Universal Probability

\( \mathbb{P} : \) class of random processes

\( \text{stat.: e.g., Bernoulli; some parametric family, Markov, \ldots} \)

Goal: find \( f_n(x_1, \ldots, x_n) \) st.

\[
\lim_{n \to \infty} \frac{1}{n} D(p(x^n) \| f(x^n)) = 0, \quad \forall p \in \mathbb{P}
\]

Such \( f \) is said to be mean universal

Desirable properties

1. Non-asymptotic performance

We will show \( f_{\text{LZ}} \) is mean universal

\( \uparrow \)

Lempel-Ziv

For every stat. erg. process.

Let \( P \) be Bernoulli:

\[
\frac{1}{n} D(p_{\text{Bernoulli}} \| f_{\text{LZ}}) \to 0 \quad \forall \text{Bernoulli}
\]

Find a good \( f \), one that achieves
\[
\min_{\beta} \max_{\rho \in \mathcal{P}} \mathcal{D}(\rho \| \pi) \quad \text{minmax universal probability}
\]

(2) Horizon independence

Sequence \((g_n(x^n))_{n=1}^\infty\) should be consistent.

\[
\sum_{x^{n+1}} g_{n+1}(x_1, \ldots, x_{n+1}) = g_n(x_1, \ldots, x_n)
\]

\(\Rightarrow\) defines a random process on its own.

- \(g_{n+2}\) is horizon independent

\[
g_n(x^n) = \int \Theta^n (\Theta)^{1-k(x^n)} \, d\Theta
\]

Where \(k(x^n)\) is \# of 1's in \(x^n\).

This is Laplace's distribution

- Is consistent in \(n\)
- Has conditional form

\[
g(x_n | x_1, \ldots, x_{n-1}) = \frac{g(x_1, \ldots, x_n)}{g(x_1, \ldots, x_{n-1})}
\]

\[
= \frac{k+1}{n+2} \quad \text{when } k=n, \frac{n+1}{n+2}
\]

\[
g(x^n) = g(x_1) g(x_2 | x_1) \cdots g(x_n | x_{n-1})
\]
3. Low computational complexity

- Laplace mixture is easy to implement
- So is $q_{Lt}$ Consider "MCMC" methods?

4. Pointwise universality

$$\frac{1}{n} \log \frac{p(x^n)}{q(x^n)} \rightarrow 0 \quad (p - a.s.)$$

for every $p \in P$.

Such $g$ is pointwise universal w.r.t. $P$.

- Graph is pw universal
- $q_{Lt}$ is pw universal.

Application of universal Probability

If $p$ were known, then our data science algorithm will be optimized under $p$.

Since we don't know $p$, we use the unin.
prob. $g$ in place of $p$ (Plug-in strategy)
Entropy Estimation

\(x_1, x_2, \ldots\) with pmf \(p(x^n)\)

Shannon–McMillan–Breiman theorem

\[
\frac{1}{n} \log \frac{1}{p(x^n)} \rightarrow \overline{H}(x) \text{ as } \text{ entropy rate}
\]

for every stat. erg. process \(X\).

Now, since \(g_{\text{LZ}}\) is universal stat. erg. processes, we instead use

\[
\frac{1}{n} \log \frac{1}{g_{\text{LZ}}(x^n)} \rightarrow \overline{H}(x)
\]

Application: Apply this to estimate entropy of a text in multiple languages — which language is most efficient? Are original languages more efficient than translations?

Brief review of information theory

Entropy

\(X\) random variable on alphabet \(X\) drawn according to probability mass fcn \(p(x)\) on \(X\),
\[ X \sim p(X) = P(X=x) \]

The entropy of \( X \) is

\[ H(X) = \sum_{x \in X} p(x) \log \frac{1}{p(x)} = \mathbb{E} \left[ \log \frac{1}{p(x)} \right] \]

sometimes written as \( H(p(x)) \)

**Example**

\[ X \sim \text{Bern}(p) \] \quad p \in [0,1]

\[ X = \begin{cases} 1, & \text{wp. } p \\ 0, & \text{wp. } 1-p \end{cases} \]

Then \( H(X) = p \log \frac{1}{p} + (1-p) \log \frac{1}{1-p} = H(p) \)

\( \text{binary entropy function} \)

Here, we define \( 0 \cdot \log(0) = 0 \)

\( (\text{so } 0 \cdot \log \frac{1}{0} = 0) \),

as \( \lim_{x \to 0} x \log x = 0 \)

**Example**

\[ X \sim \text{Geometric}(p) \]

\[ H(X) = \sum_{i=1}^{\infty} p \cdot (1-p)^{i-1} \log \frac{1}{p(1-p)^{i-1}} \]

\[ = \log \frac{1}{p} + \frac{1-p}{p} \log \frac{1}{1-p} \]

**Properties of Entropy**
1. \( H(X) \geq 0 \), with equality if and only if \( X = c \) up to \( 1 \)

2. \( H(X) \) is a concave func in \( p(x) \), i.e.-
\[
H(\lambda p_i + (1-\lambda) p_0) \geq \lambda H(p_i) + (1-\lambda) H(p_0)
\]

3. \( H(X) = \log |X| \)
   - cardinality of \( X \)
   - How many values \( X \) can take

4. For any func \( f \),
\[
H(f(x)) \leq H(X)
\]
   With equality if and only if \( f \) is one-to-one on the support of \( X \)
   - set of \( x \in X \) with nonzero probability
   \[
   \{ x \in X : \rho(x) > 0 \}
   \]

- Jensen's ineq

Let \( f \) be a convex func,
Then
\[
E[f(X)] = f(E[X])
\]
Let $g$ be concave, then

$$E[g(x)] \leq g(E[x])$$

Now $H(X) = E\left[ \log \frac{1}{p(x)} \right] \leq \log E \left[ \frac{1}{p(x)} \right]$

$$= \sum_{x : p(x) > 0} p(x) \cdot \frac{1}{p(x)} = \sum_{x : p(x) > 0} 1$$

$$= |\{x : p(x) > 0\}| \leq \lvert X \rvert.$$

so $H(X) \leq \log \lvert X \rvert$