Mutual Information for a Deletion Channel

Michael Drmota, Wojciech Szpankowski, and Krishnamurthy Viswanathan
TU Wien, Purdue University, HPL

July 2, 2012

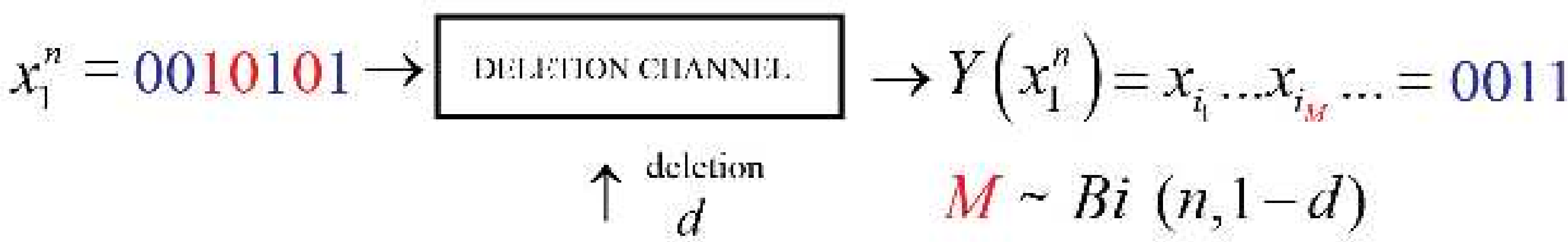
Center for Science of Information
NSF Science & Technology Center

ISIT, MIT, 2012

Outline

Deletion Channel

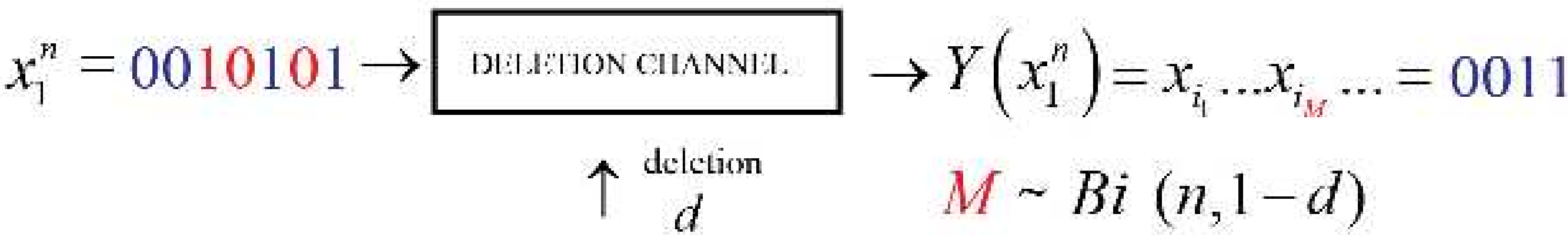
A deletion channel with parameter d:



input: a binary sequence $x := x_1^n = x_1 \cdots x_n$,
channel: deletes each symbol independently with probability d,
output: a *subsequence* $Y = Y(x) = x_{i_1} \dots x_{i_M}$ of x;
M follows the binomial distribution $\mathrm{Bi}(n, (1 - d))$ and indices $i_1, \dots, i_M$ correspond to undeleted bits.

Deletion Channel

A deletion channel with parameter d:



input: a binary sequence $x := x_1^n = x_1 \cdots x_n$,
channel: deletes each symbol independently with probability d,
output: a *subsequence* $Y = Y(x) = x_{i_1} \dots x_{i_M}$ of x;
M follows the binomial distribution $\mathrm{Bi}(n, (1 - d))$ and indices $i_1, \dots, i_M$ correspond to undeleted bits.

The channel capacity is

$$C(d) = \lim_{n \to \infty} \frac{1}{n} \sup_{P_{X_1^n}} I(X_1^n; Y(X_1^n)),$$

where $I(X_1^n; Y(X_1^n))$ is the mutual information between the input and output of the deletion channel.

Hidden Pattern Matching & Deletion Channel

Let $w = w_1 w_2 \ldots w_m \in \{0,1\}^m$, $m \le n$, be a given binary sequence and $\Omega_x(w)$ be the number of occurrences of w as a *subsequence* (not consecutive symbols) in $x := x_1^n$:

$$\Omega_x(w) = \sum_{1 \le i_1 < i_2 < \cdots < i_m \le n} \mathbf{I}_{[x_{i_1} = w_1]} \mathbf{I}_{[x_{i_2} = w_2]} \cdots \mathbf{I}_{[x_{i_m} = w_m]},$$

where $\mathbf{I}_A = 1$ if A is true and zero otherwise.

Example. Word **date** occurs four times in hidden pattern.

This problem is known as the hidden pattern matching problem studied in Flajolet, W.S., and Vallee, (2006) and Bourdon and Vallee, (2008).

Hidden Pattern Matching & Deletion Channel

Let $w = w_1 w_2 \dots w_m \in \{0,1\}^m$, $m \le n$, be a given binary sequence and $\Omega_x(w)$ be the number of occurrences of w as a *subsequence* (not consecutive symbols) in $x := x_1^n$:

$$\Omega_x(w) = \sum_{1 \le i_1 < i_2 < \cdots < i_m \le n} \mathbf{I}_{[x_{i_1} = w_1]} \mathbf{I}_{[x_{i_2} = w_2]} \cdots \mathbf{I}_{[x_{i_m} = w_m]},$$

where $\mathbf{I}_A = 1$ if A is true and zero otherwise.

Example. Word **date** occurs four times in hidden pattern.

This problem is known as the hidden pattern matching problem studied in Flajolet, W.S., and Vallee, (2006) and Bourdon and Vallee, (2008).

Our first main result shows a relation between the mutual information and $\Omega_x(w)$.

Theorem 1. *For any random input X_1^n, the mutual information satisfies*

$$I(X_1^n; Y(X_1^n)) = \sum_w d^{n-|w|}(1-d)^{|w|} \Big(\mathbf{E}[\Omega_{X_1^n}(w) \log \Omega_{X_1^n}(w)] - \mathbf{E}[\Omega_{X_1^n}(w)] \log \mathbf{E}[\Omega_{X_1^n}(w)] \Big).$$

Hidden Pattern Matching – Fixed m

For **fixed** m (fixed pattern) and memoryless sources the problem was studied by Flajolet, Vallee and W.S. (2001, 2006).

Proposition 1 (Flajolet, W.S., Vallee, 2006)**.** *The mean and the variance are:*

$$E[\Omega_n(w)] \quad = \quad \binom{n}{m} P(w) \sim \frac{P(w)}{m!} n^m \left(1 + O(\frac{1}{n})\right),$$

$$\mathbf{Var}_n[\Omega_n(w)] \quad = \quad \frac{P^2(w)}{(2m-1)!} \kappa^2(w) \cdot n^{2m-1} \left(1 + O(\frac{1}{n})\right),$$

where

$$\kappa^2(w) := \sum_{1 \le r,s \le m} \binom{r+s-2}{r-1} \binom{2m-r-s}{m-r} \mathbf{I}(w_r = w_s) \left(\frac{1}{P(w_r)} - 1 \right).$$

Furthermore, for any fixed x

$$\frac{\Omega_n(w) - \mathbf{E}[\Omega_n(w)]}{\sqrt{\mathbf{Var}[\Omega_n(w)]}} \to N(0,1)$$

where $N(0,1)$ is the standard normal distribution.

Bourdon and Vallee (2008) extended it to dynamic sources (e.g., Markov).

Hidden Pattern Matching – Large m

Let now assume that $m = \theta n$, $\theta = 1 - d$, and $p = P(1)$.
For memoryless sources (with parameter p) we know that

$$\mathbf{E}[\Omega_n(w)] = \binom{n}{m} P(w)$$

and for a typical w we have: $\mathbf{E}[\Omega_n(w)] \sim 2^{n(H(\theta)-\theta H(p))}$,
where $H(x)$ is the binary entropy. Notice that $\mathbf{E}[\Omega_n(w)]$ may exponentially increase or decrease depending whether $H(\theta) - \theta H(p) > 0$ or not.

Hidden Pattern Matching – Large m

Let now assume that $m = \theta n$, $\theta = 1 - d$, and $p = P(1)$.
For memoryless sources (with parameter p) we know that

$$\mathbf{E}[\Omega_n(w)] = \binom{n}{m} P(w)$$

and for a typical w we have: $\mathbf{E}[\Omega_n(w)] \sim 2^{n(H(\theta) - \theta H(p))}$,
where $H(x)$ is the binary entropy. Notice that $\mathbf{E}[\Omega_n(w)]$ may exponentially increase or decrease depending whether $H(\theta) - \theta H(p) > 0$ or not.

Variance. In general,

$$\mathbf{Var}[\Omega_n(w)] = \sum_{k=1}^{m} \binom{n}{2m-k} P^2(w) \kappa^2(w)$$

where

$$\kappa^2(w) = \sum_{\substack{1 \le r_1 < \cdots r_k \le m \\ 1 \le s_1 < \cdots s_k \le m}} \binom{r_1 + s_1 - 2}{r_1 - 1} \binom{r_2 - r_1 + s_2 - s_1 - 2}{r_2 - r_1 - 1} \cdots \binom{2m - r_k - s_k}{m - r_k} \cdot$$

$$\cdot \quad \prod_{i=1}^{k} \mathbf{I}(w_{r_i} = w_{s_i}) \left(\frac{1}{P(w_{r_1} \cdots w_{r_k})} - 1 \right)$$

Some Additional Observations

1. For special patterns we can estimate asymptotically the variance. For example for $w = 0^m$ we we have

$$P\left(\Omega_n(0^m) = \binom{n-k}{m}\right) = \binom{n}{k} p^k (1-p)^{n-k}$$

and then

$$\mathbf{Var}[\Omega_n(0^m)] \sim \mathbf{E}[\Omega_n^2(0^m)] \sim 2^{n\beta((1-d)p)}$$

where with $\theta = (1-d)$ and

$$\beta(\theta, p) = (2(q+\theta p-\delta)H(\theta/(q+\theta p-\delta))+H((1-\theta)p+\delta)+((1-\theta)p+\delta)\log p+(q+\theta p-\delta)$$

Some Additional Observations

1. For special patterns we can estimate asymptotically the variance. For example for $w = 0^m$ we we have

$$P\left(\Omega_n(0^m) = \binom{n-k}{m}\right) = \binom{n}{k} p^k (1-p)^{n-k}$$

and then

$$\mathbf{Var}[\Omega_n(0^m)] \sim \mathbf{E}[\Omega_n^2(0^m)] \sim 2^{n\beta((1-d)p)}$$

where with $\theta = (1-d)$ and

$$\beta(\theta, p) = (2(q+\theta p-\delta)H(\theta/(q+\theta p-\delta))+H((1-\theta)p+\delta)+((1-\theta)p+\delta)\log p+(q+\theta p-\delta)$$

2. When $\mathbf{E}[\Omega_n(w)] \to \infty$ we have

$$\mathbf{E}[\Omega_n(w)\log \Omega_n(w)] - \mathbf{E}[\Omega_n(w)]\log \mathbf{E}[\Omega_n(w)] \sim \frac{\mathbf{Var}[\Omega_n(w)]}{2\mathbf{E}[\Omega_n(w)]}$$

which can be used to estimate the mutual information, if one knows how to compute the variance.

Proof of Theorem 1

We use $I(X;Y) = H(Y) - H(Y|X)$. Notice that

$$\begin{aligned} P(Y(X_1^n) = w | X_1^n = x_1^n) &= \Omega_n(w) d^{n-|w|}(1-d)^{|w|} \\ P(Y = w) &= \sum_{x \in \mathcal{A}^n} P(X = x)\Omega_n(w) d^{n-|w|}(1-d)^{|w|}. \end{aligned}$$

Hence

$$H(Y) = -\sum_w d^{n-|w|}(1-d)^{|w|} \left(\mathbf{E}[\Omega_X(w)] \log \mathbf{E}[\Omega_X(w)] + \mathbf{E}[\Omega_X(w)] \log(d^{n-|w|}(1-d)^{|w|})\right)$$

Since $P(x, y) = P(x)\Omega_x(y) d^{n-m}(1-d)^m$ also have

$$H(Y|X) = -\sum_w d^{n-|w|}(1-d)^{|w|} \left(\mathbf{E}[\Omega_X(w) \log \Omega_X(w)] + \mathbf{E}[\Omega_X(w)] \log d^{n-|w|}(1-d)^{|w|}\right)$$

and the proof is complete.

Proof of Theorem 1

We use $I(X;Y) = H(Y) - H(Y|X)$. Notice that

$$\begin{aligned} P(Y(X_1^n) = w | X_1^n = x_1^n) &= \Omega_n(w) d^{n-|w|}(1-d)^{|w|} \\ P(Y = w) &= \sum_{x \in \mathcal{A}^n} P(X = x) \Omega_n(w) d^{n-|w|}(1-d)^{|w|}. \end{aligned}$$

Hence

$$H(Y) = -\sum_w d^{n-|w|}(1-d)^{|w|} \left(\mathbf{E}[\Omega_X(w)] \log \mathbf{E}[\Omega_X(w)] + \mathbf{E}[\Omega_X(w)] \log(d^{n-|w|}(1-d)^{|w|}) \right)$$

Since $P(x, y) = P(x)\Omega_x(y) d^{n-m}(1-d)^m$ also have

$$H(Y|X) = -\sum_w d^{n-|w|}(1-d)^{|w|} \left(\mathbf{E}[\Omega_X(w) \log \Omega_X(w)] + \mathbf{E}[\Omega_X(w)] \log d^{n-|w|}(1-d)^{|w|} \right)$$

and the proof is complete.

Note. It is easy to see that

$$I(X_1^n; Y(X_1^n)) \leq n(1-d)$$

and hence $C(d) \leq 1 - d$.

Memoryless Sources

From now on we assume that the source is **memoryless** over $\mathcal{A} = \{0, 1\}$ with $p = P(1)$ and $q = 1 - p$. Extension to Markov sources possible.

Define

$$S_1 = \sum_w d^{n-|w|}(1-d)^{|w|}\mathbf{E}[\Omega_{X_1^n}(w)\log\Omega_{X_1^n}(w)],$$

$$S_2 = \sum_w d^{n-|w|}(1-d)^{|w|}\mathbf{E}[\Omega_{X_1^n}(w)]\log\mathbf{E}[\Omega_{X_1^n}(w)],$$

so that $I((X_1^n; Y(X_1^n)) = S_1(X_1^n, Y(X_1^n)) - S_2(X_1^n, Y(X_1^n)) := S_1 - S_2$.

Furthermore, let

$$\begin{aligned} I(d,p) &= \lim_{n\to\infty}\frac{1}{n}I(X_1^n; Y(X_1^n)) \\ \lambda(d,p) &= \lim_{n\to\infty}\frac{1}{n}S_1(X_1^n, Y(X_1^n)). \end{aligned}$$

Second Main Result

Theorem 2. *The limit* $I(d,p)$ *and* $\lambda(d,p)$ *exist and*

$$\lim_{n\to\infty} \frac{1}{n} S_2(X_1^n, Y(X_1^n)) = H(1-d) - (1-d)H(p)$$

as well as

$$I(d,p) = \lambda(d,p) + (1-d)H(p) - H(1-d).$$

Furthermore,

$$\begin{aligned} I(d,p) &= \inf_{n\geq 1} \frac{1}{n} I(X_1^n; Y(X_1^n)) \\ \lambda(d,p) &= \sup_{n\geq 1} \frac{1}{n} S_1(X_1^n, Y(X_1^n)) \end{aligned}$$

where $H(p) = -p\log p - (1-p)\log(1-p)$ *is the binary entropy function.*

Second Main Result

Theorem 2. *The limit* $I(d,p)$ *and* $\lambda(d,p)$ *exist and*

$$\lim_{n\to\infty} \frac{1}{n} S_2(X_1^n, Y(X_1^n)) = H(1-d) - (1-d)H(p)$$

as well as

$$I(d,p) = \lambda(d,p) + (1-d)H(p) - H(1-d).$$

Furthermore,

$$\begin{aligned} I(d,p) &= \inf_{n\geq 1} \frac{1}{n} I(X_1^n; Y(X_1^n)) \\ \lambda(d,p) &= \sup_{n\geq 1} \frac{1}{n} S_1(X_1^n, Y(X_1^n)) \end{aligned}$$

where $H(p) = -p\log p - (1-p)\log(1-p)$ *is the binary entropy function.*

Thus, $I(d,p) \leq \frac{1}{n} I(X_1^n; Y(X_1^n))$ for all $n \geq 1$. In particular,

$$I(d,p) \leq (1-d)H(p),$$

$$I(d,p) \leq d(1-d)(H(p) + p^2 + q^2 - 1) + (1-d)^2 H(p),$$

Special Cases: $d \to 1$ and $d \to 0$

Theorem 3. *As* $d \to 1$

$$I(d,p) \le K(1-d)^{4/3} \log \frac{1}{1-d}$$

where the constant $K > 0$ *is absolute.*

Note. The capacity $C(d) = \Theta(1-d)$ but it cannot be achieved by a memoryless source.

Special Cases: $d \to 1$ and $d \to 0$

Theorem 3. *As* $d \to 1$

$$I(d, p) \leq K(1 - d)^{4/3} \log \frac{1}{1 - d}$$

where the constant $K > 0$ *is absolute.*

Note. The capacity $C(d) = \Theta(1 - d)$ but it cannot be achieved by a memoryless source.

Theorem 4. *As* $d \to 0$,

$$I(d, p) \geq (1 - d)H(p) + d \log d - d \log(e) + d(q^2 f(p) + p^2 f(p)) + O\left(d^{2-\varepsilon}\right)$$

where $f(x) = \sum_{\ell \geq 2} x^\ell \, \ell \log \ell$. *Furthermore, as* $d \to 0$,

$$I(d, p) \leq H(p) + d \log d + O(d \log \log(1/d)).$$

Special Cases: $d \to 1$ and $d \to 0$

Theorem 3. *As* $d \to 1$

$$I(d,p) \leq K(1-d)^{4/3} \log \frac{1}{1-d}$$

where the constant $K > 0$ *is absolute.*

Note. The capacity $C(d) = \Theta(1-d)$ but it cannot be achieved by a memoryless source.

Theorem 4. *As* $d \to 0$,

$$I(d,p) \geq (1-d)H(p) + d \log d - d \log(e) + d(q^2 f(p) + p^2 f(p)) + O\left(d^{2-\varepsilon}\right)$$

where $f(x) = \sum_{\ell \geq 2} x^\ell \, \ell \log \ell$. *Furthermore, as* $d \to 0$,

$$I(d,p) \leq H(p) + d \log d + O(d \log \log(1/d)).$$

We can recover (a weaker) Kanoria & Montanari result

$$C(d) = I(d, 0.5) + O(d^{3/2-\varepsilon}) = 1 + d \log d - Ad + O(d^{2-\varepsilon})$$

where $A = \log(2e) - \sum_{\ell \geq 1} 2^{-\ell-1} \ell \log \ell$.
Note that symmetric memoryless distribution is asymptotically optimal.

Proof of Theorem 2

1. We first observe that

$$\Omega_{x_1^{n+k}}(w) = \sum_{w_1 w_2 = w} \Omega_{x_1^n}(w_1)\Omega_{x_{n+1}^{n+k}}(w_2),$$

for any $x^{n+k} \in \mathcal{A}^{n+k}$.

2. Then we establish subadditivity property of the mutual information

$$I(X_1^{n+k}; Y(X_1^{n+k})) \leq I(X_1^n; Y(X_1^n)) + I(X_1^k; Y(X_1^k)).$$

This follows from 1 above and

$$\sum_{m=1}^{M} z_m \log \frac{\sum_{m=1}^{M} z_m}{\sum_{m=1}^{M} a_m} \leq \sum_{m=1}^{M} z_m \log \frac{z_m}{a_m}.$$

By Fekete's lemma we then have

$$I(d,p) = \inf_{n \geq 1} \frac{1}{n} I(X_1^n; Y(X_1^n)).$$

Proof of Theorem 2

3. It is easy to see that $(\mathbf{E}[\Omega_n(w)] = \binom{n}{m} P(w))$

$$S_2 = \sum_w d^{n-|w|}(1-d)^{|w|}\mathbf{E}[\Omega_{X_1^n}(w)] \log \mathbf{E}[\Omega_{X_1^n}(w)] \sim n\cdot(H(1-d) - (1-d)H(p)) \, .$$

4. Let $a_n = S_1(X_1^n, Y(X_1^n))$. Then we can prove superadditivity property of a_n, that is

$$a_{n+k} \geq a_n + a_k$$

which immediately implies that

$$0 \leq \lambda(d, p) := \sup_{n \geq 1} \frac{a_n}{n} \leq H(1-d).$$

This completes the proof of Theorem 2.

Proof of Theorem 3

1. Recall that

$$I(X;Y) = \sum_{w\in\{0,1\}^n} d^{n-|w|}(1-d)^{|w|}\left(\mathbf{E}[\Omega_{X_1^n}(w)\log\Omega_{X_1^n}(w)] - \mathbf{E}[\Omega_{X_1^n}(w)]\log\mathbf{E}[\Omega_{X_1^n}(w)]\right).$$

We compute now the sum for all words w of length $|w| = 1$ and $|w| \geq 2$. For example, if $w = 0$ and if $X = X_1^n = 0^m 1^{n-m}$, then $\Omega_X(w) = m$

2. Let T_1 be the above sum for $|w| = 1$ (i.e, $w = 0$ or $w = 1$). We have

$$\begin{aligned} T_1 :=& d^{n-1}(1-d)\left(\sum_{m=1}^{n} m\log m\binom{n}{m}\left(p^m q^{n-m} + p^{n-m}q^m\right)\right) \\ &- d^{n-1}(1-d)\left(np\log(np) + nq\log(nq)\right) \end{aligned}$$

But $\log m = \log(np) + \log\left(1 + \frac{m-np}{np}\right) \leq \log(np) + \frac{m-np}{np}$, thus

$$\sum_{m=1}^{n} m\log m\binom{n}{m}p^m q^{n-m} \leq \log(np)np + \frac{npq}{np} = np\log(np) + q.$$

Summing up $T_1 \leq d^{n-1}(1-d) \leq (1-d)$.

Proof of Theorem 3

3. Now consider $|w| \geq 2$ and use $\Omega_X(w) \leq \binom{n}{|w|}$. Then it contributes to T_2

$$T_2 \leq 2\sum_{\ell=2}^{n} d^{n-\ell}(1-d)^{\ell}\binom{n}{\ell}\log\binom{n}{\ell} \leq 2d^n \log n \frac{n(1-d)}{d}\left(e^{n(1-d)/d} - 1\right)$$

thus $T_2 \leq C_1 n^2 (1-d)^2 \log n$.

4. Summing up

$$I(d,p) \leq \frac{1}{n} I(X_1^n; Y(X_1^n)) \leq \frac{1-d}{n} + C_1 n (1-d)^2 \log n.$$

Set now $n = \lfloor (1-d)^{-1/3} \rfloor$ we arrive at

$$I(d,p) \leq K\,(1-d)^{4/3} \log \frac{1}{1-d}$$

which completes the proof.

Sketch of Proof for Theorem 4

1. Recall that $a_n = S_1(X, Y)$ and we lower bound it by considering only words $|w| = n - 1$ so that

$$S_1 \geq d(1-d)^{n-1} \sum_{|w|=n-1} \mathbf{E}[\Omega_X(w) \log \Omega_X(w)].$$

Sketch of Proof for Theorem 4

1. Recall that $a_n = S_1(X,Y)$ and we lower bound it by considering only words $|w| = n-1$ so that

$$S_1 \geq d(1-d)^{n-1} \sum_{|w|=n-1} \mathbf{E}[\Omega_X(w) \log \Omega_X(w)].$$

2. We consider $w = 0^{i_1}1^{j_1}0^{i_2}1^{j_2}\cdots 0^{i_K}1^{j_K}$ of length $n-1$. Then $\Omega_n(w) = \ell$ iff exists r such that

$$i_r = \ell - 1 \quad \text{and} \quad X = 0^{i_1}1^{j_1}\cdots 1^{j_{r-1}}0^{i_{r+1}}1^{j_r}\cdots 0^{i_K}1^{j_K} \quad \text{(missing } 0^{i_r}1^{j_r}\text{)}$$

or

$$j_r = \ell - 1 \quad \text{and} \quad X = 0^{i_1}1^{j_1}\cdots 0^{i_r}1^{j_{r+1}}0^{i_{r+1}}\cdots 0^{i_K}1^{j_K} \quad \text{(missing } 1^{j_r}0^{i_{r+1}}\text{)}$$

This leads to

$$\sum_{|w|=n-1} \mathbf{E}[\Omega_X(w) \log \Omega_X(w)] = \sum_{\ell \geq 2} \ell \log \ell \sum_{|w|=n-1} P(w) \sum_{r \geq 1} (p\mathbf{I}_{[i_r(w)=\ell-1]} + q\mathbf{I}_{[j_r(w)=\ell-1]}).$$

3. Since $\sum_{r\geq 1}(p\mathbf{I}_{[i_r(w)=\ell-1]} + q\mathbf{I}_{[j_r(w)=\ell-1]}) \sim npq\left(pp^{\ell-2}q + qq^{\ell-2}p\right)$ we conclude that

$$\sum_{|w|=n-1} \mathbf{E}[\Omega_X(w) \log \Omega_X(w)] \sim n \sum_{\ell \geq 2} \ell \log \ell \left(p^{\ell}q^{2} + q^{\ell}p^{2}\right) = n\left(q^{2}f(p) + p^{2}f(q)\right).$$

That's IT