CS-6190: Homework 0

James Brissette

September 9, 2019

1 Warm Up

1.

$$p(A \cup B) \leq p(A) + p(B)$$

We know from set probability theory that $p(A \cup B) = p(A) + p(B) - p(A \cap B)$, which allows us to examine two possible scenarios about set A and set B – one in which the events are independent and one in which they're not. In either case, we know $0 \leq p(A \cap B) \leq 1$, and the following holds:

$$p(A \cup B) = p(A) + p(B) \text{ for A,B independent}$$
$$p(A \cup B) < p(A) + p(B) \text{ for } p(A \cap B) > 0$$

$$p(A \cap B) \leq p(A)$$

Again, we can imagine sketching out two sets and observing that at either extreme, $0 \leq p(A \cap B) \leq 1$. Under conditions of independence, we know $p(A \cap B) = p(A)p(B)$ and furthermore since we know $0 \leq p(B) \leq 1$ it follows that when $p(B) = 0$, $p(A \cap B) = 0$ and when $p(B) = 1$, $p(A \cap B) = p(A)$. Given this, we can see the for any value of $p(B)$ we have that $0 \leq p(A \cap B) \leq p(A)$ and so $p(A \cap B) \leq p(A)$.

$$p(A \cap B) \leq p(B)$$

We just showed when you hold $p(A)$ constant and vary $p(B)$ you get the $p(A \cap B) \leq p(A)$. Using that same logic and the fact that under conditions of independence, $p(A \cap B) = p(A)p(B)$, if we hold $p(B)$ constant and vary only the value of $p(A)$ it follows naturally from above that $p(A \cap B) \leq p(B)$.

2. For the following induction, we will first prove the base case when $i = 2$, and then the general case using $i = n - 1$

$$p\Big(\bigcup_{i=1}^{n} A_i \Big) \leq \sum_{i=1}^{n} p(A_i)$$

Base case:

$$p(A_1 \cup A_2) = p(A_1) + p(A_2) - p(A_1 \cap A_2)$$

which we know from Q1 above is

$$p(A_1 \cup A_2) \leq p(A_1) + p(A_2)$$

Next, we assume our base case is correct and generalize our induction step as follows:

$$p(\bigcup_{i=1}^{n-1} A_i) \le \sum_{i=1}^{n-1} p(A_i)$$

At the point where we run into $i = n$, based on our base case we get:

$$p(\bigcup_{i=1}^{n} A_i) = p(\bigcup_{i=1}^{n-1} A_i) + p(A_n) - p(\bigcap_{i=0}^{n} A_i)$$

$$p(\bigcup_{i=1}^{n} A_i) \le p(\bigcup_{i=1}^{n-1} A_i) + p(A_n)$$

Substituting in our induction step we get:

$$p(\bigcup_{i=1}^{n} A_i) \le \sum_{i=1}^{n-1} p(A_i) + p(A_n)$$

which is equivalent to:

$$p(\bigcup_{i=1}^{n} A_i) \le \sum_{i=1}^{n} p(A_i)$$

3. For this question we have the following:

	$Y = 0$	$Y = 1$
$X = 0$	3/10	1/10
$X = 1$	2/10	4/10

(a) i. Marginal Distributions:

	$p(X)$	$p(Y)$
0	40%	50%
1	60%	50%

ii. Conditional Distributions:

	$p(X|Y)$	$p(Y|X)$
0	10%	20%
1	40%	40%

iii. Expectation/Variance:

$$\mathbb{E}(X) = 0\left(\frac{4}{10}\right) + 1\left(\frac{6}{10}\right) = 60\%$$

$$\mathbb{V}(X) = 0^2\left(\frac{4}{10}\right) + 1^2\left(\frac{6}{10}\right) - \left(\frac{3}{5}\right)^2 = 36\%$$

$$\mathbb{E}(Y) = 0\left(\frac{5}{10}\right) + 1\left(\frac{5}{10}\right) = 50\%$$

$$\mathbb{V}(Y) = 0^2\left(\frac{5}{10}\right) + 1^2\left(\frac{5}{10}\right) - \left(\frac{1}{2}\right)^2 = 25\%$$

iv. Conditional Expectation/Variance:

$$\mathbb{E}(Y|X=0) = 0\left(\frac{3}{10}\right) + 1\left(\frac{1}{10}\right) = 10\%$$

$$\mathbb{V}(Y|X=0) = 0^2\left(\frac{3}{10}\right) + 1^2\left(\frac{1}{10}\right) - \left(\frac{1}{10}\right)^2 = 9\%$$

$$\mathbb{E}(Y|X=1) = 0\left(\frac{2}{10}\right) + 1\left(\frac{4}{10}\right) = 40\%$$

$$\mathbb{V}(Y|X=1) = 0^2\left(\frac{2}{10}\right) + 1^2\left(\frac{4}{10}\right) - \left(\frac{4}{10}\right)^2 = 24\%$$

v. Covariance:

$$=(0)(0)(3/10) + (0)(1)(1/10)+$$
$$(1)(0)(2/10) + (1)(1)(4/10) - (3/5)(5/10)$$
$$=(4/10) - (3/10) = 10\%$$

(b) X and Y are NOT independent, and we can see this from the fact that there exists a non-zero covariance between variables.

(c) When X is not assigned a specific value then $\mathbb{E}(Y|X)$ and $\mathbb{V}(Y|X)$ become Expectations and Variances of a Random Variable, making them Random Variables themselves. If X is non-constant (e.g. a Random Variable), then $\mathbb{E}(Y|X)$ and $\mathbb{V}(Y|X)$ are NOT constants.

4. (a)

$$\mathbb{E}(Y) = \int_{-\infty}^{\infty} g(x)p(x)dx$$

$$= \int_{-\infty}^{\infty} e^{-x^2}\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{-(x-\mu)^2}{2\sigma^2}}dx$$

$$= \int_{-\infty}^{\infty} \frac{e^{-x^2}e^{\frac{-x^2}{2}}}{\sqrt{2\pi}}dx$$

$$= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-3x^2}dx$$

$$= \frac{1}{\sqrt{3}}$$

(b)

$$\mathbb{V}(Y) = \mathbb{E}[X^2] - \mu^2$$

$$= \int_{-\infty}^{\infty} [g(x)]^2 p(x)dx - \mu^2$$

$$= \int_{-\infty}^{\infty} \frac{[e^{-x^2}]^2 e^{\frac{-x^2}{2}}}{\sqrt{2\pi}}dx - \frac{1}{3}$$

$$= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-2x^2}e^{\frac{-x^2}{2}}dx - \frac{1}{3}$$

$$= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{\frac{-5x^2}{2}}dx - \frac{1}{3}$$

$$= \frac{1}{\sqrt{5}} - \frac{1}{3}$$

5. (a) By definition, when we take a function that transforms a $RV(X)$ into a $RV(Y)$, then the PDF of Y is given by:

$$p(y) = \left| \frac{d}{dy}\left(f^{-1}(y)\right) \right| p\left(f^{-1}(y)\right)$$

For $Y = X^3$ and $X \sim \mathcal{N}(X|0,1)$ we see:

$$p(y) = \left| \frac{d}{dy}(y^{\frac{1}{3}}) \right| p(y^{\frac{1}{3}})$$

$$= \left| \frac{1}{3}(y^{\frac{-2}{3}}) \right| \frac{1}{\sqrt{2\pi}} e^{\frac{-y^{2/3}}{2}}$$

(b) Assuming the matrix G is invertible, then from $Y = GX$, we get $X = G^{-1}Y$ and the PDF of Y becomes:

$$f_y(y) = f_x(G^{-1}y)/det(G)$$

6. (a) Law of Total Expectation:

$$\mathbb{E}(Y|X) = \sum_y y * P(y|x)$$

$$\mathbb{E}(\mathbb{E}(Y|X)) = \sum_x \sum_y (y * P(y|x)) * P(x)$$

$$= \sum_y \sum_x y * P(y|x) * P(x)$$

$$= \sum_y y \sum_x P(y|x) * P(x)$$

$$= \sum_y y \sum_x P(y,x)$$

$$= \sum_y y * P(y) = \mathbb{E}(Y)$$

(b) Law of Total Variance:

$$\mathbb{V}(Y) = \mathbb{E}(Y^2) - (\mathbb{E}(Y))^2$$
$$\mathbb{V}(Y|X) = \mathbb{E}[Y^2|X] - \mathbb{E}[Y|X]^2$$
$$\mathbb{E}(\mathbb{V}(Y|X)) = \mathbb{E}(\mathbb{E}(Y^2|X)) - \mathbb{E}(\mathbb{E}(Y|X)^2)$$
$$= \mathbb{E}(Y^2) - \mathbb{E}(\mathbb{E}(Y|X)^2)$$
$$\mathbb{V}(\mathbb{E}(Y|X)) = \mathbb{E}(\mathbb{E}(Y|X)^2) - \mathbb{E}(\mathbb{E}(Y|X))^2$$
$$= \mathbb{E}(\mathbb{E}(Y|X)^2) - E(Y)^2$$
$$\mathbb{E}(\mathbb{V}(Y|X)) + \mathbb{V}(\mathbb{E}(Y|X)) = \mathbb{E}(Y^2) - \mathbb{E}(\mathbb{E}(Y|X)^2) + \mathbb{E}(\mathbb{E}(Y|X)^2) - E(Y)^2$$
$$= \mathbb{E}(Y^2) - E(Y)^2$$

7. (a) We can take the gradient of the function as follows:

$$f = (1 + exp(-a^T x))^{-1}$$
$$\nabla f = -(1 + exp(-a^T x))^{-2} * exp(-a^T x) * -a^T$$
$$= \frac{exp(-a^T x) * a^T}{1 + exp(-a^T x))^2}$$

(b) We can then derive the next gradient as follows using the quotient rule for derivatives and to simplify we'll let $a = exp(-a^T x)$; $b = (1 + exp(-a^T x))$; $c = a^T$:

$$\left(\frac{f}{g}\right)' = \frac{f'g - g'f}{g^2}$$

$$\nabla^2 f = \frac{(a * b^2 * -c) - (a^2 * 2b * -c^2)}{b^4}$$

$$= \frac{(a * b * -c) + (a^2 * 2 * c^2)}{b^3}$$

(c) We know the log of the sigmoid function is concave, and from the second derivative we can deduce similar. So the -log would be convex.

8. (a) If $f(x) = -log(x)$, then we can define the Frenchel conjugate as follows:

$$g(\lambda) = \max_x \lambda x - f(x)$$

$$= \max_x \lambda x + log(x)$$

We can find the max by taking the derivative and setting it equal to zero, and then solve by plugging our answer back into $g(\lambda)$:

$$\lambda + \frac{1}{x} = 0$$

$$x = -\frac{1}{\lambda}$$

$$g(\lambda) = -1 + log(-\frac{1}{\lambda}), for \lambda < 0$$

(b) If $f(x) = x^T A^{-1} x$ where $A \succ 0$ then the Frenchel conjugate is defined as follows:

$$g(\lambda) = \max_x \lambda x - x^T A^{-1} x$$

where

$$\frac{\partial x^T A^{-1} x}{\partial x} = 2A^{-1} x$$

$$\lambda - 2A^{-1} x = 0$$

$$x = A\frac{\lambda}{2}$$

9. (a)

$$-\frac{1}{2} log|S\Sigma S^T| - \frac{1}{2} tr((a - A\mu)^T (S\Sigma S^T)^{-1}(a - A\mu))$$

$$-\frac{1}{2} dlog|S\Sigma S^T| = -\frac{1}{2} tr(S\Sigma S^T)^{-1} d(S\Sigma S^T)$$

$$= -\frac{1}{2} tr((S\Sigma S^T)^{-1} S d\sigma S^T)$$

$$= -\frac{1}{2} tr(S^T (S\Sigma S^T)^{-1} S d\Sigma)$$

$$d(-\frac{1}{2} tr((a - A\mu)^T (S\Sigma S^T)^{-1}(a - A\mu))$$

$$= -\frac{1}{2} tr((a - Amu)^T d(S\Sigma S^T)^{-1}(a - A\mu))$$

$$= -\frac{1}{2} tr((a - Amu)^T - (S^T) d(S\Sigma S^T)(S\Sigma S^T)^{-1}(a - A\mu))$$

$$= -\frac{1}{2} tr(S^T (S\Sigma S^T)^{-1}(a - A\mu))(a - Amu)^T - (S^T) S d\Sigma$$

(b)

10. If we first define a second function $h(z) = log(|X + zV|)$ s.t. $X + zV$ is a positive definite matrix, then the following holds true:

$$h(z) = log(|X + zV|)$$
$$= log(|X^{\frac{1}{2}}X^{\frac{1}{2}} + zX^{\frac{1}{2}}X^{\frac{-1}{2}}VX^{\frac{-1}{2}}X^{\frac{1}{2}}|)$$
$$= log(|X^{\frac{1}{2}}(I + zX^{\frac{-1}{2}}VX^{\frac{-1}{2}})X^{\frac{1}{2}})$$

By doing this, we aren't changing anything, but we are able to use the rule that the determinant of products equals the product of determinants:

$$h(z) = log(|X| * |I + tX^{\frac{-1}{2}}VX^{\frac{-1}{2}}|)$$
$$= log(|X|) + log(|I + tX^{\frac{-1}{2}}VX^{\frac{-1}{2}}|)$$

Almost there, from here we add that we know the following:

$$log(|I + tX^{\frac{-1}{2}}VX^{\frac{-1}{2}}|) = log(\prod_i)(1 + z\lambda_i) = \sum_i log(1 + z\lambda_i)$$

To wrap it all up we substitute that back in and take the second derivative of the negative of the function:

$$h(z) = log(|X|) + \sum_i log(1 + z\lambda_i)$$
$$-h''(z) = \sum_i \frac{\lambda_i^2}{(1 + z\lambda_i)^2} \geq 0$$

Since $-h(z)$ is convex, we know that $-log(|X|)$ is as well, meaning we can conclude that $log(|X|)$ is in fact concave. Derived with some help from some online resources.