# CS 542, Spring 2014

# Assignment 1

Shan Sikdar

## 1 2.1

Verify $\sum_{x=0}^{1} p(x|\mu) = 1$

Proof. $\sum_{x=0}^{1} p(x|\mu) = p(x=0|\mu) + p(x=1|\mu) = \mu + 1 - \mu = 1$

Verify: $\mathbb{E}[x] = \mu$

Proof. $\mathbb{E}[x] = \sum_{x} x \;\; p(x|\mu) = 1 * \mu + 0 * (1-\mu) = \mu$

Verify $var[x] = \mu(1-\mu)$

Proof.

$var(x) = \mathbb{E}[(x - E[x])^2]$

$= (1-\mu)(0-\mu)^2 + \mu(1-\mu)^2$

$= (1-\mu)(\mu)^2 + \mu(1-\mu)^2$

$= \mu - \mu^2 = \mu(1-\mu)$

Show entropy of H[x] of a Bernoulli distributed random variable x is given by : $H[x] = -\mu\, ln\mu - (1-\mu)ln(1-\mu)$

By definition of entropy: $H[x] = -\sum_{x} p(x|\mu)\; ln[p(x|\mu)]$

$= -[(1-\mu)ln(1-\mu) + (\mu\, ln\mu)] = -\mu\, ln\mu - (1-\mu)ln(1-\mu)$

## 2 2.2

Show that the distribution (2.261) is normalized and evaluate its mean, variance, and entropy.

Show normalized: $\sum x = -1, 1\ p(x|\mu) = 1$
proof.
$\sum_{x=-1,1} p(x|\mu) = \frac{(1-\mu)^1}{2}\frac{(1+\mu)^0}{2} + \frac{(1-\mu)^0}{2}\frac{(1+\mu)^1}{2} = \frac{(1-\mu)}{2} + \frac{(1+\mu)}{2} = \frac{2}{2} = 1$

Evaluate its mean:
$= \sum_x f(x)\ \ p(x) = (-1)(\frac{(1-\mu)}{2}) + (1)(\frac{(1+\mu)}{2}) = \frac{2\mu}{2} = \mu$
Evaluate its Variance:
$var[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = (-1)^2(\frac{1-\mu}{2}) + (1)^2(\frac{1+\mu}{2}) + \mu^2 = 1 - \mu^2$
Evaluate its Entropy:
$H[x] = -\sum_x p(x|\mu)\, ln[p(x|\mu)] = -[((\frac{1-\mu}{2})\, ln((\frac{1-\mu}{2})) + ((\frac{1+\mu}{2})\, ln((\frac{1+\mu}{2}))] =$
$-((\frac{1-\mu}{2})\, ln((\frac{1-\mu}{2}))\ -\ ((\frac{1+\mu}{2})\, ln((\frac{1+\mu}{2}))$

## 3 2.5

$\Gamma(a)\Gamma(b) = \int_0^\infty e^{-x}x^{a-1}dx \int_0^\infty e^{-y}y^{b-1}dy$
$= \int_0^\infty \int_0^\infty e^{-(x+y)}x^{a-1}y^{b-1}dydx$
Change of variable t = y + x, so y = t - x and dy = dt. The limits of integration change as well:
$= \int_0^\infty \int_x^\infty e^{-(t)}x^{a-1}(t-x)^{b-1}dtdx$
Interchange the order of integration, the limits will have to change so the same area is being integrated:
$= \int_0^\infty \int_0^t e^{-(t)}x^{a-1}(t-x)^{b-1}dxdt$
Make change of variables $x = \mu t$ so $dx = td\mu$ The limits of integration change again as well:
$= \int_0^\infty \int_0^1 e^{-(t)}(t\mu)^{a-1}(t-t\mu)^{b-1}t\ d\mu dt$

$$= \int_0^\infty e^{-(t)} t^{a-1} t^{b-1} t dt \int_0^1 (\mu)^{a-1} (1-\mu)^{b-1} d\mu$$
$$= \int_0^\infty e^{-(t)} t^{(a+b)-1} dt \int_0^1 (\mu)^{a-1} (1-\mu)^{b-1} d\mu$$
$$= \Gamma(a+b) \int_0^1 (\mu)^{a-1} (1-\mu)^{b-1} d\mu$$
So:
$$\Gamma(a)\Gamma(b) = \Gamma(a+b) \int_0^1 (\mu)^{a-1} (1-\mu)^{b-1} d\mu$$
$$\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} = \int_0^1 (\mu)^{a-1} (1-\mu)^{b-1} d\mu$$

# 4 2.7

Show that the posterior mean value of x lies between the prior mean and the maximum likelihood estimate for $\mu$. To do this show the posterior mean can be written as $\lambda$ times the prior mean plus $(1-\lambda)$ times the maximum likelihood estimate where $0 \leq \lambda \leq 1$.

From 2.15 the mean of the prior is $\frac{a}{a+b}$
From 2.20 the mean of the posterior is $\frac{m+a}{m+a+l+b}$
From class the maximum likelihood estimate is $\frac{m}{m+l}$

So want to show $\lambda(\frac{a}{a+b}) + (1-\lambda)(\frac{m}{m+l}) = \frac{a+m}{a+b+l+m}$ where $0 \leq \lambda \leq 1$.
So solve for $\lambda$ and argue its range is $0 \leq \lambda \leq 1$:

$\lambda(\frac{a}{a+b}) + (1-\lambda)(\frac{m}{m+l}) = \frac{a+m}{a+b+l+m}$
$(\frac{a\lambda}{a+b}) + (\frac{m-m\lambda}{m+l}) = \frac{a+m}{a+b+l+m}$
$\frac{(m+l)(a\lambda)+(m-m\lambda)(a+b)}{(a+b)(m+l)} = \frac{a+m}{a+b+l+m}$
$\frac{a\lambda l+am+bm-b\lambda m}{(a+b)(m+l)} = \frac{a+m}{a+b+l+m}$
$\frac{a\lambda l-b\lambda m}{(a+b)(m+l)} + \frac{am+bm}{(a+b)(m+l)} = \frac{a+m}{a+b+l+m}$
$\frac{a\lambda l-b\lambda m}{(a+b)(m+l)} + \frac{m}{(m+l)} = \frac{a+m}{a+b+l+m}$
$\frac{a\lambda l-b\lambda m}{(a+b)(m+l)} = \frac{a+m}{a+b+l+m} - \frac{m}{(m+l)}$
$\frac{a\lambda l-b\lambda m}{(a+b)(m+l)} = \frac{a+m}{a+b+l+m} - \frac{m}{(m+l)}$

$\frac{a\lambda l - b\lambda m}{(a+b)(m+l)} = \frac{(a+m)(m+l)}{(a+b+l+m)(m+l)} - \frac{m(a+b+l+m)}{(a+b+l+m)(m+l)}$
$\frac{a\lambda l - b\lambda m}{(a+b)(m+l)} = \frac{(al-bm)}{(a+b+l+m)(m+l)}$
$\lambda = \frac{(al-bm)(a+b)(m+l)}{(al-bm)(a+b+l+m)(m+l)}$
$\lambda = \frac{(a+b)}{(a+b+l+m)}$
$\lambda = \frac{1}{1+\frac{l+m}{a+b}}$

All variables in this equation are greater than 0. If $\frac{l+m}{a+b}$ gets larger, $\lambda$ will tend to 0. If $\frac{l+m}{a+b}$ gets smaller, lambda will tend to 0. So $0 \leq \lambda \leq 1$

# 5 2.8

Prove $\mathbb{E}[x] = \mathbb{E}_y[\mathbb{E}_x[x|y]]$ and $var[x] = \mathbb{E}_y[var_x[x|y]] + var_y[\mathbb{E}_x[x|y]]$

Prove $\mathbb{E}[x] = \mathbb{E}_y[\mathbb{E}_x[x|y]]$.

$E[x] = \int_x p(x)\ x\ dx$
$= \int_x \int_y p(x|y)\ \ p(y)\ \ x\ dy\ dx$ (using product rule)
$= \int_y \int_x p(x|y)\ \ p(y)\ \ x\ dx\ dy$
$= \int_y p(y) \int_x p(x|y)\ x\ dx\ dy$
$= \int_y p(y)\ E_x[x|y]\ dy$
$= E_y[E_x[x|y]]$

Prove: $var[x] = \mathbb{E}_y[var_x[x|y]] + var_y[\mathbb{E}_x[x|y]]$
$var[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2$
$= \mathbb{E}_y[\mathbb{E}_x[x^2|y]] - \mathbb{E}_y[\mathbb{E}_x[x|y]]^2$
$= \mathbb{E}_y[\mathbb{E}_x[x^2|y]] - \mathbb{E}_y[\mathbb{E}_x[x|y]^2] + \mathbb{E}_y[\mathbb{E}_x[x|y]^2] - \mathbb{E}_y[\mathbb{E}_x[x|y]]^2$
$= \mathbb{E}_y[\mathbb{E}_x[x^2|y]] - \mathbb{E}_y[\mathbb{E}_x[x|y]^2] + var_y[\mathbb{E}_x[x|y]]$
Since Expectation is a linear operator:
$= \mathbb{E}_y[\mathbb{E}_x[x^2|y] - \mathbb{E}_x[x|y]^2] + var_y[\mathbb{E}_x[x|y]]$
$= \mathbb{E}_y[var_x[x|y]] + var_y[\mathbb{E}_x[x|y]]$

# 6 2.27

Let x and z be two independent random vectors, so that p(x,z) = p(x)p(z).

Show that the mean of their sum y = x + z is given by the sum of the means of each of the variables seperately.
Proof: Let $x = [x_1, ....., x_n]$ and $z = [z_1, ....., z_n]$ then:

$$\frac{x_1+x_2+....+x_n}{n} + \frac{z_1+z_2+....+z_n}{n} = \frac{(x_1+z_1)+(x_2+z_2)+....+(x_n+z_n)}{n} = \frac{y_1+y_2+...+y_n}{n}$$

Show that the covariance matrix of y is given by the sum of the covariance matrices x and z.
Proof: Using formula 2.63:
$cov[y] = \mathbb{E}[(y - E[y])(y - E[y])^T]$
$= \mathbb{E}[((x + z) - E[(x + z)])((x + z)) - E[(x + z)])^T]$
$= \mathbb{E}[(x - E[x] + z - E[x])(x - E[x] + z - E[x])^T] = \mathbb{E}[(\ (x - E[x]) + (z - E[x])\ )(\ (x - E[x])^T + (z - E[x])^T)] = \mathbb{E}[(x - E[x])(x - E[x])^T + (x - E[x])(z - E[x])^T + (z - E[x])(x - E[x])^T + (z - E[x])(z - E[x])^T]$
$= \mathbb{E}[(x-E[x])(x-E[x])^T]+\mathbb{E}[(x-E[x])(z-E[x]^T)]+\mathbb{E}[(z-E[x])(x-E[x])^T] + \mathbb{E}[(z - E[x])(z - E[x])^T)]$
$= cov[x] + cov[z] + cov[x, z] + cov[z, x]$
$= cov[x] + cov[y]$

Facts I used:
(1) x,z independent, E is a linear operator
(2) $(A + B)^T = A^T + B^T$
(3) Since x and z are independent, the covariance matrices with respect to each other will be 0.

Confirm that this result agrees with that of excercise 1.10.
You can consider one variable to be a vector with one element so the first proof shows that 1.28 holds. In the single variable case, covariance reduces to just normal varince so 1.29 holds.

# 7 2.28

Consider a joint distribution over the variable $z = \binom{x}{y}$ whoose mean and covairiance are given by (2.108) and (2.105) respectively.

I. By making use of the results (2.92) and (2.93) show that the marginal distribution p(x) is given by $p(x) = N(x|\mu, \Lambda^{-1})$
So find the mean and covariance of the marginal distribution:
By 2.92 we know that marginal ditribution p(x) has the mean given by $E[x] = \mu_x$ but we also know that the mean vector for z is givien by $E[z] = \binom{\mu}{A\mu+b}$. Therefore $\mu_x = \mu$ and so the mean of the marginal distribution p(x) is $\mu$. From 2.93, we know that the covariance of the marginal distribution p(x) is given by $cov[x] = \Sigma_{xx}$. But from 2.105 we know that covariance matrix for z is $cov[z] = \begin{pmatrix} \Lambda^{-1} & \Lambda^{-1}A^T \\ A\Lambda^{-1} & L-1+A\Lambda^{-1}A^T \end{pmatrix}$
So: $\Lambda^{-1} = \Sigma_{xx} = cov[x]$ then $\Lambda^{-1}$ is the covariance of the marginal distribution. Therefore $p(x) = N(x|\mu, \Lambda^{-1})$

II. Similarly, by making useof the results (2.81) and (2.82) show that the conditional distribution $p(y|x) = N(y|Ax + b, L^{-1})$
So find the mean and covariance of the conditonal distribution:
By 2.81:
$\mu_{y|x} = \mu_y + \Sigma_{xy}\Sigma_{xx}^{-1}(x - \mu_x)$
$= A\mu + b + (A\Lambda^{-1})(\Lambda^{-1})^{-1}(x - \mu)$ (using 2.108 and 2.105)
$= A\mu + b + (A\Lambda^{-1})(\Lambda)(x - \mu)$
$= A\mu + b + (A)(x - \mu)$
$= A\mu + b + (Ax - A\mu)$
$= Ax + b$
And so the mean for the conditonal distribution is $Ax + b$

By 2.82:
$\Sigma_{y|x} = \Sigma_{yy} - \Sigma_{yx}\Sigma_{yx}^{-1}\Sigma_{xy}$
$= L^{-1} + A\Lambda^{-1}A^T - (A\Lambda^{-1})(\Lambda^{-1})^{-1}(\Lambda^{-1}A^T)$ (using 2.105)
$= L^{-1} + A\Lambda^{-1}A^T - (A\Lambda^{-1})(\Lambda)(\Lambda^{-1}A^T)$
$= L^{-1} + A\Lambda^{-1}A^T - (A\Lambda^{-1}A^T)$
$= L^{-1}$
And so the variance for the conditonal distribution is $L^{-1}$ Therefore the conditonal distribution is $N(y|Ax + b, L^{-1})$

# 8 2.31

Find an expression for the marginal distribution p(y) by considering the linear-Gaussian model compromising the product of the marginal distribution p(x) and the conditonal distribution $p(y|x)$
Since: p(x) is gaussian, $p(x) = N(x|\mu_x, \Sigma_x)$
From page 91 in the book that $p(x) = N(x|\mu, \Lambda^{-1})$
So $\Sigma_x = \Lambda^{-1}$ and $\mu_x = \mu$.
Since y is made up of x and z, $p(y|x)$ needs to be written in terms of x and z so $p(y|x) = N(y|\mu_z + x, \Sigma_z)$
Also from page 91 $p(y|x) = N(y|Ax + b, L^{-1})$
So: $Ax + b = x + \mu_z$ and $L^{-1} = \Sigma_z$ then A = I and $b = \mu_z$
Now using formula 2.115 and plugging in found values:
$p(y) = N(x|A\mu + b, L^{-1} + A\Lambda^{-1}A^T)$
$p(y) = N(x|I\mu_x + \mu_z, \Sigma_z + I\Sigma_x I^T)$
$p(y) = N(x|\mu_x + \mu_z, \Sigma_z + \Sigma_x)$

# 9 2.34

log of the likelihood function:

$$lnp(X|\mu,\Sigma) = -\frac{ND}{2}ln(2\pi) - \frac{N}{2}ln|\Sigma| - \frac{1}{2}\sum_{n=1}^{N}(x_n-\mu)^T\Sigma^{-1}(x_n-\mu)$$

To find the where the max will be, take the derivative and set it equal to 0. When the derivative is take with respect to $\Sigma$ the first term will become 0. Using (c.21) the second term will become $-\frac{N}{2}(\Sigma^{-1})^T = -\frac{N}{2}(\Sigma^{-1})$
For the third term, since there are derivative rules in the appendix, it easier to find an alternate form if the term in terms of the trace. So as long as the result ends up the same (i.e the same entries in the vectors and matrices get multiplied and added correctly) the term can be rewritten as trace. So rewrite $\sum_{n=1}^{N}(x_n-\mu)^T\Sigma^{-1}(x_n-\mu)$ with $Tr[\Sigma^{-1}\sum_{n=1}^{N}(x_n-\mu)^T(x_n-\mu)]$
So:

$-\frac{1}{2}\sum_{n=1}^{N}(x_n-\mu)^T\Sigma^{-1}(x_n-\mu)$
$= -\frac{1}{2}Tr[\Sigma^{-1}\sum_{n=1}^{N}(x_n-\mu)^T(x_n-\mu)]$
$= -\frac{1}{2}\frac{N}{N}Tr[\Sigma^{-1}\sum_{n=1}^{N}(x_n-\mu)^T(x_n-\mu)]$
$= -\frac{1}{2}\frac{N}{N}Tr[\Sigma^{-1}\sum_{n=1}^{N}(x_n-\mu)^T(x_n-\mu)]$
$= -\frac{1}{2}NTr[\Sigma^{-1}\frac{1}{N}\sum_{n=1}^{N}(x_n-\mu)^T(x_n-\mu)]$
$= -\frac{1}{2}NTr[\Sigma^{-1}Q]$ (hide terms with Q to make the math simpler to read)
So when the derivative is taken this term:
$\frac{d}{d\Sigma_{i,j}} - \frac{1}{2}NTr[\Sigma^{-1}Q]$
$= -\frac{1}{2}NTr[\frac{d}{d\Sigma_{ij}}\Sigma^{-1}Q]$ (2.26) used here
$= -\frac{1}{2}NTr[\frac{d}{d\Sigma_{ij}}\Sigma^{-1}Q\Sigma^{-1}]$
$= \frac{1}{2}NTr[\Sigma^{-1}Q\Sigma^{-1}]$
$= \frac{1}{2}N(\Sigma^{-1}Q\Sigma^{-1})_{ij}$

So : $\frac{1}{2}N(\Sigma^{-1}Q\Sigma^{-1}] - \frac{N}{2}(\Sigma^{-1})_{ij} = 0$
$\frac{1}{2}N(\Sigma^{-1}Q\Sigma^{-1})_{ij} = \frac{N}{2}(\Sigma^{-1}) = 0$
$\frac{1}{2}N(\Sigma^{-1}Q\Sigma^{-1})_{ij} = \frac{N}{2}(\Sigma^{-1}) = 0$

$\Sigma = \frac{1}{N} \sum_{n=1}^{N} (x_n - \mu)^T (x_n - \mu)$ which is the same equation as 1.22.

## 10 2.40

Given a prior distribution $p(\mu) = N(\mu|\mu_0, \Sigma_0)$ find the corresponding posterior distribution $p(\mu|x)$.

We know that posterior is proportional to prior and likelihood. The prior is given as $p(\mu) = N(\mu|\mu_0, \Sigma_0)$. And the likelihoodfor N observations is $\prod^{N} p(x_n|\mu, \Sigma)$.

So $p(\mu|x) = p(\mu)\prod^{N} p(x_n|\mu, \Sigma) = N(\mu|\mu_0, \Sigma_0)\prod^{N} N(x_n|\mu, \Sigma)$ Using 2.71 this can be rewritten as:

$= -\frac{1}{2}(\mu - \mu_0)^T\Sigma_0^{-1}(\mu - \mu_0) + -\frac{1}{2}\sum_{n=1}^{N}(x_n - \mu)^T\Sigma^{-1}(x_n - \mu)$ Using the second equality in 2.71:

$= -\frac{1}{2}\mu^T\Sigma_0^{-1}\mu + \mu^T\Sigma_0^{-1}\mu_0 - \frac{1}{2}\sum^{N} x_n^T\Sigma_x^{-1} + \frac{1}{2}\sum^{N} x_n^T\Sigma^{-1}\mu + const$

$= -\frac{1}{2}\mu^T\Sigma_0^{-1}\mu - \frac{1}{2}\sum^{N} x_n^T\Sigma^{-1}\mu + \mu^T\Sigma_0^{-1}\mu_0 - \frac{1}{2}\sum^{N} x_n^T\Sigma_x^{-1} + const$

$= -\frac{1}{2}\mu^T(\Sigma_0^{-1} + N\Sigma^{-1})\mu + \mu^T\Sigma_0^{-1}\mu_0 - \frac{1}{2}\sum^{N} x_n^T\Sigma_x^{-1} + const$

$= -\frac{1}{2}\mu^T(\Sigma_0^{-1} + N\Sigma^{-1})\mu + \mu^T(\Sigma_0^{-1}\mu_0 + \Sigma^{-1}\sum^{N} x_n) + const$

So this is the distribution for the posterior