EE5585 Data Compression February 8, 2013

Lecture 7

Instructor: Arya Mazumdar *Scribe: Seth Hays*

Lemmas

Lemma 1

AEP: given the sequence: $x_1x_2x_3...x_n$; an independant identically distributed (iid) random variable, then:

$\frac{1}{n}\log p(x_1x_2x_3...x_n) \to \mathbf{H}(X)$ in probability (proved in Lecture 6)

Lemma 2

Given a binary sequence of length n parsed so that all phrases are distinct. The number of distinct phrases, $c(n)$, is:
$c(n) \leq \frac{n}{(1-\varepsilon_n)\log n}$
where $\varepsilon_n \to 0$ as $n \to \infty$

Lemma 3

X is a random variable that takes non-negative integer values and $\mathbf{E}(X) = \mu$ then $\mathbf{H}(X) \leq \mathbf{H}(Z)$, where Z is a geometric random variable and

$Pr(Z = k) = (1 - \frac{1}{\mu+1})^k \frac{1}{\mu+1}$; $k = 0, 1, 2, ...$

Verify that:

$\sum_{k=1}^{\infty} Pr(Z = k) = 1$ and $\mathbf{E}(z) = \mu$
$\mathbf{H}(Z) = (\mu + 1)\log\mu + 1 - \mu\log\mu$
(Proved later)

Prove L.Z. Code is Optimal

Theorem: L.Z. Code is optimal:
$\frac{c(n)[\log c(n)+1]}{n} \to \mathbf{H}(X)$

True for any file of length n that is produced from the source alphabet, $\mathcal{X}$ and random variable, X:
$X_1, X_2, ..., X_n$; X_i iid with X

Proof of the theorem:
Start with Lemma 1:

$\frac{1}{n}\log p(X_1X_2X_3...X_n)$

take $p(X_1X_2X_3...X_n)$ with parsed phrases of $S_1, S_2, ...S_{c(n)}$

therefore $\log[p(S_1, S_2, ...S_{c(n)})]$

$= \sum_{i=1}^{c(n)} \log p[S_i] = \sum_{l=1}^{l_{max}} [\sum_{l(S)=l} \log p(S)]$

$= \sum_{l=1}^{l_{max}} c_l \sum_{l(S)=l} \frac{1}{c_l} \log[p(S)]$
where c_l = the number of phrases of length l

$\leq \sum_{l=1}^{l_{max}} c_l \log \sum_{l(S)=l} \frac{1}{c_l}[p(S)] \rightarrow$ Jensen's Inequality

$\leq \sum_{l=1}^{l_{max}} c_l \log \frac{1}{c_l}$

$= c(n) \sum_{l=1}^{l_{max}} \frac{c_l}{c(n)} \log \frac{c(n)}{c_l} - \sum_{l=1}^{l_{max}} c_l \log[c(n)]$
take $\sum_{l=1}^{l_{max}} c_l \log[c(n)] = \log c(n) \sum_{l=1}^{l_{max}} c_l = c(n) \log c(n)$

$= c(n) \sum_{l=1}^{l_{max}} \frac{c_l}{c(n)} \log \frac{c(n)}{c_l} - c(n) \log c(n)$

$= c(n) \sum_{l=1}^{l_{max}} \pi_l \log \frac{1}{\pi_l} - c(n) \log c(n)$
where $\pi_l = \frac{c(n)}{c_l}$

$= c(n)\mathbf{H}(\Pi) - c(n) \log c(n)$

Using $\sum_{l=1}^{l_{max}} l\pi_l = \frac{\sum lc_l}{c(n)} = \frac{n}{c(n)} = \mu$:
$c(n)\mathbf{H}(\Pi) - c(n) \log c(n) \leq c(n)((\frac{n}{c(n)} + 1) \log(\frac{n}{c(n)} + 1) - \frac{n}{c(n)} \log \frac{n}{c(n)}) - c(n) \log c(n)$

Going back to the beginning:
$\frac{1}{n} \log p(x_1x_2x_3...x_n) \geq \frac{c(n) \log c(n)}{n}$

$\frac{-c(n)}{n}[(\frac{n}{c(n)} + 1) \log(\frac{n}{c(n)} + 1) - \frac{n}{c(n)} \log \frac{n}{c(n)}]$

$= (1 + \frac{c(n)}{n}) \log \frac{n}{c(n)} - \log \frac{n}{c(n)}$

$= \log(1 + \frac{c(n)}{n} + \frac{c(n)}{n} \log \frac{n}{c(n)} + 1$

$\lim_{n\rightarrow\infty} \log(1 + \frac{c(n)}{n} + \frac{c(n)}{n} \log \frac{n}{c(n)} + 1$
take $\frac{n}{c(n)} = m$

$= \log(1) + \lim_{m\rightarrow\infty} \frac{\log m+1}{m} = 0$

So as $n \rightarrow \infty$; $\frac{c(n)[\log c(n)+1}{n} \rightarrow \mathbf{H}(X)$

Prove Lemma 3

Take $Pr(X = k) = p$ and $Pr(Z = k) = q$

$\mathbf{H}(Z) - \mathbf{H}(X) = -\sum(q \log q) + \sum(p \log p)$

$= \sum(q[k \log 1 - \frac{1}{\mu+1}] - \log \mu + 1] + \sum(p \log p)$

$= \sum(kq[\log 1 - \frac{1}{\mu+1}]) + \sum(q \log \mu + 1) + \sum(p \log p)$

$= -\sum(kp[\log 1 - \frac{1}{\mu+1}]) + \sum(p \log \mu + 1) + \sum(p \log p)$

$= \sum(p[\log (1 - \frac{1}{\mu+1})^k [\frac{1}{\mu+1}]) + \sum(p \log p)$

$= -\sum(p \log q + p \log p)$

$= \sum(p \log \frac{p}{q}) = D(X||Z) \geq 0$

Now Prove:
$\mathbf{H}(Z) = (\mu + 1) \log (\mu + 1) - \mu \log \mu$

$\mathbf{H}(Z) = -\sum_k (1 - \frac{1}{\mu+1})^k \frac{1}{\mu+1} [k \log (1 - \frac{1}{\mu+1}) - \log \mu + 1]$

$= -\sum(kq \log (1 - \frac{1}{\mu+1})) + \sum(q \log (\mu + 1))$

$= -\mu \log \frac{\mu}{\mu+1} + \log (\mu + 1)$

$= (\mu + 1) \log (\mu + 1) - \mu \log \mu$