Chapter 2 - Questions and Solutions

Edwin Fennell

2.1 Derive the mean and variance for the binomial distribution

The binomial distribution $B(n,p)$ is defined on $\mathbb{N}$ by the probability mass function

$$p(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & x = 0, 1, ..., n \\ 0 & else \end{cases}$$

More abstractly, it is the distribution of total successes of n independent Bernoulli trials with success probability p. This immediately gives the expectation as

$$np$$

since the expectation of a sum of random variables is just the sum of the expectation. Since the Bernoulli trials are independent, the variance of their sum is also just the sum of their variances, so

$$np(1-p)$$

2.2 Derive the mean and variance for the uniform distribution

The uniform distribution $N(a,b)$ (for $a < b$) is defined on $\mathbb{R}$ by the probability density function

$$p(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & else \end{cases}$$

We therefore calculate the expectation as

$$\int_a^b \frac{x}{b-a} dx = \left[\frac{x^2}{2(b-a)} \right]_a^b = \frac{a+b}{2}$$

This result is also reasonably clear from a symmetry argument.

We also have

$$\int_a^b \frac{x^2}{b-a} dx = \left[\frac{x^3}{3(b-a)} \right]_a^b = \frac{a^2+ab+b^2}{3}$$

which gives us the variance as

$$\frac{a^2+ab+b^2}{3} - \frac{(a+b)^2}{4} = \frac{a^2-2ab+b^2}{12} = \frac{(b-a)^2}{12}$$

2.3 Derive the mean and covariance matrix for the multivariate normal distribution

The multivariate Gaussian $N(\mu, \Sigma)$ is defined on $\mathbb{R}^k$ by the probability density function

$$p(x) = \frac{1}{(2\pi)^{\frac{k}{2}}|\Sigma|^{\frac{1}{2}}} exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

Then the ith element of the expectation is

$$\int x_i \frac{1}{(2\pi)^{\frac{k}{2}}|\Sigma|^{\frac{1}{2}}} exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right) dx$$

Σ is symmetric and positive definite, so by Cholesky composition we can write

$$\Sigma^{-1} = H^T H$$

If we set $y = H(x-\mu)$ then $dx = |H^{-1}|dy = |\Sigma|^{\frac{1}{2}}dy$, and changing variables in the above integral gives us

$$\int (H^{-1}_{ik}y_k + \mu_i) \frac{1}{(2\pi)^{\frac{k}{2}}} exp\left(-\frac{1}{2}y \cdot y\right) dy$$

We now use symmetry to discard the Hy and consider all elements to obtain our expectation as

$$\mu$$

The (i,j)th element of the covariance matrix is given by

$$\int (x_i-\mu_i)(x_j-\mu_j) \frac{1}{(2\pi)^{\frac{k}{2}}|\Sigma|^{\frac{1}{2}}} exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right) dx$$

Performing the previous substitution gives

$$\int H^{-1}_{ia}y_a H^{-1}_{jb}y_b \frac{1}{(2\pi)^{\frac{k}{2}}} exp\left(-\frac{1}{2}y \cdot y\right) dy$$

where we are implicitly summing over both a and b. Consider what happens in our implicit sum whenever $a \neq b$ - the integrand is an odd function of y_a (and y_b - it doesn't matter which one we pick) - and therefore the whole integral evaluates to 0. Therefore in order to evaluate the integral,

we just sum over all cases where $a = b$. This gives the value of the integral as

$$\sum_a \left(H_{ia}^{-1}(H_{aj}^T)^{-1} \cdot \int \frac{y_a \cdot y_a}{(2\pi)^{\frac{1}{2}}} exp\left(-\frac{1}{2} y_a \cdot y_a\right) dy_a \cdot \prod_{b \neq a} \left(\int \frac{1}{(2\pi)^{\frac{1}{2}}} exp\left(-\frac{1}{2} y_b \cdot y_b\right) dy_b \right) \right)$$

All terms in the rightmost product evaluate to 1 (consider the pdf of a 1D Gaussian). The middle term also evaluates to 1 (consider the variance of a 1D Gaussian), and therefore the sum evaluates to

$$\sum_a H_{ia}^{-1}(H_{aj}^T)^{-1} = (HH^T)_{ij}^{-1} i = \Sigma_{ij}$$

And so the covariance matrix is just

$$\Sigma$$

2.4 Show that the mean and variance of the beta distribution with parameters a and b are

$$\frac{\boldsymbol{a}}{\boldsymbol{a+b}}$$

and

$$\frac{\boldsymbol{ab}}{\boldsymbol{(a+b)^2(a+b+1)}}$$

respectively

The beta distribution $Beta(a, b)$ is defined on $[0, 1]$ by the probability density function

$$\frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}$$

where

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

The expectation is given by

$$\int_0^1 \frac{x^a(1-x)^{b-1}}{B(a,b)} dx = \int_0^1 \frac{\Gamma(a+b)x^a(1-x)^{b-1}}{\Gamma(a)\Gamma(b)} dx$$

We note that $\Gamma(a+b+1) = (a+b)\Gamma$ and $\Gamma(a+1) = a\Gamma(a)$ so our integral becomes

$$\frac{a}{a+b} \int_0^1 \frac{\Gamma(a+b+1)x^a(1-x)^{b-1}}{\Gamma(a+1)\Gamma(b)} dx$$

The integrand is just the density function of $Beta(a+1,b)$, so the above expression just reduces to

$$\frac{a}{a+b}$$

Similarly, we have

$$\int_0^1 \frac{x^{a+1}(1-x)^{b-1}}{B(a,b)}dx = \int_0^1 \frac{\Gamma(a+b)x^{a+1}(1-x)^{b-1}}{\Gamma(a)\Gamma(b)}dx$$

We have $\Gamma(a+b+2) = (a+b+1)(a+b)\Gamma(a+b)$, $\Gamma(a+2) = (a+1)(a)\Gamma(a)$, so this integral is just equal to

$$\frac{a(a+1)}{(a+b)(a+b+1)}\int_0^1 \frac{\Gamma(a+b+2)x^{a+1}(1-x)^{b-1}}{\Gamma(a+2)\Gamma(b)}dx = \frac{a(a+1)}{(a+b)(a+b+1)}$$

and therefore the variance is

$$\frac{a(a+1)}{(a+b)(a+b+1)} - \frac{a^2}{(a+b)^2} = \frac{ab}{(a+b)^2(a+b+1)}$$

2.5 Show that the normalising constant in the beta distribution with parameters a,b is given by

$$\boldsymbol{\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}}$$

The beta distribution $Beta(a,b)$ is defined on $[0,1]$ by the probability density function

$$\frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}$$

We are required to prove that

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

This is equivalent to showing

$$\int_{\mathbb{R}} x^{a-1}(1-x)^{b-1}dx = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

since any probability distribution must integrate to 1 over its domain. The gamma function is defined by

$$\Gamma(a) = \int_{\mathbb{R}} x^{a-1}e^{-x}dx$$

so we can multiply through by $\Gamma(a+b)$ and expand to obtain our proposition as an equality between two double integrals:

$$\int_{\mathbb{R}}\int_{\mathbb{R}} x^{a-1}(1-x)^{b-1}y^{a+b-1}e^{-y}dxdy = \int_{\mathbb{R}}\int_{\mathbb{R}} u^{a-1}e^{-u}v^{b-1}e^{-v}dudv$$

On the right-hand side we can substitute $u = xy$ and $v = (1-x)y$, which gives us the following relation between area elements:

$$dudv = \begin{vmatrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} \\ \frac{\partial v}{\partial x} & \frac{\partial v}{\partial y} \end{vmatrix} dxdy = \begin{vmatrix} y & x \\ -y & (1-x) \end{vmatrix} dxdy = y \cdot dxdy$$

Using this and fully substituting leaves the right-hand side as

$$\int_{\mathbb{R}}\int_{\mathbb{R}} (xy)^{a-1}((1-x)y)^{b-1}y \cdot dxdy$$

which is just the left-hand side of the equation. □

2.6 Show that the mean and variance of the pdf of the gamma distribution with parameters a,b

$$\boldsymbol{p(x) = \begin{cases} \frac{b^a}{\Gamma(a)}x^{a-1}e^{-bx} & x > 0 \\ 0 & else \end{cases}}$$

are given by

$$\boldsymbol{\frac{a}{b}}$$

and

$$\boldsymbol{\frac{a}{b^2}}$$

respectively.

The expectation is given by

$$\int_0^\infty \frac{b^a}{\Gamma(a)}x^a e^{-bx}dx$$

We can rewrite this as

$$\frac{a}{b}\int_0^\infty \frac{b^{a+1}}{a\Gamma(a)}x^a e^{-bx}dx$$

Since $a\Gamma(a) = \Gamma(a+1)$, the integrand is simply the pdf of the gamma distribution with parameters $a+1, b$, and therefore the integral evaluates to 1. The desired result immediately follows. The variance is given by

$$\int_0^\infty \frac{b^a}{\Gamma(a)}x^{a+1}e^{-bx}dx - \frac{a^2}{b^2}$$

We rewrite this expression as

$$\frac{a(a+1)}{b^2}\int_0^\infty \frac{b^{a+2}}{a(a+1)\Gamma(a)}e^{-bx}dx - \frac{a^2}{b^2}$$

Using the fact that $\Gamma(a+2) = a(a+1)\Gamma(a)$, we see that the integrand is just the pdf of a gamma distribution with parameters $a+2, b$. Therefore we calculate the original variance as

$$\frac{a(a+1)-a}{b^2} = \frac{a}{b^2}$$

2.7 Show that the mean and covariance of a Dirichlet pdf with parameters a_k, $k = 1, 2, ..., K$ are given by

$$\frac{\boldsymbol{a}}{\bar{\boldsymbol{a}}}$$

and

$$\frac{\bar{\boldsymbol{a}}\boldsymbol{\delta_3 a} - \boldsymbol{a a^T}}{\bar{\boldsymbol{a}}^2(\boldsymbol{1}+\bar{\boldsymbol{a}})}$$

respectively

($\bar{a}$ is the sum of all elements of a and δ_3 is the 3D Kronecker delta)

The pdf of this Dirichlet distribution over $\mathbb{R}^k$ is defined by

$$\begin{cases} \frac{\Gamma(\bar{a})}{\Gamma(a_1)...\Gamma(a_k)}\prod_{k=1}^K x_k^{a_k-1} & (\forall i, 0 \le x_i \le 1) \wedge (\sum_{i=1}^K x_i = 1) \\ 0 & else \end{cases}$$

We won't write out complete integrals here because they are horrible - we will instead use some neat trickery to rewrite the integrands of expectations we want in terms of other Dirichlet pdfs, which conveniently integrate to 1 across our domain. We will therefore write our compound integral using a single integral sign over our domain D, and just write dx for our hypervolume element.

The ith component of the mean is given by

$$\int_D x_i \frac{\Gamma(\bar{a})}{\Gamma(a_1)...\Gamma(a_k)}\prod_{k=1}^K x_k^{a_k-1}dx$$

If we define a vector b s.t. $b_j = a_j$ for $j \neq i$, and $b_i = a_i + 1$, then we note that the above integral is just

$$\int_D \frac{\Gamma(\bar{a})}{\Gamma(a_1)...\Gamma(a_k)}\prod_{k=1}^K x_k^{b_k-1}dx$$

We now define $\bar{b}$ as the sum of the elements of b. $\bar{b} = \bar{a} + 1$, so we can use the our nifty relation on gammas

$$\Gamma(x+1) = x\Gamma(x)$$

to rewrite our integral as

$$\frac{a_i}{\bar{a}} \int_D \frac{\Gamma(\bar{b})}{\Gamma(b_1)...\Gamma(b_k)} \prod_{k=1}^{K} x_k^{b_k - 1} dx$$

The integrand is just the pdf of the Dirichlet distribution with parameter vector b, and so the integral evaluates to 1. This gives our mean vector as just

$$\frac{a}{\bar{a}}$$

The i, jth element of the covariance is given by

$$\int_D x_i x_j \frac{\Gamma(\bar{a})}{\Gamma(a_1)...\Gamma(a_k)} \prod_{k=1}^{K} x_k^{a_k - 1} dx - \frac{a_i a_j}{\bar{a}^2}$$

If $i \neq j$, we can define a vector c s.t. $c_k = a_k$ for $k \neq i, j$, and $c_k = a_k + 1$ for $k = i, j$. Then we again use our gamma relation to rearrange the above integral to

$$\frac{a_i a_j}{\bar{a}(\bar{a}+1)} \int_D \frac{\Gamma(\bar{c})}{\Gamma(c_1)...\Gamma(c_k)} \prod_{k=1}^{K} x_k^{c_k - 1} dx = \frac{a_i a_j}{\bar{a}(\bar{a}+1)}$$

since the integrand above is just the pdf of the Dirichlet distribution with parameter vector c. This gives the value of the i, jth element as

$$\frac{a_i a_j}{\bar{a}} \left(\frac{1}{\bar{a}+1} - \frac{1}{\bar{a}} \right) = -\frac{a_i a_j}{\bar{a}^2(\bar{a}+1)}$$

If instead $i = j$, then define d s.t. $d_k = a_k$ for $k \neq i$ and $d_i = a_i + 2$. So our expression for the i, ith element becomes

$$\frac{a_i(a_i+1)}{\bar{a}(\bar{a}+1)} \int_D \frac{\Gamma(\bar{d})}{\Gamma(d_1)...\Gamma(d_k)} \prod_{k=1}^{K} x_k^{d_k - 1} dx - \frac{a_i^2}{\bar{a}^2}$$

As with the previous examples, the integral just evaluates to 1, so this simplifies to

$$\frac{a_i}{\bar{a}(\bar{a}+1)} + \frac{a_i^2}{\bar{a}} \left(\frac{1}{\bar{a}+1} - \frac{1}{\bar{a}} \right) = \frac{\bar{a}a_i - a_i^2}{\bar{a}^2(\bar{a}+1)}$$

When we consider this across all i, j, we get the proposed expression for covariance.

2.8 Show that the sample mean of N i.i.d. samples, is an unbiased estimator with variance that tends to zero asymptotically as $N \to \infty$

Suppose we are sampling from an underlying distribution A, and our ith sample is denoted by A_i - all independent and distributed as A. Our sample mean is defined by

$$\bar{A} = \frac{1}{n}\sum_{i=1}^{K} A_i$$

Expectation distributes over all sums, therefore

$$E(\bar{A}) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{K} A_i\right) = \frac{1}{n}\sum_{i=1}^{K}\mathbb{E}(A_i) = \frac{n * \mathbb{E}(A)}{n} = \mathbb{E}(A)$$

since all the A_i are distributed as A is. Therefore the sample mean is unbiased. This tells us that the variance of the sample mean is given by

$$\mathbb{E}((\bar{A} - \mathbb{E}(A)(\bar{A} - \mathbb{E}(A)))$$

Using again the distribution of expectation over addition and scalar multiples, we can rewrite this as

$$\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbb{E}\left(\frac{A_i - \mathbb{E}(A)}{n}\frac{A_j - \mathbb{E}(A)}{n}\right)$$

By independence of the A_i and the fact that all the A_i have mean $\mathbb{E}(A)$, the cross-terms cancel and we are left with

$$\sum_{i=1}^{j}\frac{Var(A_i)}{n^2} = \frac{n * Var(A_i)}{n^2} = \frac{Var(A)}{n}$$

This converges to 0 as $N \to \infty$

2.9 Show that for WSS processes

$$\boldsymbol{r(0) \geq |r(k)|, \forall k \in \mathbb{Z}}$$

and that for jointly-WSS processes

$$\boldsymbol{r_u(0)r_v(0) \geq |r_{uv}(k)|, \forall k \in \mathbb{Z}[sic]}$$

We first consider the second part of the assertion. Note that it is not true (consider two processes with a constant value of $\frac{1}{2}$). The mathematical expression should probably read

$$r_u(0)r_v(0) \geq |r_{uv}(k)|^2$$

The left-hand side of the inequality is equal to

$$\mathbb{E}(u_k u_k^*)\mathbb{E}(v_0 v_0^*)$$

while the right-hand side is equal to

$$|\mathbb{E}(u_k v_0^*)|^2$$

Therefore the result follows trivially from the Cauchy-Schwarz inequality, since $\langle u, v\rangle = E(uv^*)$ is an inner product space. The first part of the assertion is a special case of this where $u = v$ (a WSS process is necessrily WSS with itself).

2.10 Show that the autocorrelation of the output of a linear system, with impulse response $w_n, n \in \mathbb{Z}$ is related to the autocorrelation of the input WSS process, via

$$\boldsymbol{r_d(k) = r_u(k) * w_k * w_{-k}^*}$$

By definition

$$r_d(k) = \mathbb{E}(d_k d_0^*) = \mathbb{E}\left(\sum_{i=-\infty}^{\infty}\sum_{j=-\infty}^{\infty} w_i^* u_{k-i} w_j u_{-j}^*\right)$$

We now make the substitution $j = i - a$, giving us

$$r_d(k) = \sum_{a=-\infty}^{\infty}\sum_{i=-\infty}^{\infty} \mathbb{E}(u_{k-i}u_{a-i}^*) w_i^* w_{i-a}$$

Note that $\mathbb{E}(u_{k-i}u_{a-i}^*) = r_u(k-a)$, and this is independent of i, so we have

$$r_d(k) = \sum_{a=-\infty}^{\infty} r_u(k-a) \sum_{i=-\infty}^{\infty} w_i^* w_{i-a}$$

The inner sum is just $w_a * w_{-a}$, so

$$r_d(k) = \sum_{a=-\infty}^{\infty} r_u(k-a)(w_a * w_{-a})$$

Now we substitute $n = k - a$ to obtain

$$r_d(k) = \sum_{n=-\infty}^{\infty} r_u(n)(w_{k-n} * w_{n-k})$$

which simplifies to

$$r_d(k) = r_u(k)^* * w_k * w_{-k}$$

We note that

$$(a^* * b^*) = (a * b)^* = (b * a)$$

Therefore we have for any a, b, c

$$(a^* * b) * c = (c * (a^* * b))^* = c^* * (a^* * b)^* = c^* * b * a^*$$

By applying this relation twice, we see that

$$r_u(k)^* * w_k * w_{-k} = w^*_{-k} * w_k * r_u(k)^* = r_u(k) * w_k * w^*_{-k}$$

as required.

2.11 Show that

$$\boldsymbol{ln(x) \leq x - 1}$$

Consider the function

$$f(x) = x - 1 - ln(x)$$

The statement we are required to prove is equivalent to f being everywhere non-negative. f is differentiable everywhere it is defined (on $(0, \infty)$), and so we have

$$f'(x) = 1 - \frac{1}{x}$$

We note that f' is negative for $x < 1$ and positive for $x > 1$. Therefore the minimum of f occurs at $x = 1$. $f(1) = 0$, therefore $f \geq 0$ everywhere. □

2.12 Show that

$$\boldsymbol{I(X, Y) \geq 0}$$

The average mutual information $I(X, Y)$ is defined by

$$I(X, Y) = \sum_{x \in X} \sum_{y \in Y} P(x, y) \log \left(\frac{P(x, y)}{P(x)P(y)} \right)$$

We can restrict this sum to combinations of x, y s.t. $P(x, y) > 0$. This condition implies $P(x)P(y) > 0$ Now we can use the solution to problem 2.13, since

$$\sum_{x,y \ s.t. \ P(x,y)>0} P(x, y) = 1$$

and

$$\sum_{x,y \ s.t. \ P(x,y)>0} P(x)P(y) \leq 1$$

and $P(x, y) \geq 0, P(x)P(y) \geq 0 \ \forall x, y$ we have

$$\sum_{x,y \ s.t. \ P(x,y)>0} P(x, y) log(P(x, y)) \geq \sum_{x,y \ s.t. \ P(x,y)>0} P(x, y) log(P(x)P(y))$$

which immediately rearranges to the required result.

2.13 Show that if $a_i, b_i, i = 1, 2, ..., M$ are positive numbers with

$$\sum_{i=1}^{M} a_i = 1, \sum_{i=1}^{M} b_i \leq 1$$

then

$$-\sum_{i=1}^{M} a_i \log(a_i) \leq -\sum_{i=1}^{M} a_i \log(b_i)$$

This follows straightforwardly from the concavity of log. Since the a_i are positive and sum to 1 we have

$$\sum_{i=1}^{M} a_i \log\left(\frac{b_i}{a_i}\right) \leq \log\left(\sum_{i=1}^{M} a_i \frac{b_i}{a_i}\right) = \log\left(\sum_{i=1}^{M} b_i\right) \leq \log(1) = 0$$

The initial proposition follows from a simple rearrangement.

2.14 Show that the maximum value of the entropy of a random variable occurs if all possible outcomes are equiprobable

First we show that a p.d.f with greatest entropy exists. We note that $-x \log x$ is bounded on $(0, 1]$, and therefore any sequence of random variables (on a finite state space) cannot have unboundedly increasing entropy.

Now, consider a sequence of random variables with monotonically increasing entropy. Note that all these random variables are members of $[0, 1]^n$, where n is the size of the state space. $[0, 1]^n$ is compact, so we can find a convergent subsequence from our sequence of random variables. Entropy is a continuous function of distribution, so the "entropy" of the limit is guaranteed to be the limit of the entropies. All that remains to check is that the limit of our sequence is a valid pdf - we only know that is is a member of $[0, 1]^n$. However, we note that sum and minimum element are both continuous functions in $[0, 1]^n$, and that the defining features of a random variable (considered as a member of $[0, 1]^n$) are that all elements individually exceed 0 and sum to 1. This is true of every single member of our convergent subsequence, and therefore holds for the limit. Therefore we have constructed a random variable with entropy greater than that of every random variable in our original sequence. Thus there must exist a random variable on this state space that maximises entropy.

Suppose that X is the random variable on M outcomes with the highest entropy. Suppose X takes outcome x_i with probability p_i for $i = 1, 2, ..., M$. Now suppose that not all the p_i are equal. By reordering, $p_1 < p_2$. Now we define a random variable Y taking the same values as X with the same probabilities, apart from x_1, x_2, which Y takes with probabilities $\frac{p_1+p_2}{2}$

each. We note that the function $f(x) = x\log(x)$ is strictly convex in the range $[0, 1]$ (second derivative is $1/x$) so we get

$$\frac{p_1 + p_2}{2} \log\left(\frac{p_1 + p_2}{2}\right) < \frac{1}{2}(p_1 \log(p_1) + p_2 \log(p_2))$$

which when we consider the expressions for entropy of both X and Y, tells us that X has a lower entropy than Y. This is a contradiction, so all the p_i must be equal. $\square$

2.15 Show that of all the pdfs that describe a random variable in an interval $[a, b]$, the uniform one maximises the entropy.

Jensen's inequality states that if (Ω, A, μ) is a probability space, then for a real-valued μ-integrable function g and real-valued convex function ϕ

$$\int_\Omega \phi \circ g \cdot d\mu \geq \phi\left(\int_\Omega g \cdot d\mu\right)$$

Now, given a random variable X with pdf p defined on $[a, b]$, we obtain the entropy of X as

$$-\int_a^b p(x)\log(p(x))dx$$

We can make a change of variables $y = \frac{x-a}{b-a}$ to rewrite this as

$$-\int_0^1 (b-a)\cdot p((b-a)y + a)\log(p((b-a)y + a))dy$$

We note that is an integral the probability space $[0, 1]$ (equipped with the uniform measure). We also note that $g(y) = p((b-a)y + a)$ is integrable with respect to this measure (integrates to $\frac{1}{b-a}$) and that $\phi(x) = x\log(x)$ is convex on the real line (we will only be applying it to positive x so this is valid). Thus the integral is exactly of the form required to use Jensen's inequality, and we find that

$$entropy(X) \leq -(b-a)I\log(I)$$

where

$$I = \int_0^1 p((b-a)y + a)dy$$

This is equal to $\frac{1}{b-a}$, so we have $entropy(X) \leq -log(\frac{1}{b-a})$ The right-hand side is the entropy of the uniform pdf on $[a, b]$, so we have proven our proposition.