Infotheory for Statistics and Learning

Lecture 12

- The method of types[1]
  - Definition empirical distribution and type class
  - Connection between types and probability theory
  - Large deviation via types
  - Joint types, V-shell, typicality

[1] based on material by I. Csiszar such as [CK, CS].

Empirical distribution and type class

Notation: Let

- $\mathcal{A} = \{a_1, a_2, \dots a_M\}$ denote a finite set with $|\mathcal{A}| = M$;
- $x^n = (x_1, x_2 \dots, x_n) \in \mathcal{A}^n$ denote a sequence of length n;
- *frequency* and *relative freq.* of letter $a \in \mathcal{A}$ in sequence x^n:

$$N(a|x^n) = \sum_{i=1}^{n} \mathbb{1}\{x_i = a\} \qquad \hat{P}_{x^n}(a) = \frac{1}{n} N(a|x^n)$$

Empirical distribution of sequence x^n is the probability vector $(\hat{P}_{x^n}(a_1), \hat{P}_{x^n}(a_2), \dots, \hat{P}_{x^n}(a_M))$. Set of all empirical distributions of length n is denoted by $\mathbb{P}_n = \{\hat{P}_{x^n} : x^n \in \mathcal{A}^n\}$.

Type and type class is $P \in \mathbb{P}_n$ with $\mathcal{T}_P^n = \{x^n : \hat{P}_{x^n} = P\}$.

- **Example:** $x^n = (1,3,2,1,3,1) \in \{1,2,3\}^6 \to \hat{P}_{x^n} = (\frac{3}{6}, \frac{1}{6}, \frac{2}{6})$
- **Example:** For $P = (\frac{3}{6}, \frac{1}{6}, \frac{2}{6})$ we have type class $\mathcal{T}_P^n = \{(1,1,1,2,3,3), (1,1,2,1,3,3), (1,1,2,3,1,3), \dots, (3,3,2,1,1,1)\}$

- Note, the number of sequences grows exponentially M^n.
- Q: For sequence length n, how many types can we have?

Lemma: (HW) For the number possible n-types we have[2]

$$|\mathbb{P}_n| = \binom{n+M-1}{M-1} \leq (n+1)^{M-1} \leq (n+1)^M.$$

- Key observation 1: We have only a sub-exponential growth of number of types!

- Key observation 2: For iid sequences, the probability of sequences of the same type are equal!
  - $\Rightarrow$ For the computation of the probability of a set of sequences of the same type we need the cardinality of the set.

- Q: How many sequences are in each type class $\mathcal{T}_P^n$?

[2] A type is described by a vector $\big(N(a_1|x^n), \ldots, N(a_M|x^n)\big) \in [0:n]^M$.

- Q: How many sequences are in each type class $\mathcal{T}_P^n$?

Lemma:[3] For any $P \in \mathbb{P}_n$ we have

$$\frac{1}{\binom{n+M-1}{M-1}} 2^{nH(P)} \leq |\mathcal{T}_P^n| \leq 2^{nH(P)}.$$

Proof: Let $k_i = N(a_i|x^n)$ for $1 \leq i \leq M$. The number of sequences of length n with k_i times element a_i is given by the number of distinct ways to permute a multiset[4] of n elements

$$|\mathcal{T}_P^n| = \frac{n!}{k_1!k_2! \cdot \dots \cdot k_M!}$$

- Next, find upper and lower bounds on RHS.

[3]Note that we write $H(P)$ for entropy $H(X)$ when RV X is distributed according to P. A stronger result can be obtained using the Stirling formula (HW). A weaker result can be obtained using $|\mathbb{P}_n| \leq (n+1)^M$ in the proof, which is often sufficient if we are interested in the asymptotic only.

[4]A multiset is a set where multiple instances of an element are allowed.

- Multinomial theorem gives

$$n^n = (k_1+k_2+\cdots+k_M)^n = \underbrace{\sum_{j_1+j_2+\cdots+j_M=n}}_{\binom{n+M-1}{M-1}\text{ terms}} \underbrace{\frac{n!}{j_1!j_2!\cdots j_M!}k_1^{j_1}k_2^{j_2}\cdots k_M^{j_M}}_{\overset{(*)}{\leq}\frac{n!}{k_1!k_2!\cdots k_M!}k_1^{k_1}k_2^{k_2}\cdots k_M^{k_M}} \tag{1}$$

(∗) since for $k_i \leq j_i$ we have $\frac{k_i!}{j_i!}k_i^{j_i-k_i} \leq 1$ and for $k_i > j_i$ we have $\frac{j_i!}{k_i!}k_i^{k_i-j_i} \geq 1$ it follows $\prod_{i:k_i\leq j_i}\frac{k_i!}{j_i!}k_i^{j_i-k_i} \leq \prod_{i:k_i>j_i}\frac{j_i!}{k_i!}k_i^{k_i-j_i}$

- Divide (1) by $k_1^{k_1}k_2^{k_2}\cdots k_M^{k_M}$, then LHS gives

$$\frac{n^n}{k_1^{k_1}k_2^{k_2}\cdots k_M^{k_M}} = \prod_{i=1}^{M}\left(\frac{k_i}{n}\right)^{-k_i} = \prod_{i=1}^{M}P(a_i)^{-nP(a_i)} = 2^{\log\prod_{i=1}^{M}P(a_i)^{-nP(a_i)}} = 2^{nH(P)}$$

- Recall $\frac{n!}{k_1!k_2!\cdots k_M!} = |\mathcal{T}_P^n|$ so that bounds follow from (1) with
  - Upper bound: RHS lower bounded by largest term in sum.
  - Lower bound: RHS upper bounded by taking $\binom{n+M-1}{M-1}$-times largest term in sum. □

Connection between types and probability theory

- Let P^n denote distribution of an iid sequence according to P, i.e., $P^n(x^n) = \prod_{i=1}^n P(x_i)$
- Entropy $H(\hat{P}_{x^n})$ is called *empirical entropy* of x^n.

Lemma:[5] For any $x^n \in \mathcal{A}^n$ and distribution P on $\mathcal{A}$ we have

$$P^n(x^n) = 2^{-n[H(\hat{P}_{x^n}) + D(\hat{P}_{x^n}||P)]}.$$

Proof:

$$\begin{aligned} P^n(x^n) &= \prod_{i=1}^n P(x_i) = 2^{\sum\limits_{i=1}^n \log P(x_i)} = 2^{\sum\limits_{a\in\mathcal{A}} N(a|x^n) \log P(a)} \\ &= 2^{-n\left[\sum\limits_{a\in\mathcal{A}} \frac{N(a|x^n)}{n} \log \frac{1}{P(a)}\right]} = 2^{-n\left[\sum\limits_{a\in\mathcal{A}} \hat{P}_{x^n}(a) \log \frac{1}{P(a)} \frac{\hat{P}_{x^n}(a)}{\hat{P}_{x^n}(a)}\right]} \\ &= 2^{-n[H(\hat{P}_{x^n}) + D(\hat{P}_{x^n}||P)]} \end{aligned}$$

□

[5] $H(\hat{P}_{x^n}) + D(\hat{P}_{x^n}||P) = -\sum_{a\in\mathcal{A}} \hat{P}_{x^n}(a) \log P(a)$ is called *inaccuracy*.

- A slight reformulation gives us the following lemma.

Lemma: For any distribution P on $\mathcal{A}$ and any n-type Q we have

$$\frac{P^n(x^n)}{Q^n(x^n)} = 2^{-nD(Q||P)}, \qquad x^n \in \mathcal{T}_Q^n.$$

Proof: For $x^n \in \mathcal{T}_Q^n$ we have $H(\hat{P}_{x^n}) = H(Q)$ and $D(\hat{P}_{x^n}||P) = D(Q||P)$ so that the previous lemma gives

$$Q^n(x^n) = 2^{-n[H(\hat{P}_{x^n})+D(\hat{P}_{x^n}||Q)]} = 2^{-n[H(Q)+D(Q||Q)]} = 2^{-nH(Q)}$$

and therewith

$$P^n(x^n) = 2^{-n[H(\hat{P}_{x^n})+D(\hat{P}_{x^n}||P)]} = \underbrace{2^{-n[H(Q)+D(Q||P)]}}_{=Q^n(x^n)2^{-nD(Q||P)}}$$

$\square$

- In particular, the lemma implies $P^n(\mathcal{T}_Q^n) = \sum_{x^n \in \mathcal{T}_Q^n} P^n(x^n) = 2^{-nD(Q||P)} \sum_{x^n \in \mathcal{T}_Q^n} Q^n(x^n) = 2^{-nD(Q||P)} Q(\mathcal{T}_Q^n)$.

- The next lemma combines the previous results.

Lemma: For any distribution P on $\mathcal{A}$ and any n-type Q we have

$$\frac{1}{\binom{n+M-1}{M-1}} 2^{-nD(Q||P)} \leq P^n(\mathcal{T}_Q^n) \leq 2^{-nD(Q||P)}.$$

Proof:

$$P^n(\mathcal{T}_Q^n) = \sum_{x^n \in \mathcal{T}_Q^n} P^n(x^n) = \sum_{x^n \in \mathcal{T}_Q^n} Q^n(x^n) 2^{-nD(Q||P)} \leq 2^{-nD(Q||P)}$$

$$P^n(\mathcal{T}_Q^n) = \sum_{x^n \in \mathcal{T}_Q^n} \underbrace{Q^n(x^n)}_{=2^{-nH(Q)}} 2^{-nD(Q||P)} = \underbrace{|\mathcal{T}_Q^n|}_{\geq \frac{1}{\binom{n+M-1}{M-1}} 2^{H(Q)}} 2^{-n[H(Q)+D(Q||P)]}$$

$\square$

- In particular $\frac{1}{\binom{n+M-1}{M-1}} \leq P^n(\mathcal{T}_P^n) \leq 1$.
- If $P \neq Q$, then $P^n(\mathcal{T}_Q^n) \overset{n\to\infty}{\longrightarrow} 0$ exponentially fast.

- As a direct consequence we have the following:

Corollary: Let $\hat{P}^n$ be the empirical distribution of a sequence of length n iid randomly drawn according to distribution P. Then

$$\mathrm{Prob}\{D(\hat{P}_{X^n}||P) \geq \delta\} \leq \binom{n+M-1}{M-1} 2^{-n\delta} \qquad \forall \delta > 0.$$

Proof:

$$\mathrm{Prob}\{D(\hat{P}_{X^n}||P) \geq \delta\} = \sum_{Q:D(Q||P)\geq\delta} \underbrace{P^n(\mathcal{T}_Q^n)}_{\leq 2^{-nD(Q||P)}} \leq \binom{n+M-1}{M-1} 2^{-n\delta}$$

$\square$

- Note[6] that we have $\binom{n+M-1}{M-1} 2^{-n\delta} \leq 2^{(M-1)\log n} 2^{-n\delta} \overset{n\to\infty}{\longrightarrow} 0$.

[6] In comparison, convergence of the weak law of large numbers: $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i \to \mu$ in probability when $n \to \infty$, i.e. $\lim_{n\to\infty} \mathrm{Prob}\{|\bar{X}_n - \mu| > \epsilon\} = 0$ for any $\epsilon > 0$.

Large deviation via types

- Let $D(\Pi||P) = \inf_{Q\in\Pi} D(Q||P)$ for a set of distributions Π.
- Q: Can we otherwise make an asymptotic statement?

Sanov's Theorem: Let Π be a set of distribution on $\mathcal{A}$ whose closure is equal to the closure of its interior.[7] Then for the empirical distribution $\hat{P}_n$ of a sample sequence iid of strictly positive distribution P on $\mathcal{A}$ we have

$$-\frac{1}{n}\log \mathrm{Prob}\{\hat{P}_n \in \Pi\} \overset{n\to\infty}{\longrightarrow} D(\Pi||P).$$

Proof: Let $\Pi_n = \Pi \cap \mathbb{P}_n$ be the set of possible n-types in Π, then we have $\mathrm{Prob}\{\hat{P}_n \in \Pi_n\} = P^n(\cup_{Q\in\Pi_n}\mathcal{T}_Q^n)$ and

- $\frac{1}{\binom{n+M-1}{M-1}}2^{-nD(\Pi_n||P)} \leq P^n(\cup_{Q\in\Pi_n}\mathcal{T}_Q^n) \leq \binom{n+M-1}{M-1}2^{-nD(\Pi_n||P)}$
- Result follows taking the limit of $-\frac{1}{n}\log$ of the RHS and LHS.

Convergence is guaranteed due to assumption on Π. $\square$

[7] For any set Π we have $\limsup_{n\to\infty}\frac{1}{n}\log \mathrm{Prob}\{\hat{P}_n \in \Pi\} \leq -D(\Pi||P)$.

Example

- Consider $f : \mathcal{A} \to \mathbb{R}$ and $\Pi = \{Q : \sum_{a\in\mathcal{A}} Q(a)f(a) > \alpha\}$
- $\hat{P}_{x^n} \in \Pi$ for sequence $x^n \in \mathcal{A}^n$ if $\frac{1}{n}\sum_{i=1}^n f(x_i) > \alpha$
  - since $\sum\limits_{a\in\mathcal{A}} \hat{P}_n(a)f(a) = \frac{1}{n}\sum\limits_{a\in\mathcal{A}} N(a|x^n)f(a) = \frac{1}{n}\sum_{i=1}^n f(x_i)$
- Using Sanov's theorem we have **large deviation result**

$$-\frac{1}{n}\log \mathrm{Prob}\left\{\frac{1}{n}\sum_{i=1}^n f(X_i) > \alpha\right\} \xrightarrow{n\to\infty} D(\Pi||P).$$

- Here $D(\Pi||P) = D(\mathrm{cl}(\Pi)||P) = \min_{Q:\sum_a Q(a)f(a)\geq\alpha} D(Q||P)$
- For $\alpha > \sum_a P(a)f(a)$ we have $D(\Pi||P) > 0$ [8]
  $\Rightarrow$ $\mathrm{Prob}\{\frac{1}{n}\sum_{i=1}^n f(X_i) > \alpha\}$ decays to 0 exponentially fast!
- Remaining task is to compute (or bound) $D(\Pi||P)$ for a specific function f and distribution P.[9]

[8] Note, if $\alpha \leq \sum_a P(a)f(a)$, then $P \in \Pi$ so that $D(\Pi||P) = D(P||P) = 0$.
[9] Basically same idea is used in achievability proofs in coding theorems.

Joint Type

- Besides $\mathcal{A}$ consider $\mathcal{B} = \{b_1, b_2, \ldots, b_N\}$ with $|\mathcal{B}| = N$.
- Consider two sequences $x^n \in \mathcal{A}^n$ and $y^n \in \mathcal{B}^n$ of length n.
- *Frequency and relative freq.* of $(a,b) \in \mathcal{A} \times \mathcal{B}$ in (x^n, y^n)

$$N(a,b|x^n,y^n) = \sum_{i=1}^{n} \mathbb{1}\{x_i = a, y_i = b\}, \quad \hat{P}_{x^n y^n}(a,b) = \frac{N(a,b|x^n,y^n)}{n}$$

Joint type is defined as type of the sequence $\{(x_i,y_i)\}_{i=1}^n$, in other words the relative frequency $\hat{P}_{x^n y^n}$ (aka empirical distribution).

- **Example:** $\mathcal{A} = \{1,2,3\}$ and $\mathcal{B} = \{\triangle, \square\}$
  $(x^6, y^6) = ((1,\triangle),(3,\triangle),(2,\square),(1,\square),(3,\square),(1,\triangle))$
  - $\hat{P}_{x^n y^n}(1,\triangle) = \frac{2}{6}$, $\hat{P}_{x^n y^n}(2,\triangle) = 0$, $\hat{P}_{x^n y^n}(2,\square) = \frac{1}{6}$,...
- Joint types are often given in terms of the type of x^n and a stochastic matrix[10] $V : \mathcal{A} \to \mathcal{B}$,

$$\hat{P}_{x^n y^n}(a,b) = \hat{P}_{x^n}(a) V(b|a)$$

[10] $\hat{P}_{x^n y^n}$ uniquely specifies $V(b|a)$ for which $N(a|x^n) \neq 0$.

Conditional type and V-shell

Conditional type: Sequence $y^n \in \mathcal{B}^n$ has *conditional type* V given sequence $x^n \in \mathcal{A}^n$ if we have

$$N(a,b|x^n,y^n) = N(a|x^n)V(b|a) \qquad \forall a \in \mathcal{A}, b \in \mathcal{B}.$$

V-shell: For any given $x^n \in \mathcal{A}^n$ and stochastic matrix V, the set of all sequences $y^n \in \mathcal{B}^n$ having conditional type V given x^n is called V-shell of x^n and is denoted by $\mathcal{T}_V^n(x^n)$.

- $\mathcal{T}_V^n(x^n)$ is uniquely defined, even if $N(a|x^n) = 0$ for some a.

Notation:

- $H(V|P) = \sum_{a\in\mathcal{A}} P(a)H(V(\cdot|a))$ denotes the average of entropies of rows of V with respect to a distribution P on $\mathcal{A}$.
- $D(V||W|P) = \sum_{a\in\mathcal{A}} P(a)D(V(\cdot|a)||W(\cdot|a))$ denotes the average of divergences between rows of stochastic matrices V and W with respect to a distribution P on $\mathcal{A}$.

- Most previous results can be straightforwardly generalized.
  - The number of sequences in a V-shell of x^n is given by:

Lemma: For every $x^n \in \mathcal{A}$ and stochastic matrix $V : \mathcal{A} \to \mathcal{B}$ such that $\mathcal{T}_V^n(x^n) \neq \emptyset$ with $|\mathcal{A}| = M$ and $|\mathcal{B}| = N$ we have

$$(n+1)^{-MN} 2^{nH(V|\hat{P}_{x^n})} \leq |\mathcal{T}_V^n(x^n)| \leq 2^{H(V|\hat{P}_{x^n})}$$

Proof: Note, $|\mathcal{T}_V^n(x^n)|$ depends on x^n only through its type $\hat{P}_{x^n}$! Thus consider vector

$$x^n = (\underbrace{a_1, \dots, a_1}_{N(a_1|x^n) \text{ times}}, \underbrace{a_2, \dots, a_2}_{N(a_2|x^n) \text{ times}}, \dots, \underbrace{a_M \dots, a_m}_{N(a_M|x^n) \text{ times}}).$$

Then $\mathcal{T}_V^n(x^n)$ is Cartesian product of sets of sequences of type $V(\cdot|a_i)$ on $\mathcal{B}^{N(a_i|x^n)}$. For each a_i apply previous lemma $\prod_{a_i \in \mathcal{A}} (N(a_i|x^n)+1)^{-N} 2^{N(a_i|x^n)H(V(\cdot|a_i))} \leq |\mathcal{T}_V^n(x^n)|$
$\leq \prod_{a_i \in \mathcal{A}} 2^{N(a_i|x^n)H(V(\cdot|a_i))}$. $\square$

- ... the *inaccuracy* result becomes:

Lemma: For every $x^n \in \mathcal{A}$ and stochastic matrices $V : \mathcal{A} \to \mathcal{B}$ and $W : \mathcal{A} \to \mathcal{B}$ such that $\mathcal{T}_V^n(x^n) \neq \emptyset$ we have

$$W^n(y^n|x^n) = 2^{-n(D(V||W|\hat{P}_{x^n})+H(V|\hat{P}_{x^n}))} \quad \text{if } y^n \in \mathcal{T}_V(x^n)$$

$$\frac{1}{(n+1)^{NM}} 2^{-nD(V||W|\hat{P}_{x^n})} \leq W^n(\mathcal{T}_V^n(x^n)|x^n) \leq 2^{-nD(V||W|\hat{P}_{x^n})}.$$

Proof idea: The results can be similarly deduced from the previous results as previously using

$$W^n(y^n|x^n) = \prod_{a_i \in \mathcal{A}, b_j \in \mathcal{B}} W(b_j|a_i)^{N(a_i,b_j|x^n,y^n)}.$$

□

Typical sequences

- Coding results can be derived using types (see Wolfowitz).
- Recently they are more often derived using the concept of (strong) typical sequences, which is very closely related to the concept of types.[11]

Typical sequences: For any distribution P on $\mathcal{A}$, a sequence $x^n \in \mathcal{A}^n$ is called P-*typical* with constant δ if

$$\left| \frac{1}{n} N(a|x^n) - P(a) \right| \leq \delta \qquad \forall a \in \mathcal{A}.$$

[11] More can be learned about this in our other graduate courses *Information Theory*, *Multiuser Information Theory*, or *Information-theoretic Security*.