EE376A/STATS376A Information Theory Lecture 13 - 02/20/2018

Lecture 13: Method of Types

Lecturer: Tsachy Weissman *Scribe: Fang Cai, Rob Jones, Yi Sun, Can Wang*

In last lecture we framed the problem of lossy compression and gave the main theorem that characterizes the tradeoff between the rate and the distortion in the context of lossy compression. With that as our motivation, this week we are going to talk about the method of types, expand our tools related to typical sequences, the notion of strong typicality and the notion of conditional types, which are not only interesting in their own right, but also serve very well when we go back to establish the main result in lossy compression. We are also going to talk about some concrete schemes for lossy compression and how they are related to clustering and machine learning problems. Today we are going to talk about Method of Types.

1 Notation

Denote: $x^n = \{x_1, \ldots, x_n\}$ with $x_i \in \mathcal{X} = \{1, \ldots, r\}$ and

$$N(a|x^n) = \sum_{i=i}^{n} \mathbb{I}_{\{x_i = a\}},$$

$$\mathrm{P}_{x^n}(a) = \frac{N(a|x^n)}{n}.$$

2 Empirical distribution and type class

Definition 1 (Empirical distribution, type class)**.** *The **empirical distribution** of x^n is the probability vector $(\mathrm{P}_{x^n}(1), \ldots, \mathrm{P}_{x^n}(r))$. $\mathbb{P}_n$ denotes the collection of all empirical distribution of sequences of length n, i.e. $\mathbb{P}_n = \{\mathrm{P}_{x^n} : x^n \in \mathcal{X}^n\}$. For $\mathrm{P} \in \mathbb{P}_n$, the **type class** or **type** of P is $T(\mathrm{P}) = \{x^n : \mathrm{P}_{x^n} = \mathrm{P}\}$. The type class of x^n is $T_{x^n} = T(\mathrm{P}_{x^n}) = \{\tilde{x}^n : \mathrm{P}_{\tilde{x}^n} = \mathrm{P}_{x^n}\}$.*

Example 2. If $\mathcal{X} = \{0, 1\}$, then $\mathbb{P}_n = \left\{(1, 0), \left(\frac{n-1}{n}, \frac{1}{n}\right), \left(\frac{n-2}{n}, \frac{2}{n}\right), \ldots, (0, 1)\right\}$
Example 3. If $\mathcal{X} = \{a, b, c\}$, $n = 5$ and $x^n = (a, a, c, b, a)$, then $\mathrm{P}_{x^n} = \left(\frac{3}{5}, \frac{1}{5}, \frac{1}{5}\right)$,
$T_{x^n} = \{(a, a, a, b, c), (a, a, a, c, b), \ldots, (c, b, a, a, a)\}$ and $|T_{x^n}| = \binom{5}{3\ 1\ 1} = \frac{5!}{3!\ 1!\ 1!} = 20$.

In the following we show that the number of different type classes induced by x^n, $|\mathbb{P}_n|$, can be upper bounded by something which is polynomial in n, which doesn't increase exponentially with n.

Theorem 4. $|\mathbb{P}_n| \leq (n+1)^{r-1}$.

Proof
Every empirical distribution P_{x^n} is determined by vector $(N(1|x^n), N(2|x^n), \cdots, N(r-1|x^n))$, where $N(a|x^n)$ means the number of times that the symbol a appears in the sequence x^n. Since $0 \leq N(a|x^n) \leq n$, each of $N(a|x^n)$ can take on no more than n+1 values.
Thus we have a vector of length $r-1$ and each element can take no more than n+1 values. Therefore there are at most $(n+1)^{r-1}$possibilities. □

Note that for the case $r = 2$, the bound is tight. But the bound is not tight for the cases $r \geq 3$ because we didn't incorporate the constraint that $\sum_{a=1}^{r-1} N(a|x^n)$ must be less than or equal to n when calculating the upper bound.

Further notation

- For probability mass function(PMF) $\mathrm{Q} = \{\mathrm{Q}(x)\}_{x\in\mathcal{X}}$, we write $H(\mathrm{Q})$ for $H(X)$ when X is distributed according to Q.
- $\mathrm{Q}(x^n) = \prod_{i=1}^n \mathrm{Q}(x_i)$. For $S \subset \mathcal{X}^n$, $\mathrm{Q}(S) = \sum_{x^n\in S} \mathrm{Q}(x^n)$

Theorem 5. *$\forall x^n, \mathrm{Q}(x^n) = 2^{-n[H(\mathrm{P}_{x^n})+D(\mathrm{P}_{x^n}\,||\,\mathrm{Q})]}$, where $H(\mathrm{P}_{x^n})$ is referred to as empirical entropy of x^n.*

Proof

$$\begin{aligned}
\mathrm{Q}(x^n) &= \prod_{i=1}^n \mathrm{Q}(x_i) \\
&= 2^{\sum_{i=1}^n \log \mathrm{Q}(x_i)} \\
&= 2^{\sum_{a\in\chi} N(a|x^n)\log \mathrm{Q}(a)} \\
&= 2^{-n\left[\sum_{a\in\chi} \frac{N(a|x^n)}{n} \log \frac{1}{\mathrm{Q}(a)}\right]} \\
&= 2^{-n\left[\sum_{a\in\chi} \mathrm{P}_{x^n}(a) \log\left(\frac{1}{\mathrm{Q}(a)} \frac{\mathrm{P}_{x^n}(a)}{\mathrm{P}_{x^n}(a)}\right)\right]} \\
&= 2^{-n[H(\mathrm{P}_{x^n})+D(\mathrm{P}_{x^n}\,\|\,\mathrm{Q})]}
\end{aligned}$$

□

The next result is about the size of the type class associated with the empirical distribution P.

Theorem 6. $\forall\, \mathrm{P} \in \mathbb{P}_n$, $\frac{1}{(n+1)^{r-1}} 2^{nH(\mathrm{P})} \le |T(\mathrm{P})| \le 2^{nH(\mathrm{P})}$.

Note: We could calculate the size of type class $|T(\mathrm{P})|$ exactly, which is

$$|T(\mathrm{P})| = \binom{n}{n\cdot\mathrm{P}(1), n\cdot\mathrm{P}(2), \cdots, n\cdot\mathrm{P}(r)}.$$

But for our purposes, what we care about are (1) the behavior of this quantity for n large on an exponential scale and (2) how it is related to quantities that are familiar and important to us, such as entropy.

Proof of upper bound in Theorem 6:

$$\begin{aligned}
1 &\ge \mathrm{P}(T(\mathrm{P})) \\
&= \sum_{x^n\in T(\mathrm{P})} \mathrm{P}(x^n) \\
&= \sum_{x^n\in T(\mathrm{P})} 2^{-n[H(\mathrm{P}_{x^n})+D(\mathrm{P}_{x^n}\,||\,\mathrm{P})]} && \text{(by Theorem 5, with Q=P)} \\
&= \sum_{x^n\in T(\mathrm{P})} 2^{-n[H(\mathrm{P})+D(\mathrm{P}\,||\,\mathrm{P})]} && \text{(all elements } x^n \in T(P) \text{ have empirical distribution P)} \\
&= |T(\mathrm{P})| \cdot 2^{-nH(\mathrm{P})}
\end{aligned}$$

Thus by simple algebraic manipulation we have:

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

□

Before proving the lower bound, we prove two lemmas.

Lemma 7. *For non-negative integers m, n, $\frac{m!}{n!} \geq n^{m-n}$.*

Proof:
If $m \geq n$,

$$\frac{m!}{n!} = \underbrace{m(m-1)\cdots(n+1)}_{(m-n) \text{ factors, each} \geq n} \geq n^{m-n}.$$

If $m < n$,

$$\frac{m!}{n!} = \frac{1}{\underbrace{n(n-1)\cdots(m+1)}_{(n-m) \text{ factors, each} \leq n}} \geq \frac{1}{n^{n-m}} = n^{m-n}.$$

□

Lemma 8. $\forall \mathrm{P}, \mathrm{Q} \in \mathbb{P}_n, \mathrm{P}(T(\mathrm{P})) \geq \mathrm{P}(T(\mathrm{Q}))$.

Proof:

$$\begin{aligned}
\frac{\mathrm{P}(T(\mathrm{P}))}{\mathrm{P}(T(\mathrm{Q}))} &= \frac{|T(\mathrm{P})|(\prod_{a\in\mathcal{X}} \mathrm{P}(a)^{n\,\mathrm{P}(a)})}{|T(\mathrm{Q})|(\prod_{a\in\mathcal{X}} \mathrm{P}(a)^{n\,\mathrm{Q}(a)})} \\
&= \frac{\binom{n}{n\,\mathrm{P}(1), n\,\mathrm{P}(2), \cdots, n\,\mathrm{P}(r)}}{\binom{n}{n\,\mathrm{Q}(1), n\,\mathrm{Q}(2), \cdots, n\,\mathrm{Q}(r)}} \prod_{a\in\mathcal{X}} \mathrm{P}(a)^{n\,\mathrm{P}(a) - n\,\mathrm{Q}(a)} \\
&= \prod_{a\in\mathcal{X}} \frac{(n\,\mathrm{Q}(a))!}{(n\,\mathrm{P}(a))!} \mathrm{P}(a)^{n[\mathrm{P}(a)-\mathrm{Q}(a)]} \\
&\geq \prod_{a\in\mathcal{X}} (n\,\mathrm{P}(a))^{n\,\mathrm{Q}(a) - n\,\mathrm{P}(a)} \mathrm{P}(a)^{n[\mathrm{P}(a)-\mathrm{Q}(a)]} && \text{(by Lemma 7)} \\
&= \prod_{a\in\mathcal{X}} n^{n[\mathrm{Q}(a)-\mathrm{P}(a)]} \\
&= n^{n\sum_{a\in\mathcal{X}}(\mathrm{Q}(a)-\mathrm{P}(a))} = 1
\end{aligned}$$

□

Proof of lower bound in Theorem 6:

$$\begin{aligned}
1 &= \mathrm{P}(\mathcal{X}^n) \\
&= \sum_{\mathrm{Q}\in\mathbb{P}_n} \mathrm{P}(T(\mathrm{Q})) \\
&\leq |\mathbb{P}_n| \cdot \max_{\mathrm{Q}\in\mathbb{P}_n} \mathrm{P}(T(\mathrm{Q})) \\
&= |\mathbb{P}_n| \cdot \mathrm{P}(T(\mathrm{P})) && \text{(by Lemma 8)} \\
&= |\mathbb{P}_n| \cdot |T(\mathrm{P})| \cdot 2^{-n[H(\mathrm{P})+D(\mathrm{P}\,||\,\mathrm{P})]} && \text{(by Theorem 5)} \\
&\leq (n+1)^{r-1} \cdot |T(\mathrm{P})| \cdot 2^{-nH(\mathrm{P})} && \text{(by Theorem 4)}
\end{aligned}$$

Thus by simple algebraic manipulation we have:

$$\frac{1}{(n+1)^{r-1}} \cdot 2^{nH(\mathrm{P})} \leq |T(\mathrm{P})|$$

□

Noting that by Theorem 5, we have that, for any probability mass function Q and any empirical distribution $\mathrm{P} \in \mathbb{P}_n$,

$$\mathrm{Q}(T(\mathrm{P})) = |T(\mathrm{P})| 2^{-n[H(\mathrm{P}) + D(\mathrm{P} \,||\, \mathrm{Q})]}.$$

Together with Theorem 6, we obtain the following theorem.

Theorem 9. $\forall PMF\ \mathrm{Q}, \forall \mathrm{P} \in \mathbb{P}_n, \frac{1}{(n+1)^{r-1}} 2^{-nD(\mathrm{P} \,||\, \mathrm{Q})} \leq Q(T(\mathrm{P})) \leq 2^{-nD(\mathrm{P} \,||\, \mathrm{Q})}$.

This shows that up to an insignificant polynomial factor ($\frac{1}{(n+1)^{r-1}}$), on an exponential scale, the probability that the sequence looks like it came from source P, if the data is generated i.i.d. from distribution Q, is exponentially unlikely. The farther away P is from Q, the more unlikely it is. Note in the expression above, the relative entropy $D(\mathrm{P} \,||\, \mathrm{Q})$ is between P, the "wrong" source, and Q, the true source, unlike in the cost of mismatch in lossless compression $D(p||q)$ (see lecture 6), p is the true source while q is the "wrong" source.