Problem 1: Moment Generating Function

In the class we had considered the logarithmic moment generating function

\[ \phi(s) := \ln E[\exp(sX)] = \ln \sum_x p(x) \exp(sx) \]

of a real-valued random variable \( X \) taking values on a finite set, and showed that \( \phi'(s) = E[X_s] \) where \( X_s \) is a random variable taking the same values as \( X \) but with probabilities \( p_s(x) := p(x) \exp(sx) \exp(-\phi(s)) \).

(a) Show that

\[ \phi''(s) = \text{Var}(X_s) := E[X_s^2] - E[X_s]^2 \]

and conclude that \( \phi''(s) \geq 0 \) and the inequality is strict except when \( X_s \) is deterministic.

(b) Let \( x_{\text{min}} := \min\{x : p(x) > 0\} \) and \( x_{\text{max}} := \max\{x : p(x) > 0\} \) be the smallest and largest values \( X \) takes. Show that

\[ \lim_{s \to -\infty} \phi'(s) = x_{\text{min}}, \quad \text{and} \quad \lim_{s \to \infty} \phi'(s) = x_{\text{max}}. \]

Solution

(a) As \( \phi(s) := \ln E[\exp(sX)] \), we have

\[ \phi'(s) = \frac{E[X \exp(sX)]}{E[\exp(sX)]} = E[X \exp(sX) \exp(-\phi(s))] = E[X_s] \]

(1)

\[ \phi''(s) = \frac{E[X^2 \exp(sX)]}{E[\exp(sX)]} - \frac{E[X \exp(sX)]E[X \exp(sX)]}{E[\exp(sX)]^2} \]

(2)

The second term is \( E[X_s]^2 \) and the first term equals \( \sum_x x^2 \exp(sx)/\exp(\phi(s)) = E[X_s^2] \). So \( \phi''(s) = \text{Var}(X_s) \). Moreover, \( \text{Var}(X_s) \geq 0 \) with equality only when \( X_s \) is deterministic. But \( X_s \) is deterministic only when \( X \) is.

(b) Observe that

\[ \phi'(s) = \frac{E[X \exp(sX)]}{E[\exp(sX)]} = \frac{E[X \exp(sX)] \exp(-sx_{\text{max}})}{E[\exp(sX)] \exp(-sx_{\text{max}})} \]

(3)

\[ = \frac{\sum_x p(x)x \exp(-s(x_{\text{max}} - x))}{\sum_x p(x) \exp(-s(x_{\text{max}} - x))} \]

(4)
In the sums above, as $s \to \infty$, all terms vanish except the ones for $x = x_{\text{max}}$. Hence we have

$$
\lim_{s \to \infty} \phi'(s) = \frac{p(x_{\text{max}})x_{\text{max}}}{p(x_{\text{max}})} = x_{\text{max}} \tag{5}
$$

Similarly, we can show that $\lim_{s \to -\infty} \phi'(s) = x_{\text{min}}$.

**Problem 2: Divergence and $L_1$**

Suppose $p$ and $q$ are two probability mass functions on a finite set $U$. (I.e., for all $u \in U$, $p(u) \geq 0$ and $\sum_{u \in U} p(u) = 1$; similarly for $q$.)

(a) Show that the $L_1$ distance $\|p - q\|_1 := \sum_{u \in U} |p(u) - q(u)|$ between $p$ and $q$ satisfies

$$
\|p - q\|_1 = 2 \max_{S,S \subseteq U} p(S) - q(S)
$$

with $p(S) = \sum_{u \in S} p(u)$ (and similarly for $q$), and the maximum is taken over all subsets $S$ of $U$.

For $\alpha$ and $\beta$ in $[0,1]$, define the function $d_2(\alpha \| \beta) := \alpha \log \frac{\alpha}{\beta} + (1 - \alpha) \log \frac{1 - \alpha}{1 - \beta}$. Note that $d_2(\alpha \| \beta)$ is the divergence of the distribution $(\alpha, 1 - \alpha)$ from the distribution $(\beta, 1 - \beta)$.

(b) Show that the first and second derivatives of $d_2$ with respect to its first argument $\alpha$ satisfy

$$
d_2'(\beta \| \beta) = 0 \quad \text{and} \quad d_2''(\alpha \| \beta) = \frac{\log e}{\alpha(1 - \alpha)} \geq 4 \log e.
$$

(c) By Taylor’s theorem conclude that

$$
d_2(\alpha \| \beta) \geq 2(\log e)(\alpha - \beta)^2.
$$

(d) Show that for any $S \subseteq U$

$$
D(p\|q) \geq d_2(p(S)\|q(S))
$$

[Hint: use the data processing theorem for divergence.]

(e) Combine (a), (c) and (d) to conclude that

$$
D(p\|q) \geq \frac{\log e}{4} \|p - q\|_1^2.
$$

(f) Show, by example, that $D(p\|q)$ can be $+\infty$ even when $\|p - q\|_1$ is arbitrarily small. [Hint: considering $U = \{0,1\}$ is sufficient.] Consequently, there is no generally valid inequality that upper bounds $D(p\|q)$ in terms of $\|p - q\|_1$.

**Solution**

(a) For any set $S$, we have

$$
p(S) - q(S) = \sum_{u \in S} p(u) - q(u) \leq \sum_{u \in S} |p(u) - q(u)|. \tag{6}
$$

Similarly for the compliment set of $S$, we also have

$$
q(S^c) - p(S^c) = \sum_{u \in S^c} q(u) - p(u) \leq \sum_{u \in S^c} |p(u) - q(u)|. \tag{7}
$$
Note that \( p(S) + p(S^c) = q(S) + q(S^c) = 1 \). Thus \( q(S^c) - p(S^c) = p(S) - q(S) \). Therefore, we have

\[
2(p(S) - q(S)) \leq \sum_{u \in S} |p(u) - q(u)| + \sum_{u \in S^c} |p(u) - q(u)| = \sum_{u \in \mathcal{U}} |p(u) - q(u)| = \|p - q\|_1
\]  

(8)

For the choice \( S = \{ u : p(u) > q(u) \} \), we have

\[
p(S) - q(S) = \sum_{u \in S} p(u) - q(u) = \sum_{u \in S} |p(u) - q(u)|
\]

(9)

\[
q(S^c) - p(S^c) = \sum_{u \in S^c} q(u) - p(u) = \sum_{u \in S^c} |p(u) - q(u)|
\]

(10)

So, for this \( S \), we have \( 2(p(S) - q(S)) = \|p - q\|_1 \).

(b): Since \( d_2(\alpha\|\beta) = \alpha \log \frac{\alpha}{\beta} + (1 - \alpha) \log \frac{1 - \alpha}{1 - \beta} \),

\[
d'_2(\alpha\|\beta) = \frac{\partial d_2(\alpha\|\beta)}{\partial \alpha} = \log \frac{\alpha}{\beta} + \log e - \log \frac{1 - \alpha}{1 - \beta} - \log e = \log \frac{\alpha(1 - \beta)}{\beta(1 - \alpha)}
\]

(11)

Therefore, we have \( d'_2(\beta\|\beta) = 0 \).

\[
d''_2(\alpha\|\beta) = \frac{\log e}{\alpha(1 - \alpha)} \geq 4 \log e
\]

(12)

where equality achieves when \( \alpha = 1/2 \).

(c): Taylor theorem says that for any \( f \) for which \( f'' \) is continuous

\[
f(\alpha) = f(\beta) + (\alpha - \beta)f'(\beta) + (1/2)(\alpha - \beta)^2 f''(x_i)
\]

(13)

where \( x_i \) is a value between \( \alpha \) and \( \beta \). With \( f(\alpha) = d_2(\alpha\|\beta) \), we thus have

\[
d_2(\alpha\|\beta) = 0 + 0 + (1/2)(\alpha - \beta)^2 f''(x_i) \geq 2 \log(e)(\alpha - \beta)^2
\]

(14)

(d) Consider a deterministic channel with binary output

\[
V = \begin{cases} 
1, & \text{if } V \in S \\
0, & \text{if } V \notin S 
\end{cases}
\]

(15)

Thus,

\[
d_2(p(S)||q(S)) = p(S) \log \frac{p(S)}{q(S)} + (1 - p(S)) \log \frac{1 - p(S)}{1 - q(S)}
\]

(16)

\[
= p(V = 1) \log \frac{p(V = 1)}{q(V = 1)} + p(V = 0) \log \frac{p(V = 0)}{q(V = 0)}
\]

(17)

\[= D(pV||qV)\]

(18)

By data processing theorem for divergence, \( D(p\|q) \geq D(pV\|qV) \)

(e) Combine (a),(c) and (d) and choosing \( S = \{ u : p(u) > q(u) \} \), we have \( \forall S \)

\[
D(p\|q) \geq d_2(p(S)||q(S)) \geq 2(\log e)(p(S) - q(S))^2 = \frac{\log e}{2} \|p - q\|_1^2
\]

(19)

(f) Let \( p \) be Bernoulli distribution with probability \( \epsilon \) to be 1 and \( q \) is 0 with probability 1. Then

\[
D(p\|q) = p(1) \log \frac{p(1)}{q(1)} + p(0) \log \frac{p(0)}{q(0)} = +\infty
\]

(20)

But \( \|p - q\|_1 = 2\epsilon \).
Problem 3: Other Divergences

Suppose $f$ is a convex function defined on $(0, \infty)$ with $f(1) = 0$. Define the $f$-divergence of a distribution $p$ from a distribution $q$ as

$$D_f(p||q) := \sum_u q(u)f(p(u)/q(u)).$$

In the sum above we take $f(0) := \lim_{t \to 0} f(t)$, $0f(0/0) := 0$, and $0f(a/0) := \lim_{t \to 0} tf(a/t) = a\lim_{t \to 0} tf(1/t)$.

(a) Show that for any non-negative $a_1, a_2, b_1, b_2$ and with $A = a_1 + a_2$, $B = b_1 + b_2$,

$$b_1f(a_1/b_1) + b_2f(a_2/b_2) \geq Bf(A/B);$$

and that in general, for any non-negative $a_1, \ldots, a_k$, $b_1, \ldots, b_k$, and $A = \sum_i a_i$, $B = \sum_i b_i$, we have

$$\sum_i b_if(a_i/b_i) \geq Bf(A/B).$$

[Hint: since $f$ is convex, for any $\lambda \in [0, 1]$ and any $x_1, x_2 > 0$ $\lambda f(x_1) + (1 - \lambda)f(x_2) \geq f(\lambda x_1 + (1 - \lambda)x_2)$; consider $\lambda = b_1/B$.]

(b) Show that $D_f(p||q) \geq 0$.

(c) Show that $D_f$ satisfies the data processing inequality: for any transition probability kernel $W(v|u)$ from $\mathcal{U}$ to $\mathcal{V}$, and any two distributions $p$ and $q$ on $\mathcal{U}$

$$D_f(p||q) \geq D_f(\tilde{p}||\tilde{q})$$

where $\tilde{p}$ and $\tilde{q}$ are probability distributions on $\mathcal{V}$ defined via $\tilde{p}(v) := \sum_u W(v|u)p(u)$, and $\tilde{q}(v) := \sum_u W(v|u)q(u)$.

(d) Show that each of the following are $f$-divergences.

i. $D(p||q) := \sum_u p(u)\log(p(u)/q(u))$. [Warning: log is not the right choice for $f$.]

ii. $R(p||q) := D(q||p)$.

iii. $1 - \sum_u \sqrt{p(u)q(u)}$

iv. $\|p - q\|_1$

v. $\sum_u (p(u) - q(u))^2/q(u)$

Solution

(a) Since $f$ is convex, for any $\lambda \in [0, 1]$ and any $x_1, x_2 > 0$ we have

$$\lambda f(x_1) + (1 - \lambda)f(x_2) \geq f(\lambda x_1 + (1 - \lambda)x_2) \quad (21)$$

By substitution $x_1 = a_1/b_1$, $x_2 = a_2/b_2$ and $\lambda = b_1/(b_1 + b_2)$:

$$\frac{b_1}{b_1 + b_2}f\left(\frac{a_1}{b_1}\right) + \left(1 - \frac{b_1}{b_1 + b_2}\right)f\left(\frac{a_2}{b_2}\right) \geq f\left(\frac{b_1}{b_1 + b_2} \frac{a_1}{b_1} + \left(1 - \frac{b_1}{b_1 + b_2}\right) \frac{a_2}{b_2}\right) \quad (22)$$

$$\Leftrightarrow b_1f\left(\frac{a_1}{b_1}\right) + b_2f\left(\frac{a_2}{b_2}\right) \geq Bf(A/B) \quad (23)$$
Let $A_k = \sum_{i=1}^{k} a_i$, $B_k = \sum_{i=1}^{k} b_i$. As we have proved that the following inequality holds for $k = 1, 2$:

$$\sum_{i=1}^{k} b_i f(a_i/b_i) \geq B_k f(A_k/B_k).$$

(24)

We assume that it also holds for $k = n$. For $k = n + 1$, we have

$$\sum_{i=1}^{n+1} b_i f(a_i/b_i) = \sum_{i=1}^{n} b_i f(a_i/b_i) + b_{n+1} f(a_{n+1}/b_{n+1})$$

(25)

$$\geq B_n f(A_n/B_n) + b_{n+1} f(a_{n+1}/b_{n+1})$$

(26)

$$\geq B_{n+1} f(A_{n+1}/B_{n+1})$$

(27)

By induction, for all any non-negative $k$, we have

$$\sum_{i=1}^{k} b_i f(a_i/b_i) \geq B_k f(A_k/B_k).$$

(28)

(b) $D_f(p||q) = \sum_u q(u) f(p(u)/q(u)) \geq (\sum_u q(u)) f\left(\frac{\sum_u p(u)}{\sum_u q(u)}\right) = f(1) = 0.$

(c)

$$D_f(p||q) = \sum_u q(u) f(p(u)/q(u)) = \sum_u \sum_v W(v|u) q(u) f(p(u)/q(u))$$

(29)

$$= \sum_u \sum_v W(v|u) q(u) f(W(v|u)p(u)/(W(v|u)q(u)))$$

(30)

$$\geq \sum_v (\sum_u W(v|u) q(u)) f\left(\frac{\sum_u W(v|u)p(u)}{\sum_u W(v|u)q(u)}\right)$$

(31)

$$= \sum_v \tilde{q}(v) f(\tilde{p}(v)/\tilde{q}(v))$$

(32)

$$= D_f(\tilde{p}||\tilde{q})$$

(33)

(d)

i. $D(p||q) := \sum_u p(u) \log(p(u)/q(u)) = \sum_u q(u) \frac{p(u)}{q(u)} \log \frac{p(u)}{q(u)}$. So $f(t) = t \log t$.

ii. $R(p||q) := D(q||p) = \sum_u p(u) \log(p(u)/q(u)) = \sum_u p(u)(- \log(q(u)/p(u)))$. So $f(t) = - \log t$.

iii. $1 - \sum_u \sqrt{p(u)q(u)} = \sum_u q(u) \left(1 - \sqrt{p(u)/q(u)}\right)$. So $f(t) = 1 - \sqrt{t}$.

iv. $\|p - q\|_1 = \sum_u |p(u) - q(u)| = \sum_u q(u)|p(u)/q(u) - 1|$. So $f(t) = |t - 1|$.

v. $\sum_u (p(u) - q(u))^2/q(u) = \sum_u q(u)(p(u)/q(u) - 1)^2$. So $f(t) = (t - 1)^2$.

**Problem 4: Entropy and pairwise independence**

Suppose $X$, $Y$, $Z$ are pairwise independent fair flips, i.e., $I(X; Y) = I(Y; Z) = I(Z; X) = 0$.

(a) What is $H(X, Y)$?

(b) Give a lower bound to the value of $H(X, Y, Z)$.
(c) Give an example that achieves this bound.

Solution

(a) Since $X$, $Y$, $Z$ are pairwise independent fair flips, $H(X) = H(Y) = H(Z) = 1$. $H(X, Y) = H(X) + H(Y|X) = H(X) + H(Y) - I(X; Y) = 2$.

(b) $H(X, Y, Z) = H(X, Y) + H(Z|X, Y) \geq H(X, Y) = 2$

(c) Let $Z = X + Y \mod 2$, then $H(Z|X, Y) = 0$ and $H(X, Y, Z) = H(X, Y)$.

Problem 5: Generating fair coin flips from biased coins

Suppose $X_1, X_2, \ldots$ are the outcomes of independent flips of a biased coin. Let $\Pr(X_i = 1) = p$, $\Pr(X_i = 0) = 1 - p$, with $p$ unknown. By processing this sequence we would like to obtain a sequence $Z_1, Z_2, \ldots$ of fair coin flips.

Consider the following method: We process the $X$ sequence in successive pairs, $(X_1, X_2)$, $(X_3, X_4)$, $(X_5, X_6)$, mapping (01) to 0, (10) to 1, and the other outcomes (00) and (11) to the empty string. After processing $X_1, X_2$, we will obtain either nothing, or a bit $Z_1$.

(a) Show that, if a bit is obtained, it is fair, i.e., $\Pr(Z_1 = 0) = \Pr(Z_1 = 1) = 1/2$.

In general we can process the $X$ sequence in successive $n$-tuples via a function $f : \{0, 1\}^n \rightarrow \{0, 1\}^*$ where $\{0, 1\}^*$ denote the set of all finite length binary sequences (including the empty string $\lambda$). [The case in (a) is the function $f(00) = f(11) = \lambda$, $f(01) = 0$, $f(10) = 1$. The function $f$ is chosen such that $(Z_1, \ldots, Z_K) = f(X_1, \ldots, X_n)$ are i.i.d., and fair (here $K$ may depend on $(X_1, \ldots, X_K)$).

(b) With $h_2(p) = -p \log p - (1 - p) \log(1 - p)$, prove the following chain of (in)equalities.

\[
\begin{align*}
\frac{n}{2} h_2(p) &= H(X_1, \ldots, X_n) \\
\geq H(Z_1, \ldots, Z_K, K) \\
= H(K) + H(Z_1, \ldots, Z_K | K) \\
= H(K) + E[K] \\
\geq E[K].
\end{align*}
\]

Consequently, on the average no more than $\frac{n}{2} h_2(p)$ fair bits can be obtained from $(X_1, \ldots, X_n)$.

(c) Find a good $f$ for $n = 4$.

Solution

(a) Since $\Pr(X_1 = 0, X_2 = 1) = \Pr(X_1 = 0) \Pr(X_2 = 1) = p(1 - p)$ and $\Pr(X_1 = 1, X_2 = 0) = \Pr(X_1 = 1) \Pr(X_2 = 0) = p(1 - p)$, the probability of $\Pr(Z_1 = 0) = \Pr(Z_1 = 1) = 1/2$. 

Since \( h_2(p) = -p \log p - (1 - p) \log (1 - p) = H(X_i) \),

\[
\begin{align*}
h_2(p) &= nH(X_i) \\
&= H(X_1, \ldots, X_n) \text{ [Independence of } X_i] \\
&\geq H(f(X_1, \ldots, X_n)) \text{ [Data Processing Inequality]} \\
&= H(Z_1, \ldots, Z_K, K) \\
&= H(K) + H(Z_1, \ldots, Z_K | K) \\
&= H(K) + \sum_k p(K = k)H(Z_1, \ldots, Z_K | K = k) \\
&= H(K) + \sum_k p(K = k)k [Z_1, \ldots, Z_k \text{ are i.i.d and fair when } K = k] \\
&\geq H(K) + E[K] \text{ [Independence of } X_i] \\
&\geq E[K] \text{ [Data Processing Inequality]}
\end{align*}
\]

(c) when \( n = 4 \), \((X_1, \ldots, X_4)\) have 16 outcomes with probabilities:

\[
\begin{align*}
1 \text{ case : } \Pr(0000) &= (1 - p)^4 \\
4 \text{ cases : } \Pr(0001) &= \cdots = \Pr(1000) = p(1 - p)^3 \\
6 \text{ cases : } \Pr(0011) &= \cdots = \Pr(1100) = p^2(1 - p)^2 \\
4 \text{ cases : } \Pr(0111) &= \cdots = \Pr(1110) = p(1 - p)^3 \\
1 \text{ case : } \Pr(1111) &= p^4
\end{align*}
\]

Now we can define the function as follows to get i.i.d. bits and produce as many bits we can:

\[
\begin{align*}
f(0000) &= f(1111) = \lambda \\
f(0011) &= 1 \\
f(1100) &= 0 \\
f(1001) &= f(1110) = f(0001) = 00 \\
f(1010) &= f(1101) = f(0010) = 01 \\
f(0110) &= f(1011) = f(0100) = 10 \\
f(0101) &= f(0111) = f(1000) = 11
\end{align*}
\]

Problem 6: Extremal characterization for Rényi entropy

Given \( s \geq 0 \), and a random variable \( U \) taking values in \( U \), with probabilities \( p(u) \), consider the distribution \( p_s(u) = p(u)^s/Z(s) \) with \( Z(s) = \sum_u p(u)^s \).

(a) Show that for any distribution \( q \) on \( U \),

\[
(1 - s)H(q) - sD(q||p) = -D(q||p_s) + \log Z(s).
\]

(b) Given \( s \) and \( p \), conclude that the left hand side above is maximized by the choice by \( q = p_s \) with the value \( \log Z(s) \),

The quantity

\[
H_s(p) := \frac{1}{1 - s} \log Z(s) = \frac{1}{1 - s} \log \sum_u p(u)^s
\]

is known as the Rényi entropy of order \( s \) of the random variable \( U \). When convenient, we will also write \( H_s(U) \) instead of \( H_s(p) \).
(c) Show that if $U$ and $V$ are independent random variables

$$H_s(UV) := H_s(U) + H_s(V).$$

[Here $UV$ denotes the pair formed by the two random variables — not their product. E.g., if $U = \{0, 1\}$ and $V = \{a, b\}$, $UV$ takes values in $\{0a, 0b, 1a, 1b\}.]

Solution

(a) We start from the left hand side of the equation:

$$\begin{align*}
(1 - s)H(q) - sD(q||p) &= (1 - s)\sum_u q(u) \log \frac{1}{q(u)} - s \sum_u q(u) \log \frac{q(u)}{p(u)} \\
&= \sum_u q(u) \left( (1 - s) \log \frac{1}{q(u)} - s \log \frac{q(u)}{p(u)} \right) \\
&= \sum_u q(u) \log \frac{p(u)^s}{q(u)} \\
&= \sum_u q(u) \log \frac{p_s(u)Z(s)}{q(u)} \\
&= \sum_u q(u) \log \frac{p_s(u)}{q(u)} + \sum_u q(u) \log Z(s) \\
&= -D(q||p_s) + \log Z(s)
\end{align*}$$

(b) We know that $D(q||p_s) \geq 0$, where equality achieves for $q = p_s$. The left hand side of above equation is maximized when $q = p_s$ and has value $\log Z(s)$.

(c) Since $U$ and $V$ are independent random variables, we have $p(u, v) = p(u)p(v)$.

$$H_s(UV) = \frac{1}{1 - s} \log \sum_{u,v} p(u,v)^s$$

$$= \frac{1}{1 - s} \log(\sum_u p(u)^s \sum_v p(v)^s)$$

$$= \frac{1}{1 - s} \log \sum_u p(u)^s + \frac{1}{1 - s} \log \sum_v p(v)^s$$

$$= H_s(U) + H_s(V)$$

Problem 7: Guessing and Rényi entropy

Suppose $X$ is a random variable taking values $K$ values $\{a_1, \ldots, a_K\}$ with $p_i = \Pr\{X = a_i\}$. We wish to guess $X$ by asking a sequence of binary questions of the type ‘Is $X = a_i$?’ until we are answered ‘yes’. (Think of guessing a password).

A guessing strategy is an ordering of the $K$ possible values of $X$; we first ask if $X$ is the first value; then if it is the second value, etc. Thus the strategy is described by a function $G(x) \in \{1, \ldots, K\}$ that gives the position (first, second, ... $K$th) of $x$ in the ordering. I.e., when $X = x$, we ask $G(x)$ questions to guess the value of $X$. Call $G$ the guessing function of the strategy.

For the rest of the problem suppose $p_1 \geq p_2 \geq \cdots \geq p_K$. 

8
(a) Show that for any guessing function $G$, the probability of asking fewer than $i$ questions satisfies

$$\Pr(G(X) \leq i) \leq \sum_{j=1}^{i} p_j$$

and equality holds for the guessing function $G^*$ with $G^*(a_i) = i$, $i = 1, \ldots, K$; this is the strategy that first guesses the most probable value $a_1$, then the next most probable value $a_2$, etc.

(b) Show that for any increasing function $f : \{1, \ldots, K\} \to \mathbb{R}$, $E[f(G(X))]$ is minimized by choosing $G = G^*$. [Hint: $E[f(G(X))] = \sum_{i=1}^{K} f(i) \Pr(G = i)$. Write $\Pr(G = i) = \Pr(G \leq i) - \Pr(G \leq i-1)$, to write the expectation in terms of $\sum_i [f(i) - f(i+1)] \Pr(G \leq i)$, and use (a).]

(c) For any $i$ and $s \geq 0$ prove the inequalities

$$i \leq \sum_{j=1}^{i} (p_j/p_i)^s \leq \sum_{j=1}^{i} (p_j/p_i)^s$$

(d) For any $\rho \geq 0$, show that

$$E[G^*(X)^\rho] \leq \left( \sum_{i} p_i^{1-s\rho} \right)^{\rho} \left( \sum_{j} p_j^{s} \right)^{\rho}.$$

for any $s \geq 0$. [Hint: write $E[G^*(X)^\rho] = \sum p_i i^\rho$, and use (c) to upper bound $i^\rho$]

(e) By a choosing $s$ carefully, show that

$$E[G^*(X)^\rho] \leq \left( \sum_{i} p_i^{1/(1+\rho)} \right)^{\frac{1+\rho}{\rho}} = \exp[\rho H_{1/(1+\rho)}(X)].$$

(f) Suppose $U_1, \ldots, U_n$ are i.i.d., each with distribution $p$, and $X = (U_1, \ldots, U_n)$. (I.e., we are trying to guess a password that is made of $n$ independently chosen letters.) Show that

$$\frac{1}{n\rho} \log E[G^*(U_1, \ldots, U_n)^\rho] \leq H_{1/(1+\rho)}(U_1)$$

[Hint: first observe that $H_{\alpha}(X) = n H_{\alpha}(U_1)$. In other words, the $\rho$-th moment of the number of guesses grows exponentially in $n$ with a rate upper bounded by in terms of the Rényi entropy of the letters.

It is possible a lower bound to $E[G^*(U_1, \ldots, U_n)^\rho]$ that establishes that the exponential upper bound we found here is asymptotically tight.

Solution

(a) The event that $G(X) \leq i$ contains the probability of $i$ distinct values.

$$\Pr(G(X) \leq i) = \sum_{j=1}^{i} \Pr(G(X) = j) \leq \sum_{j=1}^{i} p_j$$

(65)

as $p_1, \ldots, p_i$ are the $i$ largest probabilities. Equality holds for $G^*$, since $\Pr(G^* = i) = p_i$. 

9
(b) Note that $\Pr(G(X) \leq 0) = 0$ and $\Pr(G(X) \leq K) = 1$.

\[
E[f(G(X))] = \sum_{i=1}^{K} \Pr(G(X) = i)f(i) \quad (66)
\]

\[
= \sum_{i=1}^{K} (\Pr(G(X) \leq i) - \Pr(G(X) \leq i-1))f(i) \quad (67)
\]

\[
= \sum_{i=1}^{K-1} \Pr(G(X) \leq i)(f(i) - f(i+1)) + f(K) \quad (68)
\]

\[
\geq \sum_{i=1}^{K-1} \sum_{j=1}^{i} p_j(f(j) + f(j+1)) + f(K) \quad (69)
\]

where each $\Pr(G(X) \leq i) \leq \sum_{j=1}^{i} p_j$ with equality holding for $G = G^*$ according to (a) and $f(i) - f(i+1) \leq 0$ since $f$ is an increasing function. Hence, $E[f(G(X))]$ is minimized when $G = G^*$.

(c) Suppose we a distribution with probabilities \{p_1, \ldots, p_K\}. For any $i \in \{1, \ldots, K\}$ and $s > 0$:

\[
i = \sum_{j=1}^{i} 1^s \leq \sum_{j=1}^{i} (p_j/p_i)^s \leq \sum_{j=1}^{K} (p_j/p_i)^s = \sum_{j=1}^{K} (p_j/p_i)^s \quad (70)
\]

where the first inequality holds because $p_j/p_i \geq 1$ for each $1 \leq j \leq i$.

(d)

\[
E[G^*(X)^\rho] = \sum_i \Pr(G^*(X) = i)^\rho = \sum_i p_i^\rho \leq \sum_i p_i \left(\sum_j \frac{p_j}{p_i}\right)^\rho = \left(\sum_i p_i^{1-s\rho}\right) \left(\sum_j p_j^s\right)^\rho \quad (71)
\]

(e) Since inequality (71) holds for any $s > 0$, we can choose $s = \frac{1}{1+\rho}$ and get

\[
E[G^*(X)^\rho] \leq \left(\sum_i p_i^{\frac{1}{1+\rho}}\right) \left(\sum_j p_j^{\frac{1}{1+\rho}}\right)^\rho \quad (72)
\]

\[
= \left(\sum_i p_i^{\frac{1}{1+\rho}}\right)^{1+\rho} \quad (73)
\]

\[
= \exp \left[(1 + \rho) \log \sum_i p_i^{\frac{1}{1+\rho}}\right] \quad (74)
\]

\[
= \exp \left[\rho \frac{1}{1 - \frac{1}{1+\rho}} \log \sum_i p_i^{\frac{1}{1+\rho}}\right] \quad (75)
\]

\[
= \exp \left[\rho H_{1/(1+\rho)}(X)\right] \quad (76)
\]

(f) Follow the hint that $H_{\alpha}(X) = nH_{\alpha}(U_1)$:

\[
\frac{1}{n\rho} \log E[G^*(U_1, \ldots, U_n)^\rho] \leq \frac{1}{n\rho} \log \exp[\rho H_{1/(1+\rho)}(X)] \quad (77)
\]

\[
= \frac{1}{n} H_{1/(1+\rho)}(X) \quad (78)
\]

\[
= H_{1/(1+\rho)}(U_1) \quad (79)
\]

10
Problem 8: Gaussian variance estimation

Consider estimating the mean $\mu$ and variance $\sigma^2$ from $n$ independent samples $(X_1, \ldots, X_n)$ of a Gaussian with this mean and variance.

(a) Show that $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ is an unbiased estimator of $\mu$.

(b) Show that

$$S_n^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

is a biased estimator of $\sigma^2$ whereas

$$S_{n-1}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

is an unbiased estimator of $\sigma^2$.

(c) Show that $S_n^2$ has a lower mean squared error than $S_{n-1}^2$. Thus it is possible that a biased estimator may be better than an unbiased one.

Solution

(a) To show that $\bar{X}$ is an unbiased estimator of $\mu$, we have to prove that $E[\bar{X}] = \mu$.

$$E[\bar{X}] = E\left[ \frac{1}{n} \sum_{i=1}^{n} X_i \right] = \frac{1}{n} \sum_{i=1}^{n} E[X_i] = \frac{1}{n} \sum_{i=1}^{n} \mu = \mu$$ (80)

(b) Without loss of generality, we can assume that $\mu = 0$, since otherwise replacing $X_i = X_i - \mu$ also replaces $\bar{X} = \bar{X} - \mu$ and leaves $S_n^2$ and $S_{n-1}^2$ unchanged.

$$E[S_n^2] = E\left[ \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2 \right] = \frac{1}{n} \sum_{i=1}^{n} E[X_i^2 - 2X_i \bar{X} + \bar{X}^2]$$ (81)

$$= E[X_1^2] - 2E[X_1 \bar{X}] + E[\bar{X}^2]$$ (82)

$$= E[X_1^2] - 2E[X_1 \frac{1}{n} \sum_{i=1}^{n} X_i] + E[\frac{1}{n} \sum_{i=1}^{n} X_i]^2]$$ (83)

$$= \frac{n-2}{n} E[X_1^2] + \frac{1}{n^2} \sum_{i=1}^{n} E[X_i^2]$$ (84)

$$= \frac{n-1}{n} \sigma^2$$ (85)

where all cross-term $E[X_i X_j] = 0$, for $i \neq j$. Thus $S_n^2$ is a biased estimator of $\sigma^2$. And obviously, it also shows that $S_{n-1}^2$ is an unbiased estimator of $\sigma^2$, since

$$E[S_{n-1}^2] = E[\frac{n}{n-1} S_n^2] = \frac{n}{n-1} E[S_n^2] = \sigma^2$$ (86)
(c) Also, we assume that the mean is 0. The mean squared error of $S_n^2$ can be computed as

\[ E[(S_n^2 - \sigma^2)^2] = E[S_n^2 - 2S_n^2\sigma^2 + \sigma^4] \]

\[ = E[S_n^2] - 2\sigma^2E[S_n^2] + \sigma^4 \] (87)

\[ = E[S_n^2] - 2\frac{n-1}{n}\sigma^4 + \sigma^4 \] (88)

\[ = E[S_n^2] - \frac{n-2}{n}\sigma^4 \] (89)

\[ = E[S_n^2] - \frac{n-2}{n}\sigma^4 \] (90)

Now we compute

\[ E[S_n^4] = \frac{1}{n^2}E[\sum_i \sum_j (X_i^2 - 2X_i\bar{X}) (X_j^2 - 2X_j\bar{X})] \] (91)

\[ = \frac{1}{n^2} \{ nE[X_1^4] + n(n-1)E[X_1^2X_2^2] - 4nE[X_1^2\bar{X}] - 4n(n-1)E[X_1^2X_2\bar{X}] \} - (2n^2 + 4n)E[X_1^2\bar{X}^2] + 4n(n-1)E[X_1X_2\bar{X}^2] - 4n^2E[X_1\bar{X}^2] + n^2E[\bar{X}^4] \] (92)

\[ = \ldots \] (93)

\[ = \frac{n^2 - 1}{n^4}\sigma^4 \] (94)

where $E[X_1^4] = 3\sigma^4$ and $E[X_1^2X_2^2] = E[X_1^2]^2 = \sigma^4$, since $X_i$ is Gaussian random variable and we assumed $\mu = 0$. Therefore,

\[ E[(S_n^2 - \sigma^2)^2] = \frac{n^2 - 1}{n^2}\sigma^4 - \frac{n-2}{n}\sigma^4 = \frac{2n-1}{n^2}\sigma^4 \] (95)

Now we can compute mean squared error for $S_{n-1}^2$ in an efficient way

\[ E[(S_{n-1}^2 - \sigma^2)^2] = E[S_{n-1}^4 - 2S_{n-1}^2\sigma^2 + \sigma^4] \]

\[ = E[S_{n-1}^4] - 2\sigma^2E[S_{n-1}^2] + \sigma^4 \] (96)

\[ = \frac{n^2}{(n-1)^2}E[S_n^4] - \sigma^4 \] (97)

\[ = \frac{2}{n-1}\sigma^4 \] (98)

Hence, for $n > 1$, we have

\[ E[(S_n^2 - \sigma^2)^2] - E[(S_{n-1}^2 - \sigma^2)^2] = \left[ \frac{2n-1}{n^2} - \frac{2}{n-1} \right]\sigma^4 \] (99)

\[ = -\frac{3n + 1}{n^2(n-1)}\sigma^4 \] (100)

\[ < 0 \] (101)

Thus $S_n^2$ has a smaller mean squared error than $S_{n-1}^2$. 

---

12