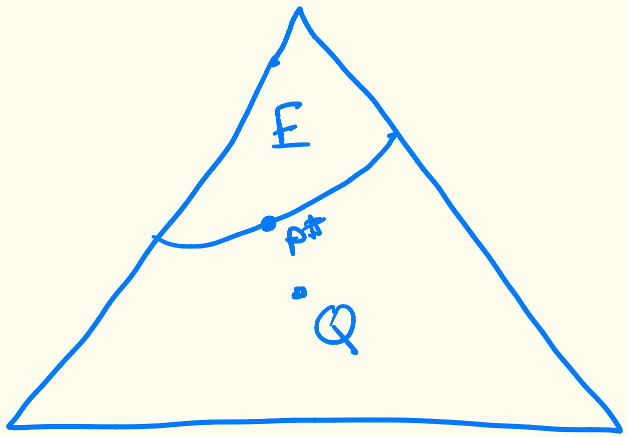
$$Q^n(E) = Q^n(E \cap \mathcal{P}_n) = \sum_{x:\, P_{\vec{x}} \in \underbrace{E \cap \mathcal{P}_n}_{\text{type classes that are possible and of interest}}} Q^n(\vec{x})$$

Sanov's theorem:

- Let $X_1, X_2, \ldots, X_n$ be drawn i.i.d. from Q.
- Let $E \subseteq \mathcal{P}$ be a set of probability distributions
  - defining a "rare" event

Then $Q^n(E) = Q^n(E \cap \mathcal{P}_n) \leq |\mathcal{P}_n| \cdot 2^{-n D(P^* \| Q)} \leq (n+1)^{|\mathcal{X}|} \, 2^{-n D(P^* \| Q)}$

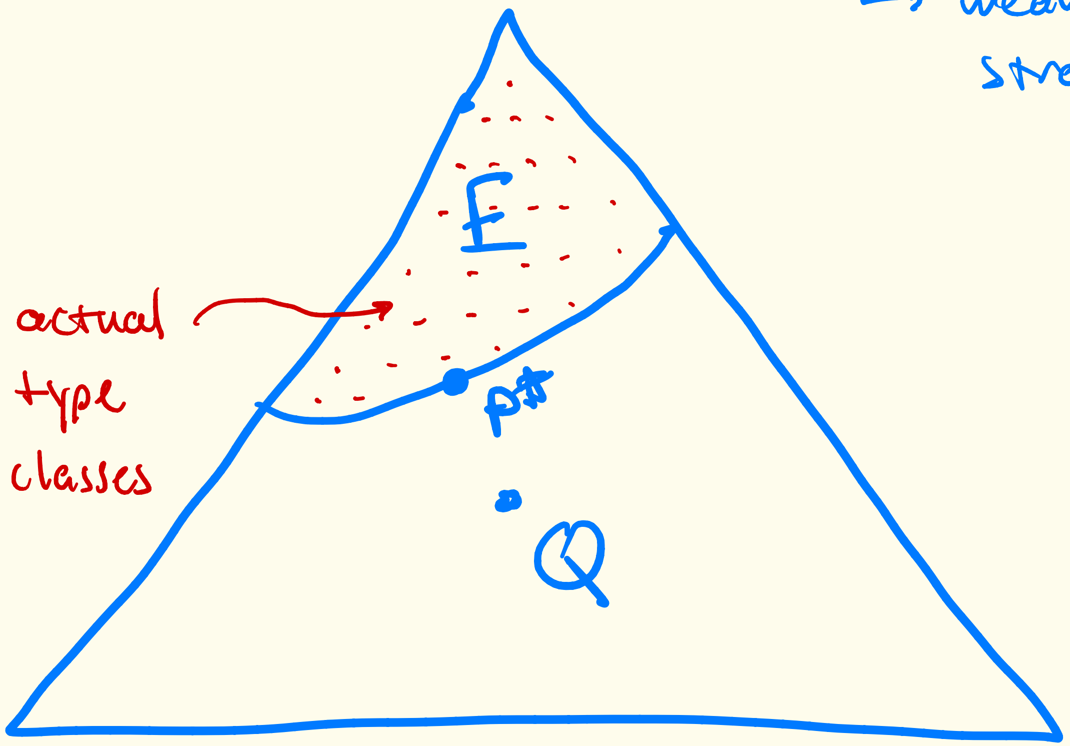
where $P^* = \arg\min_{P \in E} D(P \| Q)$ ← distribution in E closest to Q in KL-divergence



Proof Idea:

- we need to bound probability of every rare event in E
- they are grouped into finite set of type classes (in red)
- prob. of type classes drop off exponentially in KL-divergence from Q
- can bound by closest dist. in E

↳ weak bound that can be strengthened



Pf :

$$Q^n(E) = \sum_{P \in E \cap \mathcal{P}_n} Q^n(T(P))$$

$$\leq \sum_{P \in E \cap \mathcal{P}_n} 2^{-n \cdot D(P \| Q)}$$

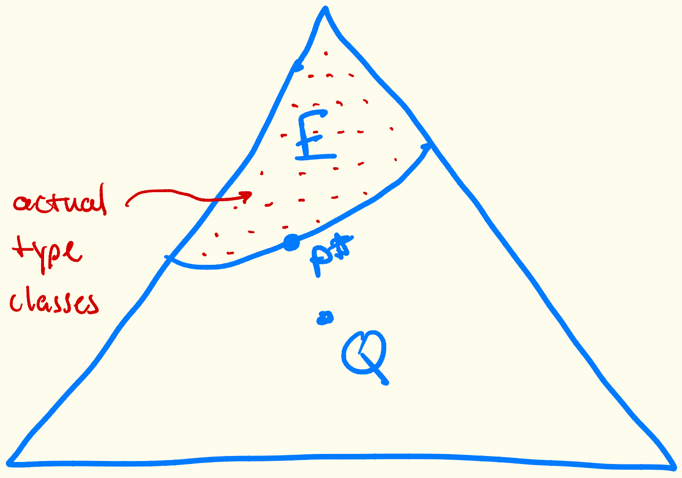
$$\leq \sum_{P \in E \cap \mathcal{P}_n} \max_{P \in E \cap \mathcal{P}_n} 2^{-n \cdot D(P \| Q)}$$

$$= \sum_{P \in E \cap \mathcal{P}_n} 2^{-n \min_{P \in E \cap \mathcal{P}_n} D(P \| Q)}$$

$$\leq \sum_{P \in E \cap \mathcal{P}_n} 2^{-n \min_{P \in E} D(P \| Q)}$$

$$= \sum_{P \in E \cap \mathcal{P}_n} 2^{-n \cdot D(P^* \| Q)}$$

$$\leq |\mathcal{P}_n| \cdot 2^{-n \cdot D(P^* \| Q)} \leq (n+1)^{|\mathcal{X}|} \cdot 2^{-n \cdot D(P^* \| Q)}$$



$$Q^n(T(P)) \leq 2^{-n D(P \| Q)}$$

Note 1: If Q is <u>uniform</u>, then $P^* = \min_{P \in E} D(P \| Q)$ is

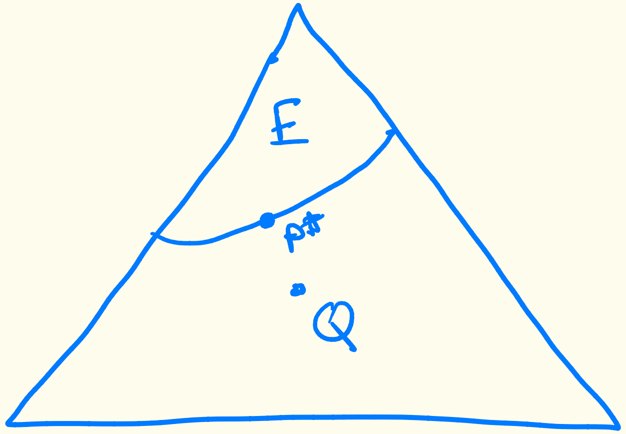
$$D(P\|Q) = \sum_{a \in \mathcal{X}} p(a) \lg \frac{p(a)}{Q(a)}$$

$$= \sum_{a \in \mathcal{X}} p(a) \lg \frac{p(a)}{1/|\mathcal{X}|}$$

$$= \lg|\mathcal{X}| + \sum_{a \in \mathcal{X}} p(a) \lg p(a)$$

$$= \lg|\mathcal{X}| - H(P)$$

So, $\min_{P \in E} D(P\|Q) = \max_{P \in E} H(P)$



$\Rightarrow$ want max. ent. dist. P subject to constraint E

Note 2: (Improved Sanov)

If $E \subseteq \mathcal{P}$ is a convex set of probability distributions,

then

$$Q^n(E) = Q^n(E \cap \mathcal{P}_n) \leq 2^{-n D(P^* \| Q)}$$

where $P^* = \arg\min_{P \in E} D(P \| Q)$

- Don't need the $|\mathcal{P}_n| \leq (n+1)^{|\mathcal{X}|}$ factor
- Convex: $(\forall P_1 \,\&\, P_2 \text{ in } E)(\forall\, 0 \leq \lambda \leq 1)\ \lambda P_1 + (1-\lambda) P_2 \in E$
- Many "rare" events of interest are convex sets

Note 3: Conditional Limit theorem

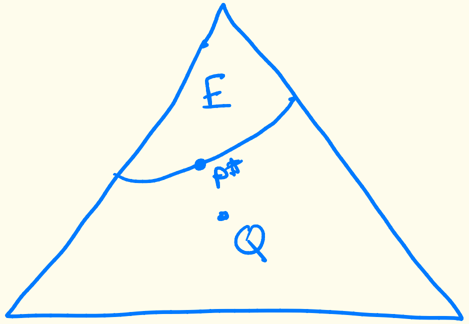
If $E \subseteq \mathcal{P}$ is a convex set of probability distributions, $x_1, x_2 \dots x_n$ are drawn i.i.d according to Q, and $P_{\vec{x}} \in E$

then

$$\Pr(x_1 = a \mid P_{\vec{x}} \in E) \to P^*(a)$$

in probability as $n \to \infty$ where

$$P^* = \arg\min_{P \in E} D(P \| Q)$$



In other words, the conditional distribution of X_1 given that $P_{\vec{x}} \in E$, is close to P^* for large n.

$\Rightarrow$ It will look like your sequence is drawn according to P^*

Example: Roll fair 6-sided die 100 times, what is probability that average die roll ≥ 4?

Solution:

- E is convex (why?)

$$\Rightarrow Q^{100}(E) \leq 2^{-100 \cdot D(P^* \| Q)}$$

where $P^* = \arg\min_{P \in E} D(P \| Q)$

- Q is uniform $\Rightarrow P^* = \arg\max_{P \in E} H(P)$

- Need max. ent. dist. $\Rightarrow$ Lagrange multipliers

$$J(P, \lambda_1, \lambda_2) = H(P) + \lambda_1 \underbrace{\left(\sum_i i \cdot p_i - 4 \right)}_{\text{mean constraint}} + \lambda_2 \underbrace{\left(\sum_i p_i - 1 \right)}_{\text{prob. dist. constraint}}$$

Solve ... $P^* = (0.1031, 0.1227, 0.1461, 0.1740, 0.2072, 0.2468)$

Compare to: $Q = (0.1666, 0.1666, 0.1666, 0.1666, 0.1666, 0.1666)$

$$D(P^* \| Q) = 0.0624$$

So, $Q^{100}(E) \leq 2^{-100 \cdot 0.0624}$

$$= 2^{-6.24}$$

$$= 0.0132$$

about 1.32%

$n = 1000$ ... $2^{-1000 \cdot 0.0624} = 2^{-60.24} = 7.344 \times 10^{-19}$!

Also, by Conditional Limit Theorem, you are likely to have seen about P^* fractions of 1s, 2s, 3s, ...

Large Deviation Theory: Bounds on tail ("rare") events

Bounds tend to be ...

1. general but weak – Markov, Chebyshev
2. strong but specific – Chernoff & Hoeffding for binomial distribution
3. Information theory – general and strong (Sanov) w/ intuition (Conditional Limit Theorem)
4. Central Limit Theorem & Normal approximation
   - reasonably strong
   - reasonably general
   - approximate: not a true bound
   - no intuition like Conditional Limit Thm.

can't handle a dice example like, "mean = 4 and no 2s".

Sanov is more general

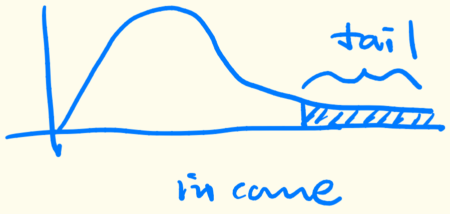
① Markov's Inequality

Let X be any non-negative r.v.

then $\Pr\{X \geq a\} \leq \frac{E\{X\}}{a}$

Pf: $$E\{X\} = \int_0^\infty x\, p(x)\, dx$$
$$\geq \int_a^\infty x\, p(x)\, dx$$
$$\geq a \cdot \int_a^\infty p(x)\, dx$$
$$= a \cdot \Pr\{X \geq a\}$$
$$\Longleftrightarrow \Pr\{X \geq a\} \leq \frac{E\{X\}}{a}$$

Distribution of Income



E.g. If mean household income is \$60K, what is bound on fraction of households w/ income at least \$240k?

$$\Pr\{X \geq \$240k\} \leq \frac{\$60k}{\$240k} = 1/4$$

② Chebyshev's Inequality

Let X be r.v. with mean μ & variance σ^2

Then $\Pr[|X-\mu| \geq \delta] \leq \frac{\sigma^2}{\delta^2}$

or, letting $\delta = k \cdot \sigma$

$$\Pr\{|X-\mu| \geq k\sigma\} \leq \frac{1}{k^2}$$

Pf: Let $Y = (X-\mu)^2$ — Y is a non-neg. r.v. so Markov applies

$$\Pr\{Y \geq \delta^2\} \leq \frac{E\{Y\}}{\delta^2} = \frac{\sigma^2}{\delta^2}$$

since $E\{(X-\mu)^2\} = \sigma^2$

$$\Rightarrow \Pr\{|X-\mu| \geq \delta\} \leq \frac{\sigma^2}{\delta^2}$$

since $Y \geq \delta^2 \Leftrightarrow |X-\mu| \geq \delta$

E.g. If $\mu = \$60k$ then $\Pr\{X \geq \$240k\} = \Pr[|X - \$60k| \geq 180k]$

$\sigma = \$30k$ $\leq (30)^2/(180)^2 = 0.02\overline{77}$

③ Chernoff / Hoeffding for Binomial Distributions (coin flips)

Common forms:

ⓐ Relative $0 \le \beta \le 1$

$$LE(p, m, (1-\beta)mp) \le e^{-\beta^2 mp/2}$$

$$GE(p, m, (1+\beta)mp) \le e^{-\beta^2 mp/3}$$

ⓑ Additive

$$LE(p, m, m(p-\alpha)) \le e^{-2\alpha^2 m}$$

$$GE(p, m, m(p+\alpha)) \le e^{-2\alpha^2 m}$$

Note: For small p, relative error bounds are better, while for large p, additive error bounds are better. Threshold is $p = 1/4$ for LE bounds & $p = 1/6$ for GE bounds

④ Normal Approximation via Central Limit Theorem

CLT: $X_1 \ldots X_N$ sequence of i.i.d. r.v. w/ μ & σ^2, then

$$X = \sum_{i=1}^{N} X_i \xrightarrow{d} N(\mu N, \sigma^2 N)$$

or $$\bar{X} = \frac{1}{N} \sum_{i=1}^{N} X_i \xrightarrow{d} N\left(\mu, \frac{\sigma^2}{N}\right)$$

$\Rightarrow$ sums or averages will eventually look like a normal (Gaussian) distribution

$\Rightarrow$ use tail of normal distribution to approximate tail of actual event

$\Rightarrow$ it's an <u>approximation</u> and tail of normal does not have a closed form

Example: · Flip a fair coin 1000 times

· what is likelihood of seeing at least 700 heads?

· should be <u>very</u> unlikely; but just how unlikely?

· Since it's so unlikely, we will look at ln prob.

① Markov $\Pr\{X \geq a\} \leq \frac{E[X]}{a}$

$$\Pr\{X \geq 700\} \leq \frac{500}{700} = \frac{5}{7}$$

$$\boxed{\ln \text{prob} = -0.336}$$

(so prob $= e^{-0.336}$)

② Chebyshev $\Pr[|X-\mu| \geq \delta] \leq \frac{\sigma^2}{\delta^2}$

$\mu = np = 500$

$\sigma^2 = np(1-p) = 1000/4 = 250$

$$\Pr[|X-500| \geq 200] \leq \frac{250}{(200)^2}$$

$$= 0.00625$$

$$\boxed{\ln \text{prob} = -5.075}$$

③ Relative (Chernoff)

$$GE(m, p, mp(1+\beta)) \le e^{-mp\beta^2/3}$$

$$\Rightarrow GE(1000, 1/2, 1000 \cdot 1/2 (1+0.4)) \le e^{-1000 \cdot 1/2 (0.4)^2/3}$$

$$\boxed{\begin{aligned} \ln \text{prob} &= -1000 \cdot \tfrac{1}{2} \cdot (0.4)^2/3 \\ &= -26.6\overline{6} \end{aligned}}$$

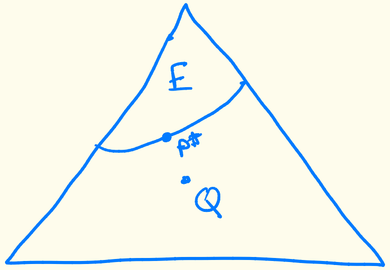
④ Additive (Hoeffding)

$$GE(m, p, m(p+\alpha)) \le e^{-2m\alpha^2}$$

$$GE(1000, 1/2, 1000(1/2+0.2)) \le e^{-2 \cdot 1000 \cdot (0.2)^2}$$

$$\boxed{\ln \text{prob} = -2 \cdot 1000 \cdot (0.2)^2 = -80}$$

⑤ Sanov : E is convex (why?)



$$\Rightarrow \quad 2^{-n\, D(p^* \| Q)}$$

$p^* = (0.7, 0.3)$ (why?)

$Q = (1/2, 1/2)$

$D(p^* \| Q) = 0.119$

$2^{-1000 \cdot 0.119} = 2^{-119}$

$$\boxed{\ln \text{prob} = -119 \cdot \ln 2 = -82.485}$$

⑥ Normal via CLT ... $\boxed{\ln \text{prob} = -83.46}$ (but an <u>approximation</u>)