Math 254A Lecture 1 Notes

Daniel Raban

March 29, 2021

1 Counting Type Classes and Introduction to Shannon Entropy

1.1 Counting type classes

Here is a basic setting we will be working with:

- A is a finite alphabet.
- $P(A) = \{p : A \to \mathbb{R} : p(a) \geq 0, \sum_a p(a) = 1\} \subseteq \mathbb{R}^A$ is the set of probability mass functions on A.
- $\|p - q\| = \sum_a |p(a) - q(a)| = 2 \sup_{B \subseteq A} |p(B) - q(B)|$ is the total variation between p and q.
- If $x \in A^n$ (for $n \in \mathbb{N}$), then $N(a \mid x) = |\{i = 1, \dots, n : x_i = a\}|$ is the number of occurrences of a in x.

Definition 1.1. The empirical distribution of x is $p_x(a) = \frac{N(x|x)}{n}$.

Definition 1.2. Given $p \in P(A)$, the **type class** of p is $T_n(p) = \{x \in A^n : p_x = p\}$.

How big is $|T_n(p)|$? Here is a basic answer:

$$|T_n(p)| = \begin{cases} \frac{n!}{(np(a))! \cdots (np(a_k))!} & np(a) \in \mathbb{N} \forall a \in A \\ 0 & \text{otherwise} \end{cases}, \qquad A = \{a_1, \dots, a_k\}$$

We are interested in the exponential asymptotic behavior of $|T_n(p)|$. Stirling's approximation tells us that

$$n! = \frac{n^n}{e^n} \sqrt{2\pi n} e^{o(1)}$$

as $n \to \infty$ (where $e^{o(1)} \to 1$ as $n \to \infty$). We will write this more crudely as

$$n! = \frac{n^n}{e^n} e^{o(n)}.$$

Inserting this into the previous expression gives

$$\begin{aligned}
|T_n(p)| &= \frac{(n^n/e^n)e^{o(n)}}{\prod_{i=1}^k((np(a_i))^{np(a_i)}/e^{np(a_i)})e^{o(n)}} \\
&= \frac{n^n}{\prod_i (np(a_i))^{np(a_i)}} \cdot \frac{e^n}{\prod_i e^{np(a_i)}} \\
&= \frac{e^{n\log n}}{\exp(\sum_i np(a_i)\log np(a_i))} \\
&= \exp\left(n\log n - \sum_i np(a_i)\log(np(a_i))\right) \\
&= \exp\left(n\log n - \sum_i np(a_i)\log n - n\sum_i p(a_i)\log p(a_i)\right).
\end{aligned}$$

In total, we have

$$\begin{aligned}
|T_n(p)| &= e^{-n\sum_i p(a_i)\log p(a_i)+o(n)} \\
&= e^{nH(p)+o(n)},
\end{aligned}$$

where $H(p) = -\sum_a p(a)\log p(a)$. This quantity is called the **Shannon entropy** of $p \in P(A)$.

Later on, high-level real analysis will allow us to make sense of redoing the above computation in more complicated variants of this problem, where we are not just looking at the empirical distribution.

Remark 1.1. We regard H as a function $P(A) \to \mathbb{R}$, with the convention that $0\log 0 = 0$.

1.2 Basic properties of Shannon entropy

Proposition 1.1. *The Shannon entropy H has the following properties:*

(a) H is continuous.

Proof. $x\log x$ is continuous for $x \in (0,1]$, and $x\log x \to 0$ as $x \to 0$. □

(b) H is strictly concave; i.e. $H(tp + (1-t)q) \geq tH(p) + (1-t)H(q)$ with equality only if either $p = q$ or $t \in \{0,1\}$.

Proof. The function $x \mapsto x\log x$ is strictly concave on $[0,1]$ (second derivative is < 0). For strictness, if $p \neq q$ and $0 < t < 1$, then there is some a such that $p(a) \neq q(a)$. Then

$$-(tp(a) + (1-t)q(a))\log(tp(a) + (1-t)q(a)) > -tp(a)\log p(a) - (1-t)q(a)\log q(a).$$

□

(c) *$H(p)$ is symmetric under permutations of A.*

(d) $0 \leq H(p) \leq \log |A|$. *Equality on the left is achieved iff $p - \delta_a$ for some $a \in A$, and quality of the right is achieved iff $p = (1/|A|, \ldots, 1/|A|)$.*

Proof. $-x \log x \geq 0$ and is > 0 unless $x = 0, 1$. So $H(p) \geq 0$, and equals 0 only if $p(a) \in \{0, 1\}$ for all a, i.e. only if $p = \delta_b$ for some b. On the other hand, by concavity and symmetry (properties (b) and (c)), H must be maximized at $p = (1/|A|, \ldots, 1/|A|)$, and then $H = \log |A|$. □

Example 1.1. Look at the image of H of the simplex $P(\{1,2,3\}) = \{(p_1, p_2, p_3) : p_i \geq 0, \sum_i p_i = 1\}$.

Remark 1.2. Suppose X is a random variable taking values in A, and let $p(a) = \mathbb{P}(X = a)$ for $a \in A$. Then $H(X) := H(p)$ is a canonical way to quantify the "uncertainty" in X.

Next time, we will loosen the counting problem to estimate the size of

$$T_{n,\delta}(p) = \{x \in A^n : \|p_x - p\| < \delta\}.$$