LECTURE 5
METHOD OF TYPES AND APPLICATIONS

Information Theory

Thomas Debris-Alazard

Inria, École Polytechnique

Relation between information theory and probability theory

$\longrightarrow$ The method of types

Powerful technique:

- Probability of rare events $\big($large deviations$\big)$
- Universal source coding
- Testing hypothesis
- etc. . .

Thought experiment:

Given i.i.d $X_i \in \mathcal{X}$ according to X, we want to estimate the $\mathbb{P}(X = a)$'s

A natural approach, observe a sequence **x** of length n and compute the empirical distribution:

$$\mathbb{P}(X = a) \approx \frac{\sharp\{i \in [1, n] : \ x_i = a\}}{n}$$

$\longrightarrow$ By the weak law of large number $\left(\text{AEP}\right)$ we know that our estimation will tend to the right one when n large enough

- How large should be n?
- What is the exact probability to make mistakes?

AEP has been a powerful tool but it does not help for rare events!

COURSE OUTLINE

1. Method of Types
2. Alternative Law of Large Numbers
3. Universal Coding
4. Large Deviation Theory
5. Chernoff's Bound
6. Sanov's Theorem
7. Some Applications of Sanov's Theorem

METHOD OF TYPES

AEP VERSUS METHOD OF TYPES

- AEP: what are the typical sequences? Their probability to appear is given by the entropy!

$\longrightarrow$ Crude tool in many situations!

- Method of types: split sequences according to their empirical distribution $\Big($the type$\Big)$

$\longrightarrow$ The event of interest is partitioned into its intersections with the type classes. But,

1. The number of types is polynomial
2. There are an exponential number of sequences
3. All sequences in a type are equiprobable $\Big($memoryless source$\Big)$

The event probability has the same exponential asymptotics as the largest one among the probabilities of the above intersections!

NOTATION

We will consider random variables $\mathsf{X}_1, \ldots, \mathsf{X}_n$ from an alphabet $\mathcal{X} = \{a_1, \ldots, a_{\sharp\mathcal{X}}\}$

Any vector $\mathbf{x}$ $\big($bold letter$\big)$ denotes a sequence $x_1, \ldots, x_n \in \mathcal{X}$

If $\mathsf{X}_1, \ldots, \mathsf{X}_n$ are i.i.d random variables distributed according to Q, *i.e.* $\mathbb{P}(\mathsf{X}_i = a) = Q(a)$, then

$$\forall \mathbf{x} \in \mathcal{X}^n, \quad Q^n(\mathbf{x}) = \prod_{i=1}^{n} \mathbb{P}(\mathsf{X}_i = x_i) = \prod_{i=1}^{n} Q(x_i)$$

Furthermore, given some event $\mathcal{E}$,

$$\mathbb{P}_{\mathbf{x} \leftarrow Q^n}(\mathbf{x} \in \mathcal{E}) \stackrel{\text{def}}{=} \sum_{\mathbf{x} \in \mathcal{E}} Q^n(\mathbf{x}) = \sum_{\mathbf{x} \in \mathcal{E}} \prod_{i=1}^{n} Q(x_i)$$

Type:

Given $\mathbf{x} \in \mathcal{X}^n$ for some $n > 0$, a type $P_{\mathbf{x}}^{\text{emp}}$ is a probability distribution over $\mathcal{X}$ defined as:

$$\forall a \in \mathcal{X}, \quad P_{\mathbf{x}}^{\text{emp}}(a) \stackrel{\text{def}}{=} \frac{\sharp\{i \in [1,n] : \; x_i = a\}}{n}$$

$\Big($A type $P_{\mathbf{x}}^{\text{emp}}$ is also called *empirical distribution of* $\mathbf{x}\Big)$

An example:

$\mathcal{X} = \{0,1\}$ and $\mathbf{x} = (1,1,1,0,1,0,0,1) \in \{0,1\}^8$,

$$P_{\mathbf{x}}^{\text{emp}}(0) = \frac{3}{8} \quad \text{and} \quad P_{\mathbf{x}}^{\text{emp}}(1) = \frac{5}{8}$$

$\longrightarrow$ Be careful: a type $P_{\mathbf{x}}^{\text{emp}}$ is defined according to the sequence $\mathbf{x}$

Type of fixed length:

Given $n > 0$,

$$\mathcal{P}_n \stackrel{\text{def}}{=} \{P_{\mathbf{x}}^{\text{emp}} : \mathbf{x} \in \mathcal{X}^n\}$$

The binary case:

Given $\mathcal{X} = \{0, 1\}$, types of length n:

$$\mathcal{P}_n = \left\{ \left(\underbrace{\frac{0}{n}}_{p(0)}, \underbrace{\frac{1}{n}}_{p(1)} \right), \left(\underbrace{\frac{1}{n}}_{p(0)}, \underbrace{\frac{n-1}{n}}_{p(1)} \right), \ldots, \left(\underbrace{\frac{n}{n}}_{p(0)}, \underbrace{\frac{0}{n}}_{p(1)} \right) \right\}$$

TYPE CLASS

Type class:

Given $P \in \mathcal{P}_n$, its type class is,

$$T(P) \stackrel{\text{def}}{=} \left\{ \mathbf{x} \in \mathcal{X}^n : \; P_{\mathbf{x}}^{\text{emp}} = P \right\}$$

The type class is the set of vectors having the same empirical distribution

Exercise:

Let $\mathcal{X} = \{0, 1\}$ and $\mathbf{x} = (1, 1, 1, 0, 1, 0, 0, 1) \in \mathcal{X}^8$. Describe $T\left(P_{\mathbf{x}}^{\text{emp}}\right)$.

The number of class type is polynomial!

Proposition:

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

Proof:

For any $\mathbf{x} \in \mathcal{X}^n$,

$$P^{\text{emp}}_{\mathbf{x}} \in \underbrace{\overbrace{\left\{ \frac{i}{n} : 0 \leq i \leq n \right\}}^{p(a_0)} \times \cdots \times \overbrace{\left\{ \frac{i}{n} : 0 \leq i \leq n \right\}}^{p(a_{\sharp \mathcal{X}})}}_{\sharp \mathcal{X} \text{ times}}$$

Therefore,

$$\sharp \mathcal{P}_n \leq \underbrace{\sharp \left\{ \frac{i}{n} : 0 \leq i \leq n \right\} \times \cdots \times \sharp \left\{ \frac{i}{n} : 0 \leq i \leq n \right\}}_{\sharp \mathcal{X} \text{ times}} = (n+1)^{\sharp \mathcal{X}}$$

$$\mathcal{P}_n = \left\{ P_{\mathbf{x}}^{\mathrm{emp}} : \mathbf{x} \in \mathcal{X}^n \right\}$$

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

$\longrightarrow$ There is a polynomial number of types of length n

But there are $\sharp \mathcal{X}^n = 2^{n \log_2 \sharp \mathcal{X}}$ sequences (exponential)

Pigeonhole principle:

It exists an exponential number of sequences having the same type!

Exercise:

Let $\mathcal{X} = \{0, 1\}$. How many different types $P_{\mathbf{x}}^{\mathrm{emp}}$ exist? How many sequences have a fixed type?

Kullback-Leiber divergence:

$$D_{\mathrm{KL}}(P||Q) = \sum_a P(a) \log_2 \frac{P(a)}{Q(a)}$$

Theorem:

Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be i.i.d according to Q, then

$$Q^n(\mathbf{x}) = 2^{-n\left(H(P_{\mathbf{x}}^{\mathrm{emp}}) + D_{\mathrm{KL}}(P_{\mathbf{x}}^{\mathrm{emp}}||Q)\right)}$$

Furthermore, if $Q \in \mathcal{P}_n$ and $\mathbf{x} \in T(Q)$ $\left(\text{ where } T(Q) = \{\mathbf{x} \in \mathcal{X}^n : \; Q_{\mathbf{x}}^{\mathrm{emp}} = Q\}\right)$,

$$Q^n(\mathbf{x}) = 2^{-nH(Q)}$$

$\longrightarrow$ It shows that, when considering a sequence $\mathbf{x}$ with its associated empirical distribution,

i.e., $P_{\mathbf{x}}^{\mathrm{emp}}$, what we loose is $D_{\mathrm{KL}}\left(P_{\mathbf{x}}^{\mathrm{emp}}||Q\right)$

(with the AEP, a typical event happens with probability $2^{-nH(\mathbf{X})}$)

Proof:

$$
\begin{aligned}
Q^n(\mathrm{x}) &= \prod_{a \in \mathcal{X}} Q(a)^{\sharp\{i \in [1,n]:\ x_i = a\}} \\
&= \prod_{a \in \mathcal{X}} Q(a)^{n P_{\mathrm{x}}^{\mathrm{emp}}(a)} \\
&= \prod_{a \in \mathcal{X}} 2^{n\left(P_{\mathrm{x}}^{\mathrm{emp}}(a) \log_2 Q(a) - P_{\mathrm{x}}^{\mathrm{emp}}(a) \log_2 P_{\mathrm{x}}^{\mathrm{emp}}(a) + P_{\mathrm{x}}^{\mathrm{emp}}(a) \log_2 P_{\mathrm{x}}^{\mathrm{emp}}(a)\right)}
\end{aligned}
$$

Therefore, $Q^n(\mathrm{x}) = 2^{-n\left(H(P_{\mathrm{x}}^{\mathrm{emp}}) + D_{\mathrm{KL}}(P_{\mathrm{x}}^{\mathrm{emp}}||Q)\right)}$. To conclude use that $D_{\mathrm{KL}}(P||Q) = 0$ if $P = Q$

Theorem:

For any $P_{\mathbf{x}}^{\text{emp}} \in \mathcal{P}_n$,

$$\frac{1}{(n+1)^{\sharp\mathcal{X}}} \; 2^{nH\left(P_{\mathbf{x}}^{\text{emp}}\right)} \leq \sharp T\left(P_{\mathbf{x}}^{\text{emp}}\right) \leq 2^{nH\left(P_{\mathbf{x}}^{\text{emp}}\right)}$$

where $T\left(P_{\mathbf{x}}^{\text{emp}}\right) = \left\{\mathbf{y} \in \mathcal{X}^n : \; P_{\mathbf{y}}^{\text{emp}} = P_{\mathbf{x}}^{\text{emp}}\right\}$

Proof:

Let $P \stackrel{\text{def}}{=} P_{\mathbf{x}}^{\text{emp}}$,

$$1 \geq \textstyle\sum_{\mathbf{x}\in T(P)} P^n(\mathbf{x}) \underbrace{=}_{\text{prev th.}} \displaystyle\sum_{\mathbf{x}\in T(P)} 2^{-nH(P)} = \sharp T(P) \; 2^{-nH(P)} \quad \text{which gives the upper bound}$$

To derive the lower bound, let us admit: $\forall Q \in \mathcal{P}_n, \;\; P\left(T\left(Q\right)\right) \leq P\left(T(P)\right)$.

$$1 = \sum_{Q\in\mathcal{P}_n} P^n(T(Q)) \leq \sum_{Q\in\mathcal{P}_n} P^n(T(P)) \leq (n+1)^{\sharp\mathcal{X}} \sum_{\mathbf{x}\in T(P)} P^n(\mathbf{x}) = (n+1)^{\sharp\mathcal{X}} \sum_{\mathbf{x}\in T(P)} 2^{-nH(P)}$$

Theorem:

For any $P \in \mathcal{P}_n$ we have,

$$\frac{1}{(n+1)^{\sharp\mathcal{X}}} \, 2^{-nD_{\mathrm{KL}}(P||Q)} \leq Q^n\left(T\left(P\right)\right) = \underset{\mathbf{x}\leftarrow Q^n}{\mathbb{P}}\left(P_{\mathbf{x}}^{\mathrm{emp}} = P\right) \leq 2^{-nD_{\mathrm{KL}}(P||Q)}$$

Proof:

$$\begin{aligned} Q^n(T(P)) &= \sum_{\mathbf{x}\in T(P)} Q^n(\mathbf{x}) \\ &= \sum_{\mathbf{x}\in T(P)} 2^{-n\left(D_{\mathrm{KL}}(P||Q)+H(P)\right)} \\ &= \sharp T(P)\, 2^{-n\left(D_{\mathrm{KL}}(P||Q)+H(P)\right)} \end{aligned}$$

Moral: sequences with empirical distribution P appear under the distribution Q with an exponentially small probability, whose exponent is given by $D_{\mathrm{KL}}(P, Q)$

- $\sharp \mathcal{P}_n \leq (n+1)^{\sharp \mathcal{X}}$ $\Big($*polynomial number of types*$\Big)$
- $\sharp T\left(P_{\mathbf{x}}^{\mathrm{emp}}\right) \overset{\text{(poly)}}{=} 2^{nH\left(P_{\mathbf{x}}^{\mathrm{emp}}\right)}$ $\Big($*exponential number of sequences in each type*$\Big)$
- $Q^n(\mathbf{x}) = 2^{-n\left(H\left(P_{\mathbf{x}}^{\mathrm{emp}}\right)+D_{\mathrm{KL}}\left(P_{\mathbf{x}}^{\mathrm{emp}}||Q\right)\right)}$ $\Big($*probability of sequence of some type under Q*$\Big)$
- $Q^n\left(T\left(P_{\mathbf{x}}^{\mathrm{emp}}\right)\right) \overset{\text{(poly)}}{=} 2^{-nD_{\mathrm{KL}}\left(P_{\mathbf{x}}^{\mathrm{emp}}||Q\right)}$

Exercise:

Explicit and compute the above equations when $\mathcal{X} = \{0,1\}$

$\longrightarrow$ These results admit many consequences that we will describe now!

ALTERNATIVE LAW OF LARGE NUMBERS

Types and type classes offer an alternative “statement‘” of the law of large numbers!

The crucial property:

Polynomial number of types and an exponential number of sequences of each type

But the probability of each type class $T(P)$ depends exponentially on $D_{\mathrm{KL}}(P||Q_{\text{true distrib.}})$

$\longrightarrow$ Type classes far from the true distribution have exponentially smaller probability!

New concept of typical sequences: close for the KL-divergence

Typical Set:

Let $\varepsilon > 0$ and a distribution Q,

$$T_Q^{(\varepsilon)} = \{\mathsf{x} \in \mathcal{X}^n : \; D_{\mathrm{KL}}(P_{\mathsf{x}}^{\mathrm{emp}} || Q) \leq \varepsilon\}$$

$\longrightarrow$ The probability of not being "empirical typical" is exponentially small! $\big($similar to AEP$\big)$

Probability of not being typical:

$$\mathbb{P}_{\mathsf{x} \leftarrow Q^n} \left(D_{\mathrm{KL}} \left(P_{\mathsf{x}}^{\mathrm{emp}} || Q\right) > \varepsilon\right) = 1 - Q^n(T_Q^{(\varepsilon)}) \leq (n+1)^{\sharp\mathcal{X}} \; 2^{-n\varepsilon}$$

Proof:

$$1 - Q^n(T_Q^{(\varepsilon)}) = \sum_{P \in \mathcal{P}_n : \; D_{\mathrm{KL}}(P||Q) \geq \varepsilon} Q^n(T(P)) \leq \sum_{P \in \mathcal{P}_n : \; D_{\mathrm{KL}}(P||Q) \geq \varepsilon} 2^{-n\varepsilon}$$

To conclude the proof, use that there are a polynomial number of types!

Law of large number with KL-divergence $\Big($admitted$\Big)$:

$X_1, \dots, X_n$ be i.i.d according to Q,

$$\mathop{\mathbb{P}}_{\mathbf{x} \leftarrow Q^n} \Big(\mathbf{x} : \; D_{\mathrm{KL}}(P_{\mathbf{x}}^{\mathrm{emp}} || Q) > \varepsilon \Big) \leq (n+1)^{\sharp \mathcal{X}} \, 2^{-n\varepsilon}$$

and, if $\mathbf{x}^{(n)} \leftarrow Q^n$, then $D_{\mathrm{KL}}(P_{\mathbf{x}}^{\mathrm{emp}} || Q) \xrightarrow[n \to +\infty]{} 0$ almost surely, *i.e.*,

$$\forall \varepsilon > 0, \quad \mathbb{P}\left(\lim_{n \to +\infty} D_{\mathrm{KL}} \left(P_{\mathbf{x}^{(n)}} || Q \right) \leq \varepsilon \right) = 1$$

UNIVERSAL CODING

Huffman compresses source with known distribution X with an amount of bits given by the entropy $H(\mathsf{X})$

$\longrightarrow$ if instead a distribution Y is assumed: a penalty of $D_{\mathrm{KL}}(\mathsf{X}||\mathsf{Y})$ is incurred!

What compression can be achieved if the true distribution X **is unknown**?

Universal coding:

A symbol code $\varphi : \mathcal{X}^n \to \{0,1\}^{nR}$ is said to be 2^{nR}-universal if we can decode it with probability tending to one, *i.e.* it exists,

$$\text{Dec} : \{0,1\}^{nR} \to \mathcal{X}^n$$

such that independently of the **memoryless** source distribution Q,

$$P_e^{(n)} \stackrel{\text{def}}{=} \mathbb{P}_{\mathbf{x} \leftarrow Q^n}\left(\text{Dec}(\varphi(\mathbf{x})) \neq \mathbf{x}\right) \xrightarrow[n \to +\infty]{} 0$$

Consequence of AEP: Shannon source coding theorem **but** for known source distribution!

Is type method enables to prove the stronger statement that universal source coding exits?

Universal source coding: it is possible!

There exists a sequence $\Big(\text{in } n\Big)$ of 2^{nR}-universal code such that $P_e^{(n)} \xrightarrow[n\to+\infty]{} 0$ independently of the memoryless source distribution Q as soon as $R > H(Q)$

Proof:

Let $R_n \overset{\text{def}}{=} R - \sharp\mathcal{X}\frac{\log_2(n+1)}{n}$. We consider sequences:

$$A = \left\{\mathbf{x} \in \mathcal{X}^n \,:\, H\left(P_{\mathbf{x}}^{\text{emp}}\right) \leq R_n\right\}$$

Then,

$$\begin{aligned}
\sharp A &= \sum_{P \in \mathcal{P}_n:\, H(P) \leq R_n} \sharp T(P) \\
&\leq \sum_{P \in \mathcal{P}_n:\, H(P) \leq R_n} 2^{nH(P)} \\
&\leq \sum_{P \in \mathcal{P}_n:\, H(P) \leq R_n} 2^{nR_n} \\
&\leq (n+1)^{\sharp\mathcal{X}}\ 2^{nR_n} \qquad \left(\textit{the method of types in action!}\right) \\
&= 2^{nR}
\end{aligned}$$

- **Encoding**:

  $$\varphi(\mathbf{x}) = \begin{cases} \text{index of } \mathbf{x} \text{ in } A & \text{if } \mathbf{x} \in A \\ \perp & \text{otherwise} \end{cases}$$

  The compression size asks $\log_2 \sharp A \leq nR$ bits!

- **Decoding**: map an index to its corresponding element

PROOF (II): ERROR IN DECODING

Proof:

$X_1, \dots, X_n$ be i.i.d according to Q where $H(Q) < R$. The probability to make a mistake during decoding verifies,

$$
\begin{aligned}
P_e^{(n)} &= 1 - Q^n(A) \\
&= \sum_{P \in \mathcal{P}_n :\ H(P) > R_n} Q^n(T(P)) \\
&\leq (n+1)^{\sharp \mathcal{X}} \max_{P:\ H(P) > R_n} Q^n(T(P)) \\
&\leq (n+1)^{\sharp \mathcal{X}}\, 2^{-n \min_{P:\ H(P) > R_n} D_{\mathrm{KL}}(P||Q)} \qquad \left(\textit{the method of types in action!}\right)
\end{aligned}
$$

But $R_n \xrightarrow[n \to +\infty]{} R$ and $H(Q) < R$. Therefore, for n sufficiently large $H(Q) < R_n$ and

$$H(P) > R_n \implies H(P) > H(Q)$$

from which we conclude that $P \neq Q$ and by Gibb's inequality $D_{\mathrm{KL}}(P||Q) > 0$

Some remark:

The error in the decoding tends to 0 exponentially fast in the code-length

Universal source coding is a huge topic

- *Elements of Information Theory, Universal Source Coding Chapter 13*, by M. Cover & Joy A. Thomas

LARGE DEVIATION THEORY

I repeated too often: a random variable is equal to its expectation...

But, if yes, why? If you don't believe me, how to convince you that I say the truth?

$\longrightarrow$ Let us study $\mathbb{P}(X \gg \mathbb{E}(X))$

First two approaches:

- Markov inequality
- Bienaymé-Tchebychev inequality

Markov's Inequality:

Given $X : \Omega \longrightarrow \mathbb{R}_+$ and $t > 0$,

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}(X)}{t}$$

Bienaymé-Tchebychev

Given $t > 0$,

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq t) \leq \frac{\mathbb{V}(X)}{t^2}$$

$\longrightarrow$ Are these inequalities tight?

- We know random variables s.t the above probabilities reach the inequalities $\Big($exercise$\Big)$

Markov and Bienaymé-Tchebychev are worst-case bounds $\Big($true for any random variable$\Big)$

$\longrightarrow$ Goal of large deviation theory: provide better bounds for a given family of random variables!

Given $X_1, \dots, X_n \in \{0, 1\}$ be i.i.d with $\mathbb{P}(X_i = 1) = p$.

$$X^{(n)} \overset{\text{def}}{=} \sum_{i=1}^{n} X_i$$

$$\mathbb{E}\left(X^{(n)}\right) = np \quad \text{and} \quad \mathbb{V}\left(X^{(n)}\right) = np(1-p)$$

Bienaymé-Tchebychev:

Let $\varepsilon > 0$,

$$\mathbb{P}\left(\left|X^{(n)} - np\right| \geq \varepsilon n\right) \leq \frac{1}{\varepsilon n}\, p(1-p) \xrightarrow[n \to +\infty]{} 0$$

It tends to 0 as $1/(\varepsilon n)$, is it the best that we can expect?

Given $X_1, \dots, X_n \in \{0,1\}$ be i.i.d with $\mathbb{P}(X_i = 1) = p$.

$$X^{(n)} \overset{\text{def}}{=} \sum_{i=1}^{n} X_i$$

$$\mathbb{E}\left(X^{(n)}\right) = np \quad \text{and} \quad \mathbb{V}\left(X^{(n)}\right) = np(1-p)$$

Bienaymé-Tchebychev:

Let $\varepsilon > 0$,

$$\mathbb{P}\left(\left|X^{(n)} - np\right| \geq \varepsilon n\right) \leq \frac{1}{\varepsilon n}\, p(1-p) \xrightarrow[n \to +\infty]{} 0$$

It tends to 0 as $1/(\varepsilon n)$, is it the best that we can expect?

$\longrightarrow$ No!

CHERNOFF'S BOUND

Given $X_1, \ldots, X_n \in \{0,1\}$ be i.i.d with $\mathbb{P}(X_i = 1) = p$.

$$X^{(n)} \stackrel{\text{def}}{=} \sum_{i=1}^{n} X_i$$

$$\mathbb{E}\left(X^{(n)}\right) = np \quad \text{and} \quad \mathbb{V}\left(X^{(n)}\right) = np(1-p)$$

Chernoff:

Let $\varepsilon > 0$,

$$\mathbb{P}\left(\left|X^{(n)} - np\right| \geq \varepsilon n\right) \leq 2\, \mathrm{e}^{-2\varepsilon^2\, n} \xrightarrow[n \to +\infty]{} 0$$

It tends to 0 as $\mathrm{e}^{-2\varepsilon^2\, n}$: exponentially better than $1/(\varepsilon n)$

But is $-2\varepsilon^2 n$ the best exponent that we can expect?

Given $X_1, \ldots, X_n \in \{0,1\}$ be i.i.d with $\mathbb{P}(X_i = 1) = p$.

$$X^{(n)} \stackrel{\text{def}}{=} \sum_{i=1}^{n} X_i$$

$$\mathbb{E}\left(X^{(n)}\right) = np \quad \text{and} \quad \mathbb{V}\left(X^{(n)}\right) = np(1-p)$$

Chernoff:

Let $\varepsilon > 0$,

$$\mathbb{P}\left(\left|X^{(n)} - np\right| \geq \varepsilon n\right) \leq 2\,\mathrm{e}^{-2\varepsilon^2\, n} \xrightarrow[n \to +\infty]{} 0$$

It tends to 0 as $\mathrm{e}^{-2\varepsilon^2\, n}$: exponentially better than $1/(\varepsilon n)$

But is $-2\varepsilon^2 n$ the best exponent that we can expect?

$\longrightarrow$ Yes as we show now thanks to the method of types!

CHERNOFF'S BOUND

- First approach: central limit theorem

  $\longrightarrow$ Poor estimations in many cases $\big($try with the binomial distribution$\big)$

- Our approach: method of type!

$\mathbf{x}_1, \ldots, \mathbf{x}_n \in \{0,1\}$ be i.i.d with $\mathbb{P}(\mathbf{x}_i = 1) = \frac{1}{3}$. **Crucial remark:** if,

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i = \frac{3}{4}, \quad \text{then } P_{\mathbf{x}}^{\mathrm{emp}} = \left(\frac{1}{4}, \frac{3}{4}\right)$$

$\longrightarrow$ We expect $\big($**by method of types**$\big)$ to obtain such sequences with probability

$$\approx 2^{-nD_{\mathrm{KL}}\left(\left(\frac{1}{4},\frac{3}{4}\right)||\left(\frac{2}{3},\frac{1}{3}\right)\right)}$$

$\big($exponentially small and we know that the exponent cannot be smaller$\big)$

We know the optimal exponent...

Theorem:

For any $P \in \mathcal{P}_n$ we have,

$$\frac{1}{(n+1)^{\sharp \mathcal{X}}} 2^{-nD_{\mathrm{KL}}(P||Q)} \leq Q^n(T(P)) = \mathop{\mathbb{P}}_{\mathbf{x} \leftarrow Q^n}\left(P_{\mathbf{x}}^{\mathrm{emp}} = P\right) \leq 2^{-nD_{\mathrm{KL}}(P||Q)}$$

$$Q = \begin{cases} 0 \text{ with probability } 1-p \\ 1 \text{ with probability } p \end{cases} \quad \text{and} \quad P = \begin{cases} 0 \text{ with probability } 1-p-\varepsilon \\ 1 \text{ with probability } p+\varepsilon \end{cases}$$

Some computation:

For $\varepsilon > 0$ small enough,

$$D_{\mathrm{KL}}(P||Q) = \frac{2\varepsilon^2}{\ln(2)} + o(1)$$

The crucial remark:

$$\mathbb{P}_{\mathbf{x} \leftarrow Q^n}(x_1 + \cdots + x_n = np + \varepsilon n) = \mathbb{P}_{\mathbf{x} \leftarrow Q^n}\left(P_{\mathbf{x}}^{\mathrm{emp}} = P\right) = Q^n(T(P))$$

From the previous theorem,

$$\frac{1}{(n+1)^2} \mathrm{e}^{-2n\varepsilon^2(1+o(1))} \leq \mathbb{P}_{\mathbf{x} \leftarrow Q^n}(x_1 + \cdots + x_n = np + \varepsilon n) \leq \mathrm{e}^{-2n\varepsilon^2(1+o(1))}$$

We almost recover the optimality of Chernoff's bound! We want to know sufficiently large deviation, not just the deviation exactly equal to $+\varepsilon n$

- By the previous bound, for all η (integer)

$$\frac{1}{(n+1)^2}\mathrm{e}^{-2n(\varepsilon+\eta/n)^2(1+o(1))} \leq \mathbb{P}_{\mathbf{x}\leftarrow Q^n}(x_1+\cdots+x_n = np+\varepsilon n+\eta) \leq \mathrm{e}^{-2n(\varepsilon+\eta/n)^2(1+o(1))}$$

- By summing all possible η (*polynomial number of types*)

$$\frac{1}{(n+1)}\mathrm{e}^{-2n\varepsilon^2(1+o(1))} \leq \mathbb{P}_{\mathbf{x}\leftarrow Q^n}(x_1+\cdots+x_n \geq np+\varepsilon n) \leq (n+1)\mathrm{e}^{-2n\varepsilon^2(1+o(1))}$$

We can obtain the same bound for

$$\mathbb{P}_{\mathbf{x}\leftarrow Q^n}(x_1+\cdots+x_n \leq np-\varepsilon n)$$

by replacing $\varepsilon \longleftrightarrow -\varepsilon$

$$Q = \begin{cases} 0 \text{ with probability } 1 - p \\ 1 \text{ with probability } p \end{cases}$$

$$\frac{1}{n} \log_2 \mathbb{P}_{\mathbf{x} \leftarrow Q^n} (|x_1 + \cdots + x_n - np| \geq \varepsilon n) = -2\varepsilon^2 / \ln(2) + o(1)$$

We obtained the best exponent by using method of types!

To obtain the "optimality" of Chernoff's bound we made the following reasoning

1. Start from the distribution Q
2. Our aim: to get an upper bound on $\underset{\mathbf{x} \leftarrow Q^n}{\mathbb{P}} \left(\sum_i x_i = \mathbb{E}(X) + \alpha\right)$
3. To this aim we introduced the distribution P s.t $P_{\mathbf{x}}^{\text{emp}} = P$ if and only if $\sum_i x_i = \mathbb{E}(X) + \alpha$

We are going to systematize this approach: Sanov's theorem!

SANOV'S THEOREM

Given i.i.d random variables $\mathbf{X}_i$'s distributed as Q, we want to estimate,

$$\mathop{\mathbb{P}}_{\mathbf{x}\leftarrow Q^n}\left(\frac{1}{n}\sum_{i=1}^{n} g(x_i) \geq \alpha\right)$$

Fundamental ideal: introduce the $\mathbf{x} \leftarrow Q^n$ such that $\sum_{a\in\mathcal{X}} g(a)P_{\mathbf{x}}^{\text{emp}}(a) \geq \alpha$

$\longrightarrow$ We expect the probability exponent to behave as $D_{\text{KL}}\left(P_{\mathbf{x}}^{\text{emp}}||Q\right)$

THE ISSUE

The fundamental ideal: introduce the $\mathbf{x} \leftarrow Q^n$ such that $\sum_{a \in \mathcal{X}} g(a) P_{\mathbf{x}}^{\mathrm{emp}}(a) \geq \alpha$

$\longrightarrow$ We expect the probability exponent to behave as $D_{\mathrm{KL}}\left(P_{\mathbf{x}}^{\mathrm{emp}} || Q\right)$

Issue:

There are many $P_{\mathbf{x}}^{\mathrm{emp}}$'s (from different classes) verifying $\sum_{a \in \mathcal{X}} g(a) P_{\mathbf{x}}^{\mathrm{emp}}(a) \geq \alpha$

$\longrightarrow$ We will show that the exponent behaves as $D_{\mathrm{KL}}(P^\star || Q)$ for $P^\star$ **minimizing** $D_{\mathrm{KL}}(P || Q)$ for the

$P \in \mathcal{P}_n$'s verifying $\sum_{a \in \mathcal{X}} g(a) P(a) \geq \alpha$

(the minimization is here to "extract" the dominant exponential term)

Goal: find the "closest" P in the constraint set for the KL-divergence to obtain the exponent!

$\longrightarrow$ We need to define the topology associated to this "KL-distance"

Probability simplex:

Subset of $[0,1]^{\sharp\mathcal{X}}$ defined as,

$$\mathcal{P} \stackrel{\text{def}}{=} \left\{ (x_1, \ldots, x_{\sharp\mathcal{X}}) \in [0,1]^{\sharp\mathcal{X}} \;:\; x_i \geq 0 \text{ and } \sum_{i=1}^{\sharp\mathcal{X}} x_i = 1 \right\}$$

$\mathcal{P} \subseteq \mathbb{R}^{\sharp\mathcal{X}}$, we will speak of closure, interior, . . . But for the D_{KL}-divergence

We will identify distributions P over $\mathcal{X}$ with elements of $\mathcal{S}$

Proposition:

$\bigcup_{n\in\mathbb{N}} \mathcal{P}_n$ is dense in $\mathcal{S}$ the set of all distributions over $\mathcal{X}$

Proof:

Given $P = (p(a_1), \ldots, p(a_m)) \in \mathcal{S}$,

$$\forall i \in [1, \sharp\mathcal{X} - 1],\; n_i \stackrel{\text{def}}{=} \frac{\lfloor np(a_i) \rfloor}{n} \quad \text{and} \quad n_{\sharp\mathcal{S}} = \frac{1 - \sum_{i=1}^{\sharp\mathcal{X}-1} n_i}{n}$$

Then, $P_n^{\mathrm{emp}} \stackrel{\text{def}}{=} (n_i)_i \in \mathcal{P}_n$ and $D_{\mathrm{KL}}\left(P_n^{\mathrm{emp}} || P\right) \xrightarrow[n\to+\infty]{} 0$

ABUSE OF NOTATION

$$P^{\text{emp}}_{\mathbf{x}} \longleftrightarrow \mathbf{x}$$

- Given $E \subseteq \mathcal{P}$,

$$Q^n(E) \stackrel{\text{def}}{=} Q^n(E \cap \mathcal{P}_n) = \sum_{\mathbf{x}:\ P^{\text{emp}}_{\mathbf{x}} \in E \cap \mathcal{P}_n} Q^n(\mathbf{x})$$

$$Q^n(E) = \mathop{\mathbb{P}}_{\mathbf{x} \leftarrow Q^n}\left(P^{\text{emp}}_{\mathbf{x}} \in E \cap \mathcal{P}_n\right)$$

But why this formalism?

Given a random variable $X = (X_1, \ldots, X_n) \in \mathcal{X}^n$ where the X_i's are i.i.d according to Q and some function g, we are interested in

$$\mathbb{P}_{X \leftarrow Q^n}\left(\frac{1}{n}\sum_{i=1}^{n} g(x_i) \geq \alpha\right)$$

We introduce:

$$E \stackrel{\text{def}}{=} \left\{P \in \mathcal{P} : \sum_{a \in \mathcal{X}} g(a)p(a) \geq \alpha\right\}$$

$$\frac{1}{n}\sum_{i=1}^{n} g(x_i) \geq \alpha \iff \sum_{a \in \mathcal{X}} g(a) P_X^{\text{emp}}(a) \geq \alpha$$

$$\iff P_X^{\text{emp}} \in E \cap \mathcal{P}_n$$

Conclusion:

$$\mathbb{P}_X\left(\frac{1}{n}\sum_{i=1}^{n} g(X_i) \geq \alpha\right) = Q^n(E) = \mathbb{P}_{X \leftarrow Q^n}\left(P_X^{\text{emp}} \in E \cap \mathcal{P}_n\right)$$

We are interested in

$$\mathop{\mathbb{P}}_{\mathbf{x}\leftarrow Q^n}\left(\frac{1}{n}\sum_{i=1}^n g(x_i) \geq \alpha\right) = Q^n(E) = \mathop{\mathbb{P}}_{\mathbf{x}\leftarrow Q^n}\left(P_{\mathbf{x}}^{\text{emp}} \in E \cap \mathcal{P}_n\right)$$

How does $Q^n(E)$ behave?

Let us use what we already know: our alternative law of large numbers $\left(\text{see Slide } 20\right)$

Let $E \subseteq \mathcal{P}$,

- Suppose that E contains a relative entropy neighbourhood of Q, *i.e.*,

$$\exists \varepsilon > 0, \quad \text{such that } \{P \in \mathcal{P} : \; D_{\mathrm{KL}}(P||Q) < \varepsilon\} \subseteq E$$

$$Q^n(E) = \underset{\mathsf{x} \leftarrow Q^n}{\mathbb{P}} \left(P_{\mathsf{x}}^{\mathrm{emp}} \in E \cap \mathcal{P}_n\right) \geq \underset{\mathsf{x} \leftarrow Q^n}{\mathbb{P}} \left(D_{\mathrm{KL}}\left(P_{\mathsf{x}}^{\mathrm{emp}}||Q\right) < \varepsilon\right) \geq 1 - (n+1)^{\sharp \mathcal{X}} 2^{-n\varepsilon} \xrightarrow[n \to +\infty]{} 1$$

- Suppose that E does not contain Q or any element of some neighbourhood of Q, *i.e.*,

$$\exists \varepsilon \geq 0 \text{ such that } \{P \in \mathcal{P} : \; D_{\mathrm{KL}}(P||Q) \leq \varepsilon\} \cap E = \emptyset$$

$$Q^n(E) = \underset{\mathsf{x} \leftarrow Q^n}{\mathbb{P}} \left(P_{\mathsf{x}}^{\mathrm{emp}} \in E \cap \mathcal{P}_n\right) \leq \underset{\mathsf{x} \leftarrow Q^n}{\mathbb{P}} \left(D_{\mathrm{KL}}\left(P_{\mathsf{x}}^{\mathrm{emp}}||Q\right) \geq \varepsilon\right) \leq (n+1)^{\sharp \mathcal{X}} 2^{-n\varepsilon} \xrightarrow[n \to +\infty]{} 0$$

It tends to zero exponentially fast, but what is the exponent?

Sanov's Theorem:

Let $X_1, \ldots, X_n$ be i.i.d. according to Q. Let $E \subseteq \mathcal{P}$. Then,

$$\mathbb{P}_{\mathbf{x} \leftarrow Q^n}\left(P_{\mathbf{x}}^{\mathrm{emp}} \in E \cap \mathcal{P}_n\right) = Q^n(E) \leq (n+1)^{\sharp \mathcal{X}} 2^{-nD_{\mathrm{KL}}\left(P^{\star} || Q\right)} \quad \text{where } P^{\star} \stackrel{\text{def}}{=} \operatorname*{arg\,min}_{P \in E} D_{\mathrm{KL}}(P||Q)$$

Furthermore, if E is the closure of its interior,

$$\frac{1}{n} \log_2 Q^n(E) \xrightarrow[n \to +\infty]{} D_{\mathrm{KL}}\left(P^{\star} || Q\right)$$

Important remark:

$D_{\mathrm{KL}}\left(P^{\star} || Q\right)$ is the exponent of the probability when E is the closure of its interior!

Proof:

$$
\begin{aligned}
Q^n(E) &= \sum_{P \in E \cap \mathcal{P}_n} Q^n(T(P)) \\
&\leq \sum_{P \in E \cap \mathcal{P}_n} 2^{-nD_{\mathrm{KL}}(P||Q)} \\
&\leq \sum_{P \in E \cap \mathcal{P}_n} \max_{P \in E \cap \mathcal{P}_n} 2^{-nD_{\mathrm{KL}}(P||Q)} \\
&\leq \sum_{P \in E \cap \mathcal{P}_n} 2^{-n \min_{P \in E \cap \mathcal{P}_n} D_{\mathrm{KL}}(P||Q)} \\
&\leq \sum_{P \in E \cap \mathcal{P}_n} 2^{-nD_{\mathrm{KL}}(P^\star||Q)} \\
&\leq (n+1)^{\sharp \mathcal{X}} \, 2^{-nD_{\mathrm{KL}}(P^\star||Q)}
\end{aligned}
$$

$\longrightarrow$ The last inequality shows the method of types in action: exponential probability of events versus polynomial number of events!

PROOF (II): LOWER BOUND

Proof:

Suppose that E is the closure of its interior. In particular, interior E° is not empty. As $\bigcup_n \mathcal{P}_n$ is dense in $\mathcal{P}$, then $E^{\circ} \cap \mathcal{P}_n$ is non-empty for n large enough $\geq n_0$. By using the density, it exists $P_n \in E^{\circ} \cap \mathcal{P}_n$ such that $D_{\mathrm{KL}}(P_n||Q) \xrightarrow[n\to+\infty]{} D\left(P^{\star}||Q\right)$. Furthermore,

$$\begin{aligned} Q^n(E) &= \sum_{P \in E \cap \mathcal{P}_n} Q^n(T(P)) \\ &\geq Q^n(T(P_n)) \\ &\geq \frac{1}{(n+1)^{\sharp \mathcal{X}}} 2^{-nD_{\mathrm{KL}}(P_n||Q)} \qquad \textit{(the method of types in action!)} \end{aligned}$$

Therefore, combining with the upper bound,

$$-D_{\mathrm{KL}}\left(P_n||Q\right) - \frac{\sharp \mathcal{X} \log_2(n+1)}{n} \leq \frac{1}{n} \log_2 Q^n(E) \leq \frac{\sharp \mathcal{X} \ \log_2(n+1)}{n} - D_{\mathrm{KL}}\left(P_n||Q\right)$$

APPLICATIONS OF SANOV'S THEOREM

We want to tightly derive an upper bound on:

$$\forall j \in [1,k], \quad \mathop{\mathbb{P}}_{\mathbf{x} \leftarrow Q^n}\left(\frac{1}{n}\sum_{i=1}^{n} g_j(x_i) \geq \alpha_j\right)$$

How to proceed: use Sanov theorem!

$$E = \left\{P \in \mathcal{P} : \ \forall j \in [1,k], \ \textstyle\sum_a P(a) g_j(a) \geq \alpha_j\right\} \qquad \Big(\text{closure of its interior}\Big)$$

The optimal exponent is given by $D_{\mathrm{KL}}\left(P^\star || Q\right)$ where $P^\star$ is given by

$$P^\star(a) \stackrel{\text{def}}{=} \frac{Q(a) \mathrm{e}^{\sum_i \lambda_i g_i(a)}}{\sum_{x \in \mathcal{X}} Q(x) \mathrm{e}^{\sum_i \lambda_i g_i(a)}}$$

where the λ_i's are chosen such that

$$\forall j \in [1,k], \ \textstyle\sum_a P^\star(a) g_j(a) = \alpha_j$$

We want to tightly derive an upper-bound on

$$\mathbb{P}_{\mathbf{x}\leftarrow Q^n}\left(\frac{1}{n}\sum_{i=1}^{n} g(x_i) \geq \alpha\right)$$

$$E = \left\{P \in \mathcal{P} : \ \textstyle\sum_a P(a)g(a) \geq \alpha\right\}$$

We wan to minimize $D_{\mathrm{KL}}(P||Q)$ over $P \in E$! To this aim introduce the constraint functions:

$$c_0(\mathbf{p}) = \textstyle\sum_{i=1}^{n} p_i - 1 \text{ and } c_1(\mathbf{p}) \stackrel{\text{def}}{=} \sum_{i=1}^{\sharp\mathcal{X}} p_i g(a_i) - \alpha \quad \left(\mathbf{p} \in \mathbb{R}^{\sharp\mathcal{X}} \text{ the distribution vector of } P\right)$$

- First step: minimize $f(\mathbf{p}) \stackrel{\text{def}}{=} D_{\mathrm{KL}}(P||Q)$ for $P \in \widetilde{E} \stackrel{\text{def}}{=} \{P \in \mathcal{P} : \ \sum_a P(a)g(a) \geq \alpha\} \cap \mathbb{R}_{>0}^{\sharp\mathcal{X}} \subseteq E$,

$$c_0(\mathbf{p}) = c_1(\mathbf{p}) = 0$$

Use Lagrange Multiplier Theorem:

It exists λ, μ such that for all $i \in [1, \sharp\mathcal{X}]$,

$$\frac{\partial f}{\partial p_i}(\mathbf{p}) = \log_2 \frac{p_i}{Q(a_i)} + \frac{1}{\ln 2} = \mu \frac{\partial c_0}{\partial p_i}(\mathbf{p}) + \frac{\partial c_1}{\partial p_i}(\mathbf{p}) = \mu + \lambda g(a_i)$$

$\longrightarrow$ We deduce that $p_i = Q(a_i)2^{-1/\ln(2)+\mu+\lambda g(a_i)}$ is a minimum of $D_{\mathrm{KL}}(P||Q)$ for $P \in \widetilde{E}$

Does $p_i = \frac{Q(a_i)2^{\lambda g(a_i)}}{C}$ where $C = \sum_i Q(a_i)2^{\lambda g(a_i)}$ give the minimum of $D_{\text{KL}}(P||Q)$ for $P \in E$ as expected?

- First computation:

$$D_{\text{KL}}(p||Q) = \sum_i p_i \log_2 \frac{Q(a_i)2^{\lambda g(a_i)}}{CQ(a_i)} = \lambda \underbrace{\sum_i p_i g(a_i)}_{=\alpha \text{ by def of } \widetilde{E}} - \log_2 C = \lambda\alpha - \log_2 C = D_{\text{KL}}(p||Q)$$

- Second computation: for any $R \in E$,

$$\sum_i R(i) \log_2 \frac{p_i}{Q(i)} \geq \lambda\alpha - \log_2 C$$

- Conclusion:

$$D_{\text{KL}}(R||Q) - D_{\text{KL}}(p||Q) \geq D_{\text{KL}}(R||Q) - \sum_i R(i) \log_2 \frac{p_i}{Q(i)} = D_{\text{KL}}(R||p) \geq 0 \text{ by Gibb's inequality.}$$

The p_i's gives the minimum of $D_{\text{KL}}(P||Q)$ for $P \in E$!

We toss a fair dice n times, what is the probability that the average of the throws is greater than or equal to 4?

By Sanov's theorem,

$$Q^n(E) \overset{\text{(poly)}}{=} 2^{-nD_{\text{KL}}(P^\star||Q)}$$

where $P^\star$ minimizes $D_{\text{KL}}(P||Q)$over all distributions P that satisfy,

$$\textstyle\sum_{i=1}^{6} iP(i) \geq 4$$

$$\forall i \in [1,6],\ \ P^\star(i) = \frac{2^{\lambda i}}{\sum_{j=1}^{6} 2^{\lambda j}} \ \text{ where } \ \lambda \ \text{ such that } \ \sum_{i=1}^{6} iP^\star(i) = 4$$

Solving numerically we obtain $\lambda = 0.2519$, therefore $D_{\text{KL}}\left(P^\star||Q\right) = 0.0624.$ After 1000 coin tosses, the probability that the average is greater than or equal to 4 is $\approx 2^{-624}$

We want to estimate the probability of observing more than 700 heads in a series of 1000 coin tosses of a fair coin

$$\mathbb{P}\left(\bar{X}_n \geq 0.7\right) \overset{\text{(poly)}}{=} 2^{-nD_{\mathrm{KL}}(P^\star || Q)}$$

where $P^\star = (0.3, 0.7)$. In that case $D_{\mathrm{KL}}(P^\star || Q) = 1 - H(0.7) = 0.119$. Our probability is $\approx 2^{-119}$

We are given a joint distribution $(\mathbf{X}, \mathbf{Y}) \leftarrow Q(x, y)$. Let $Q(x), Q(y)$ be the associated distributions formed by the marginals

Let Q_0 be the distribution product of marginals $\Big($"as if $\mathbf{X}$ and $\mathbf{Y}$ were independent"$\Big)$

$$Q_0(x, y) \stackrel{\text{def}}{=} Q(x)Q(y)$$

We want to estimate the probability that $(x^n, y^n) \leftarrow Q_0^n(\cdot, \cdot)$ looks to be picked according to $Q(x^n, y^n)$

Estimating the probability that $P_{x^n, y^n} \in E \cap \mathcal{P}_n(\mathbf{X}, \mathbf{Y})$ when $(x^n, y^n) \leftarrow Q^n(\cdot, \cdot)$ and where

$$E \stackrel{\text{def}}{=} \Big\{ P(x, y) : \left| -\sum_{x,y} P(x, y) \log_2 Q(x) - H(\mathbf{X}) \right| \leq \varepsilon, \\ \left| -\sum_{x,y} P(x, y) \log_2 Q(y) - H(\mathbf{Y}) \right| \leq \varepsilon, \\ \left| -\sum_{x,y} P(x, y) \log_2 Q(x, y) - H(\mathbf{X}, \mathbf{Y}) \right| \leq \varepsilon \Big\}$$

MUTUAL DEPENDENCE(II)

Using Sanov's theorem,

$$Q_0^n(E) = 2^{-nD_{\mathrm{KL}}(P^\star || Q_0)}$$

where $P^\star$ is the closest distribution satisfying the constraints. Then $P^\star \xrightarrow[\varepsilon \to 0]{} Q$ $\Big($exercise session$\Big)$. The probability becomes:

$$2^{-nD_{\mathrm{KL}}(Q(x,y)||Q(x)Q(y))} = 2^{-nI(\mathbf{X},\mathbf{Y})}$$

EXERCISE SESSION