EE376A Information Theory Lecture 1 - 02/24/2015

Lecture 14

Lecturer: Tsachy Weissman *Scribe: Sunil Pai, Han Altae-Tran, and Vasu Gupta*

1 Recap on Types

Last lecture, we discussed the Method of Types. This lecture, we will briefly review this method, and prove a version of Sanov's Theorem, which bounds the probability that a function's empirical mean exceeds some value α. Consider the sequence $x^n \in \mathcal{X}^n$, where $\mathcal{X}$ is a finite alphabetl Let P_{x^n} be the empirical distribution and $\mathscr{P}_n$ the set of all empirical distributions over sequences of length n. Then we define the type to be:

$$T(P) = \{x^n : P_{x^n} = P\},$$

for $P \in \mathscr{P}_n$. In the previous lecture, we have shown that:

1. $|\mathscr{P}_n| \leq (n+1)^{|x|}$
2. $Q^n(x^n) = 2^{-n[H(P_{x^n})+D(P_{x^n}||Q)]}$
3. $\frac{1}{(n+1)^{|x|}} 2^{nH(p)} \leq T(P) \leq 2^{nH(p)}$
4. $\frac{1}{(n+1)^{|x|}} 2^{-nD(P_{x^n}||Q)} \leq Q(T(P)) \leq 2^{-nD(P_{x^n}||Q)}$

2 A Version of Sanov's Theorem

Theorem 1. *Sanov's Theorem. For sufficiently large n, we have*

$$\frac{1}{(n+1)^{|x|}} 2^{-n \min D(P||Q)} \leq \left(\frac{1}{n} \sum_{i=1}^{n} f(X_i)) \geq \alpha \right) \leq (n+1)^{|x|} 2^{-n \min D(P||Q)},$$

where the min *is over the set* $\{P : P \in \mathcal{P}_n, \langle P_{x^n}, f \rangle \geq \alpha\}$.

Proof First observe that we can write

$$\begin{aligned} \frac{1}{n} \sum_{i=0}^{n} f(x) &= \frac{1}{n} \sum_{a \in \mathcal{X}} N(a|x^n) f(a) \\ &= \sum_{a \in \mathcal{X}} P_{x^n}(a) f(a) \\ &= \langle P_{x^n}, f \rangle \end{aligned}$$

where we have used the Euclidean inner product, defined as $\langle a, b \rangle := \sum_{i=1}^{n} a_i b_i$ for $a, b \in \mathbf{R}$. Then by the Law of Large numbers,

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^{n} f(X_i) &\approx \mathbf{E}_{X \sim Q} f(X) \\ &= \sum_{a \in \mathcal{X}} Q(a) f(a) \\ &= \langle Q, f \rangle \end{aligned}$$

We can proceed to find the desired upper bound.

$$\begin{aligned} P\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i) \geq \alpha\right) &= P(\langle P_{x^n}, f\rangle) \\ &= Q^n\left(\bigcup_{P\in\mathcal{P}_n, \langle P_{x^n}, f\rangle \geq \alpha} T(P)\right) \\ &= \sum_{P\in\mathcal{P}_n, \langle P_{x^n}, f\rangle \geq \alpha} Q^n(T(P)) \\ &\leq |\mathcal{P}_n| \max Q^n(T(P)) \\ &\leq (n+1)^{|x|} \max 2^{-nD(P||Q)} \\ &= (n+1)^{|x|} 2^{-n \min D(P||Q)} \end{aligned}$$

Now we solve for the lower bound:

$$\begin{aligned} P\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i) \geq \alpha\right) &\geq \max Q^n(T(P)) \\ &\geq \max \frac{1}{(n+1)^{|x|}} 2^{-nD(P||Q)} \\ &= \frac{1}{(n+1)^{|x|}} 2^{-n \min D(P||Q)} \end{aligned}$$

Note that we are taking the $\min, \max$ over the set $\{P : P \in \mathcal{P}_n, \langle P_{x^n}, f\rangle \geq \alpha\}$. Therefore, we have, up to a polynomial in n, that $P\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i) \geq \alpha\right) \doteq 2^{-nD^*(\alpha)}$ where $D^*(\alpha) = \min D(P||Q)$. This is exactly waht we were looking for. □

Example 2. Take $X_i \sim \text{Ber}\left(\frac{1}{2}\right)$. Then:

$$P\left(\text{fraction of ones in } X_1X_2\cdots X_n \geq \alpha\right) = P\left(\frac{1}{n}\sum_{i=1}^{n} X_i \geq \alpha\right) \doteq 2^{-nD^*(\alpha)}$$

where $D^*(\alpha) = \min D(\text{Ber}(\alpha)||\text{Ber}(1/2))$. This gives:

$$D^*(\alpha) = \begin{cases} 0 & 0 \leq \alpha \leq \frac{1}{2} \\ D(\text{Ber}(\alpha)||\text{Ber}(1/2)) & \frac{1}{2} < \alpha \leq 1 \\ \infty & \alpha > 1 \end{cases}$$

and since

$$\begin{aligned} D(\text{Ber}(p)||\text{Ber}(1/2)) &= \alpha \log \frac{\alpha}{1/2} + (1-\alpha)\log\frac{1-\alpha}{1/2} \\ &= 1 - h(\alpha) \end{aligned}$$

where $h(\cdot)$ is the binary entropy function. Thus we can write

$$D^*(\alpha) = \begin{cases} 0 & 0 \leq \alpha \leq \frac{1}{2} \\ 1 - h(\alpha) & \frac{1}{2} < \alpha \leq 1 \\ \infty & \alpha > 1 \end{cases}$$

Interestingly, this function explodes at $\alpha = 1$ which makes sense because the probability that the mean of random variables which take values up to 1 is greater than 1 is impossible. Furthermore, we have a region where the cost of mismatch is zero since we are guaranteed that one of the probabilities is always going to be $\geq 1/2$ so we would expect our mean to be so as well.

Definition 3. *A sequence x^n is said to be strongly δ typical with P if for every $a \in \mathcal{X}$,*

$$|P_{x^n}(a) - P(a)| \leq \delta P(a)$$

Definition 4. *The strongly δ-typical set, $T_\delta(P)$, is the set of all strongly δ typical sequences. That is*

$$T_\delta(P) = \{x^n : |P_{x^n}(a) - P(a)| \leq \delta P(a)\}$$

For reference, recall that the weakly ϵ-typical set is:

$$A_\epsilon(P) = \left\{ x^n : \left| -\frac{1}{n} \log P(x^n) - H(P) \right| \leq \epsilon \right\}$$

In the homework, you will show that $\forall\, \delta > 0$ there exists $\epsilon = \delta H(p)$ such that $T_\delta(P) \subseteq A_\epsilon(P)$.

Example 5. Consider the following extreme example where $P(a) = \frac{1}{|\mathcal{X}|}$ is uniform with $a \in \mathcal{X}$. So for all x^n,

$$\begin{aligned} P(x^n) &= \frac{1}{|\mathcal{X}|^n} \\ &= 2^{-n \log |\mathcal{X}|} \\ &= 2^{-nH(p)} \end{aligned}$$

Therefore, for all $\epsilon > 0$, $A_\epsilon^{(n)}(P) = \mathcal{X}^n$! This makes sense because this is a uniform distribution so you would always expect the typical sequence to include all possibilities regardless of n. In the homework, you will *also* show that there exists $\epsilon(\delta)$ such that for all n sufficiently large:

$$2^{n[H(P)-\epsilon(\delta)]} \leq |T_\epsilon(P)| \leq 2^{n[H(P)+\epsilon(\delta)]}$$

where $\epsilon(\delta) \to 0$ as $\delta \to 0$.

Definition 6. *For $x^n = (x_1, x_2, \ldots x_n) \in \mathcal{X}^n$ and $y^n = (y_1, y_2, \ldots y_n) \in \mathcal{Y}^n$, the joint empirical distribution is defined as*

$$\begin{aligned} P_{x^n, y^n}(x, y) &= \frac{1}{n} |\{i \in \{1, \ldots, n\} : x_i = x, y_i = y\}| \\ &= \frac{1}{n} N(x, y \mid x^n, y^n) \end{aligned}$$

where we have defined $N(x, y \mid x^n, y^n) := |\{i \in \{1, \ldots, \} : x_i = x, y_i = y\}|$

Then, we can make the following definition.

Definition 7. *For $(X, Y) \sim P$, the jointly δ typical set is given by*

$$T_\delta(P) = \{(X^n, Y^n) : |P_{x^n, y^n}(x, y) - P(x, y)| < \delta P(x, y) \forall (x, y) \in \mathcal{X} \times \mathcal{Y}\}$$

Often as is the case with $H(X)$ vs $H(P)$, we will write $T_\delta(X)$ for $T_\delta(P)$ when $P \sim X$ and $T_\delta(X, Y)$ for $T_\delta(P)$ when $P \sim (X, Y)$. The notion of strong vs weak typicality is important. For example, consider the random variable $G_n = \frac{1}{n} \sum_{i=1}^n g(x_i)$. If X^n is strongly typical, then G_n is close to $\mathbf{E}[g(x)]$ for large n. On the other hand, this would not necessarily have been the case if X^n were only weakly typical.