Method of Types and Large Deviation Theory

Saurabh Khanna,

Signal Processing for Communication, ECE, IISc

Outline

- Method of types
  - Definitions
  - Basic properties

- Large deviation theory
  - Sanov's theorem
  - Conditional limit theorem

References

- Information Theory and Statistics, Chapter-12 of Elements in Information Theory, Cover and Thomas
- Short course on Information Theory and Statistics, Mauro Barni, Univ. of Siena, Italy

Method of types (MoT)

Type or empirical probability distribution

- Let $\mathbf{x}^n = (x_1, x_2 \ldots x_n)$ be n length sequence drawn from the alphabet set $\mathcal{A}$
- Alphabet set $\mathcal{A} = \left\{a_1, a_2 \ldots a_{|\mathcal{A}|}\right\}$
- Type or empirical probability distribution of the seq $\mathbf{x}^n$:

$$P_{\mathbf{x}^n}(a) = \frac{N(a \mid \mathbf{x}^n)}{n}, \quad a \in \mathcal{A}$$

  - Type $P_{\mathbf{x}^n}(a)$ is a pmf on $\mathcal{A}$
- Example: $\mathcal{A} = \{0, 1\}$ and, $\mathbf{x}^8 = (0, 0, 1, 0, 1, 1, 0, 0)$

$$\text{Type} P_{\mathbf{x}^8} = \left(\frac{5}{8}, \frac{3}{8}\right)$$

Type or empirical probability distribution

- Set $\mathcal{P}_n$ contains all possible types (empirical probability distributions) for n length sequences

- Example: Say, $\mathcal{A} = \{0, 1\}$

$$\text{Then, } \mathcal{P}_n = \left\{ \left(\frac{0}{n}, \frac{n}{n}\right), \left(\frac{1}{n}, \frac{n-1}{n}\right) \dots \left(\frac{n}{n}, \frac{0}{n}\right) \right\}$$

Type class

- Type class: set of all sequences of same type

$$T(P) = \left\{ \mathbf{x}^n \in \mathcal{A}^n \text{ such that } P_{\mathbf{x}^n} = P \right\}$$

- Example: Say, $\mathcal{A} = \{0, 1\}$, $n = 5$ and $P = \left(\frac{3}{5}, \frac{2}{5}\right)$

$$T(P) = \left\{ \begin{array}{c} (1,1,0,0,0), (1,0,1,0,0), (1,0,0,1,0), (1,0,0,0,1), \\ (0,1,1,0,0), (0,1,0,1,0), (0,1,0,0,1), (0,0,1,1,0), \\ (0,0,1,0,1), (0,0,0,1,1) \end{array} \right\}$$

  - All sequences in type class $T(P)$ are permutations of one another

Size of a type class

- For a type class $P \in \mathcal{P}_n$, its size is given by

$$
\begin{aligned}
|T(P)| &= \text{No. of n length sequences of type P} \\
&= \frac{n!}{(nP(a_1)!)(nP(a_2)!)\dots(nP(a_{|\mathcal{A}|})!)}
\end{aligned}
$$

- Exact size is diffcult to work with

- Exponential upper and lower bounds exist for $|T(P)|$

$$
\frac{1}{(n+1)^{|\mathcal{A}|}} \cdot 2^{nH(P)} \leq |T(P)| \leq 2^{nH(P)}
$$

Number of types

- The number of types grows polynomially with n

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

- Proof: Trivial

Observations

- For fixed n,
  1. The number of sequences is exponential in n
  2. There are only polynomial number of types
- There is at least one type $P \in \mathcal{P}_n$ with exponential many sequences in the type class $T(P)$
- As $n \to \infty$, the largest type class has essentially the same number of sequences as the entire set of sequences (upto first order in exponent)

Why is MoT useful?

- As n increases, a structure is revealed about the set of types associated with observed sequences
- Some types are observed much more frequently than others
- MoT is useful in expressing the properties of an observed sequence in terms of its type.

Probability of a sequence

- The probability of a sequence $\mathbf{x}^n$ emitted by a DMS with pmf $Q : \mathcal{A} \to [0, 1]$ is given by

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

- Proof: We work out

- Sequences whose type does not match with Q (in KL diverence sense) are exponential less likely to occur

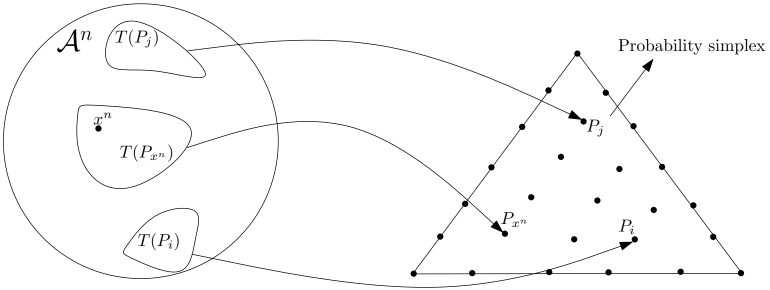
Probability of a type class

- The probability that a DMS emits a sequence belonging to type class $T(P)$ can be bounded as:

$$\frac{1}{(n+1)^{|\mathcal{A}|}} \cdot 2^{-nD(P||Q)} \leq Q^n(T(P)) \leq 2^{-nD(P||Q)}$$

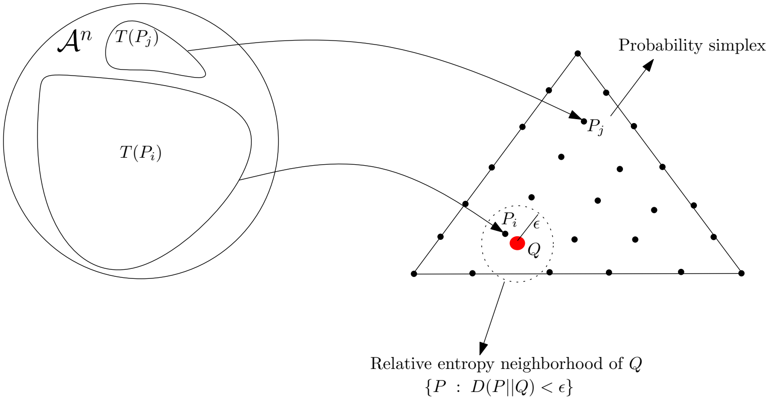
- Proof: We work out

Summary of main results



1. $|\mathcal{P}^n| \leq (n+1)^{|\mathcal{A}|}$ — Polynomial number of types
2. $Q^n(\mathbf{x}^n) = 2^{-n(H(P)+D(P||Q))}$ — Exact prob. of seqn of type P under Q
3. $|T(P)| \approx 2^{nH(P)}$ — Approx no. of sequence of each type
4. $Q^n(T(P)) \approx 2^{-nD(P||Q)}$ — Approx. prob. of type T(P) under Q

For large n



$\mathcal{A}^n$
$T(P_j)$
$T(P_i)$
Probability simplex
P_j
P_i
ϵ
Q
Relative entropy neighborhood of Q
$\{P\ :\ D(P||Q) < \epsilon\}$

Weak law of large numbers

Typical set

- Probability of n length sequence belonging to type class $T(P)$
$$Q^n(T(P)) \approx 2^{-nD(P||Q)}$$

- Sequences of type P with large $D(P||Q)$ are exponentially less likely to occur

- Sequences of type P within small relative entropy distance of source Q occur with very high probability

- We define a typical set of sequences T_Q^ϵ as
$$T_Q^\epsilon = \left\{ \mathbf{x}^n \in T(P) \mid P \in \mathcal{P}_n \text{ and } D(P||Q) \leq \epsilon \right\}$$

- As $n \to \infty$, $\mathbb{P}(\mathbf{x}^n \notin T_Q^\epsilon) \to 0$

Law of large numbers (MoT perspective)

- We show that as $n \to \infty$, $\mathbb{P}(\mathbf{x}^n \notin T_Q^\epsilon)$ tends to 0

$$
\begin{aligned}
\mathbb{P}(\mathbf{x}^n \notin T_Q^\epsilon) &= \sum_{P\,:\,D(P||Q)>\epsilon} Q^n\left(T(P)\right) \\
&\leq \sum_{P\,:\,D(P||Q)>\epsilon} 2^{-nD(P||Q)} \\
&\leq \sum_{P\,:\,D(P||Q)>\epsilon} 2^{-n\epsilon} \\
&\leq \sum_{\mathcal{P}_n} 2^{-n\epsilon} \\
&\leq (n+1)^{|\mathcal{A}|} 2^{-n\epsilon} \\
&\leq 2^{-n\left(\epsilon - \frac{|\mathcal{A}| log(n+1)}{n}\right)} \underset{n \longrightarrow \infty}{\rightarrow} 0
\end{aligned}
$$

Large deviation theory

Large deviation theory (LDT)

- LDT studies the probability of rare events i.e. events not covered by law of large numbers
- Examples:
  - What is the probability that 800 times head occurs in 1000 fair coin tosses?
  - What is the probability that mean of a sequence (emitted by DMS X) is larger than T, where T is much larger than $E(X)$

Large deviation theory (LDT)

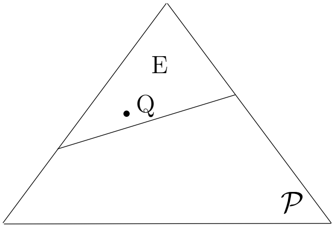
- A more general question answered by LDT:

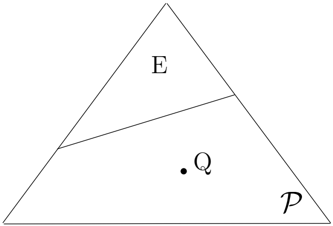
  Let E be a subset of pmf's and let Q be the source distribution. Then, what is the probability that Q emits a sequence whose type belongs to E

- In other words, LDT talks about $Q(E) = \sum_{\mathbf{x}^n : P_{\mathbf{x}^n} \in E} Q^n(\mathbf{x}^n)$

Large deviation theory (LDT)

- If E contains a relative entropy neighborhood of Q, then $Q(E) \to 1$
- If E does not contain Q, then $Q(E) \to 0$. The question is: how fast ?





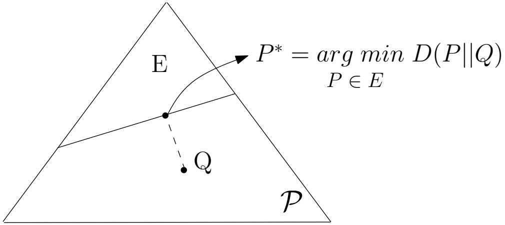
Sanov's theorem

- Let $x_1, x_2 \ldots x_n$ be i.i.d. $Q(x)$
- If $E \subset \mathcal{P}_n$ be a closed convex set of probability distributions
- Then,

$$Q^n(E) \approx 2^{-nD(P^*||Q)}$$

where

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



Example of Sanov's theorem

- Consider $x_1, x_2 \dots x_n$ to be emitted by DMS according to pmf Q

- Question:

  What can we say about $\mathbb{P}(\frac{1}{n}\sum_{i=1}^{n} g_j(x_i) \leq \alpha_j,\ j = 1, 2, \dots k)$?

- Define set E as

$$E = \left\{ P : \sum_a P(a) g_j(a) \leq \alpha_j,\ j = 1, 2 \dots k \right\}$$

- We find closest distribution $P^* \in E$ to Q

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

- From Sanov's theorem, desired probability is $\approx\ 2^{-nD(P^*||Q)}$

Example of Sanov's theorem

- Finding $P^* \in E$ closest to Q is a constrained convex optimization problem

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

- Solved using Lagrangian multipliers method:

$$\begin{aligned} L(P, \boldsymbol{\lambda}, \nu) &= \sum_{a \in \mathcal{A}} P(a) \log \frac{P(a)}{Q(a)} + \\ &\sum_{j=1}^{k} \lambda_j \left(\alpha_j - \sum_{a \in \mathcal{A}} P(a) g_j(a) \right) + \nu \left(\sum_{a \in A} P(a) - 1 \right) \end{aligned}$$

- $P^*(a) = \frac{1}{Z} Q(a) e^{\sum_{j=1}^{k} \lambda_j g_j(a)}, \quad a \in \mathcal{A}$

Conditional limit theorem

- Let E be a closed convex subset of $\mathcal{P}_n$
- Let $x_1, x_2 \ldots x_n$ be i.i.d. $\quad Q(x) \notin E$
- Then, as $n \longrightarrow \infty$

$$\mathbb{P}(x_1 = a \mid P_{\mathbf{x}^n} \in E) \underset{p}{\longrightarrow} P^*(a)$$

where

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