

Problem Sheet 1

CMPE 58K, Bayesian Statistics and Machine Learning

Instructor: A. T. Cemgil

Due: 15 Oct 2008, 10:00.

Exercises are labelled with greek characters α, β, γ . Each label denotes the type of the question and roughly corresponds to its difficulty with α the hardest. I **don't** expect you to solve all the questions but you should solve **at least one** question of each type before handing in your work. A π denotes questions that have a programming component. Don't send any executables, just the source code and an example run output is sufficient. However, write your programs clearly as some of these will be used as subroutines in later exercises.

While some of the exercises are originals, many of the exercises are verbatim copies or slight modified versions of exercises from MacKay or Bishop. Often, the solutions can be found on the web or from "other" sources. It is OK to lookup the public domain solutions but resisting the temptation and attempting them first yourself would be a lot more useful for your understanding.

A1.1 (π) (**log sum exp**) Implement a function in matlab with the following specification:

```
LOG_SUM_EXP Numerically stable computation of log(sum(exp(X), dim))
```

```
[r] = log_sum_exp(X, dim)
```

```
Inputs :
```

```
  X : Array
```

```
    dim : Sum Dimension <default = 1>
```

```
          Row vector sums should be calculated
```

```
          by transposing or specifying dim=2
```

```
Outputs:
```

```
  r : log(sum(exp(X), dim))
```

```
Usage Example : [s] = log_sum_exp([-10 -9]');
```

```
log(sum(exp([-1213 -1214])))
```

```
Warning: Log of zero.
```

```
log_sum_exp([-1213 -1214], 2)
```

```
ans = -1.2127e+003
```

A1.2 (π) (**randgen**) Implement a function in matlab that generates independent random samples from a specified distribution:

```
RANDGEN Random samples with replacement from a specified distribution
```

```
Y = RANDGEN(S, Siz, P) returns a weighted sample, using positive weights P. P is often a vector of probabilities but can be unnormalised.
```

```
If P is absent we assume a uniform distribution
```

Example: Generate a random sequence of the characters ACGT, with replacement, according to specified probabilities.

```
R = randgen('ACGT',48, [0.15 0.35 0.35 0.15])
```

Example: Generate a random 3 by 3 matrix with independent entries from $S = [1 \ 2 \ 5]$ according to specified weights.

```
R = randgen([1 2 5], [3 3], [2 2 1])
```

So on average there should be about twice as many one's as five's.

[Hint: Don't use Matlab statistics toolbox function `randsample` as a subroutine. However, you are welcome to read the source code and use the ideas in your implementation.]

A1.3 (γ) (**Bayes Theorem**) Suppose that we have three coloured boxes r (red), b (blue), and g (green). Box r contains 3 apples, 4 oranges, and 3 limes, box b contains 1 apple, 1 orange, and 0 limes, and box g contains 3 apples, 3 oranges, and 4 limes.

- If a box is chosen at random with probabilities $p(r) = 0.2$, $p(b) = 0.2$, $p(g) = 0.6$, and a piece of fruit is removed from the box (with equal probability of selecting any of the items in the box), then what is the probability of selecting an apple?
- If we observe that the selected fruit is in fact an orange, what is the probability that it came from the green box?
- Choose the appropriate random variables and draw a directed graphical model for this problem.

A1.4 (α) (**Game show**) On a game show, a contestant is told the rules as follows: There are three doors, labelled 1, 2, 3. A single prize has been hidden behind one of them. You get to select one door. Initially your chosen door will not be opened. Instead, the gameshow host will open one of the other two doors, and he will do so in such a way as not to reveal the prize. For example, if you first choose door 1, he will then open one of doors 2 and 3, and it is guaranteed that he will choose which one to open so that the prize will not be revealed. At this point, you will be given a fresh choice of door: you can either stick with your first choice, or you can switch to the other closed door. All the doors will then be opened and you will receive whatever is behind your final choice of door.

- Imagine that the contestant chooses door 1 first; then the gameshow host opens door 3, revealing nothing behind the door, as promised. Should the contestant (a) stick with door 1, or (b) switch to door 2, or (c) does it make no difference?
- Imagine that the game happens again and just as the gameshow host is about to open one of the doors a violent earthquake rattles the building and one of the three doors flies open. It happens to be door 3, and it happens not to have the prize behind it. The contestant had initially chosen door 1. Repositioning his toupée, the host suggests, 'OK, since you chose door 1 initially, door 3 is a valid door for me to open, according to the rules of the game; I'll let door 3 stay open. Let's carry on as if nothing happened.' Should the contestant stick with door 1, or switch to door 2, or does it make no difference? Assume

that the prize was placed randomly, that the gameshow host does not know where it is, and that the door flew open because its latch was broken by the earthquake.

- (c) A similar alternative scenario is a gameshow whose confused host forgets the rules, and where the prize is, and opens one of the unchosen doors at random. He opens door 3, and the prize is not revealed. Should the contestant choose what's behind door 1 or door 2? Does the optimal decision for the contestant depend on the contestant's beliefs about whether the gameshow host is confused or not?
- (d) Formally derive the results defining the appropriate random variables and using the Bayes rule.

A1.5 (γ) (**Twenty-faced dice**) A die is selected at random from two twenty-faced dice on which the symbols 1–10 are written with nonuniform frequency as follows.

Symbol	1	2	3	4	5	6	7	8	9	10
Number of faces of die A	6	4	3	2	1	1	1	1	1	0
Number of faces of die B	3	3	2	2	2	2	2	2	1	1

- (a) The randomly chosen die is rolled 7 times, with the following outcomes:

5, 3, 9, 3, 8, 4, 7.

What is the probability that the die is die A?

- (b) Assume that there is a third twenty-faced die, die C, on which the symbols 1–20 are written once each. As above, one of the three dice is selected at random and rolled 7 times, giving the outcomes: 3, 5, 4, 8, 3, 9, 7.

What is the probability that the die is die A, die B or die C?

- (c) Choose the appropriate random variables and draw directed graphical models for both problems.

A1.6 (α) (**Coin**) Suppose a biased coin with $p(\text{head}) = \pi$ is thrown N times. The number of times head shows up is 4. Assume all π and N are a-priori equally likely.

- (a) What is the most likely value of N as a function of π ?

A1.7 (α) (**Sums of Random Variables**)

- (a) Two ordinary dice with faces labelled 1...6 are thrown. What is the probability distribution of the sum of the values? What is the probability distribution of the absolute difference between the values?
- (b) One hundred ordinary dice are thrown. What, roughly, is the probability distribution of the sum of the values? Sketch the probability distribution and estimate its mean and standard deviation.

[Hint: This exercise is intended to help you think about the central-limit theorem, which says that if independent random variables x_1, \dots, x_N have means μ_n and finite variances σ_n^2 , then, in the limit of large N , the sum $\sum_n x_n$ has a distribution that tends to a normal (Gaussian) distribution with mean $\sum_n \mu_n$ and variance $\sum_n \sigma_n^2$.]

A1.8 (β) (**Jacobians**) Consider a probability density $p_x(x)$ of a continuous random variable x . Suppose we make a nonlinear change of variable using $x = g(y)$, so that the density transforms according to

$$\begin{aligned}
 p_y(y) &= \left| \frac{dx}{dy} \right| p_x(x) \\
 &= |g'(y)| p_x(g(y))
 \end{aligned}
 \tag{1}$$

(a) By differentiating Eq.1, show that the location y^* of the maximum of the density (in y) is **not** in general related to the location x^* of the maximum of the density over x by the simple functional relation $x^* = g(y^*)$.

Note: This as a consequence of the Jacobian factor. This shows that the maximum of a probability density (in contrast to a simple function) is dependent on the choice of variable.

(b) Verify that, in the case of a linear transformation, the location of the maximum transforms in the same way as the variable itself.

A1.9 (γ) (**Covariance**) We are given two random variables x and y

(a) Show that if x and y are independent, then their covariance is zero.

(b) Give an example joint density $p(x, y)$ where the covariance is zero but the variables are not independent (i.e. observing one gives information about the other).

A1.10 (β) (**Counting States**) Suppose x_i for $i = 1 \dots 4$ are discrete random variables, each with 10 states.

(a) For each of the below graphical models, specify the implied factorisation of the joint distribution $p(x_1, x_2, x_3, x_4)$ and calculate the number of free parameters one should specify

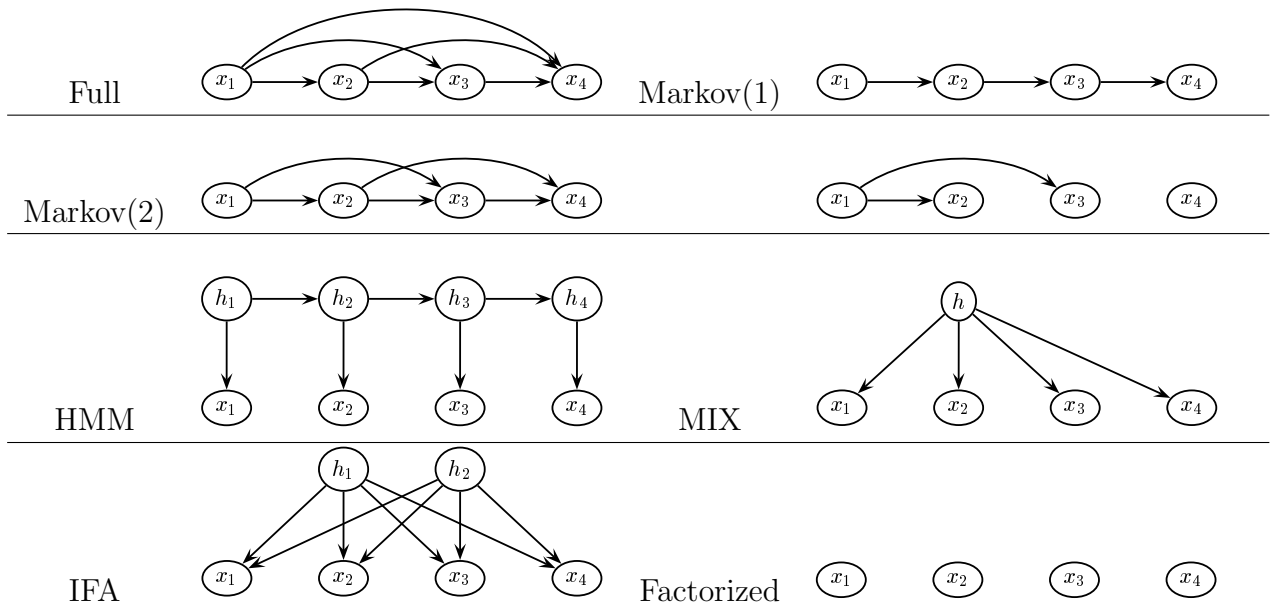
Model	Structure	factorization
Full		
Markov(2)		
Markov(1)		
Factorized		

[Hint: Be picky and calculate a minimal parametrisation. For example, if x_1 would be independent from the rest, $p(x_1)$ has only 9 free parameters.]

- (b) For each model, draw an associated factor graph and an equivalent undirected graphical model

A1.11 (β) (**Models**)

- For the following Graphical models, write down the factors of the joint distribution and plot an equivalent factor graph and an undirected graph.



A1.12 (β) (**Counting DAG's**) This is a tedious exercise but should give an idea about the search space when learning the model structure from data. You need a large piece of paper. Let us call the set of all directed acyclic graphs with N nodes $\text{DAG}(N)$.

- (a) How many directed acyclic graphs are there with 3 nodes ?
- (b) Draw each graph in $\text{DAG}(3)$ and write down the corresponding factorisation of a probability distribution for x_1, x_2 and x_3 .
- (c) Assume each random variable x_i has the same number of states. Find the partial ordering, where the binary relation for ordering two graphs \mathcal{G}_1 and \mathcal{G}_2 in $\text{DAG}(3)$ is defined if the factorisation corresponding to \mathcal{G}_1 is a special case of the one corresponding to \mathcal{G}_2 . For example, $p(x_1)p(x_2)$ is a special case of $p(x_1|x_2)p(x_2)$, whereas $p(x_1|x_2)p(x_2)p(x_3)$ and $p(x_1|x_3)p(x_2)p(x_3)$ are not comparable.
- (d) Draw the Hasse diagram. (See partially ordered set entry in wikipedia.)

A1.13 (γ) (**Chest Clinic**) A distribution factorises according to the following factorisation

$$p(A, B, D, F, T, L, M, X) = p(F|T, L)p(M)p(T|A)p(B|M)p(X|F)p(L|M)p(D|F, B)p(A)$$

- (a) Draw the corresponding directed graphical model
- (b) Draw an equivalent factor graph and undirected graphical model

A1.14 (γ) (**Hierarchical Hidden Markov Model**) A process is given by the following specification

$$\begin{aligned}x_0 &\sim p(x_0) \\z_0 &\sim p(z_0) \\x_k &\sim p(x_k|x_{k-1}) \\y_k &\sim p(y_k|x_k) \\z_k &\sim p(z_k|z_{k-1}, y_k)\end{aligned}$$

- (a) Draw the corresponding directed graphical model
- (b) Draw an equivalent factor graph and undirected graphical model

A1.15 (β) (**The Gamma Function**). The Gamma function is defined by the Integral:

$$\Gamma(x) \equiv \int_0^{\infty} u^{x-1} e^{-u} du$$

- (a) Show that $\Gamma(1) = 1$
- (b) Using integration by parts, show that

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

[Hint: Informally, integration by part follows from the chain rule as

$$\begin{aligned}(uv)' &= u'v + uv' \\ \int (uv)' &= \int u'v + \int uv' \\ \int u'v &= uv - \int uv'\end{aligned}$$

]

A1.16 (π) (**log(gamma) versus gammaln**)

- (a) In numeric computations, we almost always work with the logarithm of the gamma function $\log(\Gamma(x))$, which is computed without explicit reference to $\Gamma(x)$ to avoid overflow. In matlab, this function is `gammaln`. Using the `gammaln` function, write functions to evaluate the logarithms of \mathcal{G} , \mathcal{IG} and \mathcal{B} densities.

A1.17 (γ) (**Exponential Distribution**) The *exponential* distribution is defined as

$$\mathcal{E}(v; \lambda) = \frac{1}{\lambda} \exp\left(-\frac{v}{\lambda}\right)$$

- (a) Verify that the exponential distribution is a special case of the Gamma distribution. Find the shape and scale parameters of the corresponding gamma distribution.

A1.18 (β) (**Gamma and Inverse Gamma**) Let

$$\begin{aligned} z &\sim \mathcal{G}(v; a, 1) \\ v &= bz \\ \lambda &= 1/v \end{aligned}$$

where $a, b > 0$ are known positive constants.

- (a) Using the transformation formula Eq.1, derive the marginal distributions $p(v)$ and $p(\lambda)$ and if possible express the result as known distributions.

A1.19 (α, π) (**Generalised gamma**) The *Generalised gamma* distribution is a three parameter family defined as (Stacey and Mihram 1965, Johnson and Kotz pp.393)

$$\mathcal{GG}(v; \alpha, \beta, c) = \frac{|c|}{\Gamma(\alpha)\beta^{c\alpha}} v^{c\alpha-1} \exp^{-(v/\beta)^c}$$

Here, α is the shape, β is the scale and c is the power parameter.

- (a) Is the Generalised Gamma distribution an exponential family? If so, give the canonical parameters and the sufficient statistics.
- (b) Verify that the inverse Gamma distribution $\mathcal{IG}(v; a_i, b_i)$ and Gamma distribution $\mathcal{G}(v; a_g, b_g)$ are special cases. Give the corresponding settings of the power parameter.
- (c) Show that if

$$\begin{aligned} v &\sim \mathcal{GG}(v; \alpha, \beta, c) \\ z &= (v/\beta)^c \end{aligned}$$

then, z has the standard $\mathcal{G}(z; \alpha, 1)$ distribution. Using this fact, and a function that samples from standard gamma, implement a function generates random samples from a generalised Gamma distribution. The matlab statistics toolbox function `gamrnd(a, 1)` samples from the standard Gamma distribution.

A1.20 (γ) (**Expectations**) You are probably familiar with the idea of computing the expectation of a function of x ,

$$\langle f(x) \rangle = \sum_x P(x) f(x).$$

Maybe you are not so comfortable with computing this expectation in cases where the function $f(x)$ depends on the probability $P(x)$. The next few examples address this concern.

- (a) Let $p_a = 0.1$, $p_b = 0.2$, and $p_c = 0.7$. Let $f(a) = 10$, $f(b) = 5$, and $f(c) = 10/7$. What is $\langle f(x) \rangle$? What is $\langle 1/P(x) \rangle$?
- (b) For an arbitrary ensemble, what is $\langle 1/P(x) \rangle$?
- (c) Let $p_a = 0.1$, $p_b = 0.2$, and $p_c = 0.7$. Let $g(a) = 0$, $g(b) = 1$, and $g(c) = 0$. What is $\langle g(x) \rangle$?

- (d) Let $p_a = 0.1$, $p_b = 0.2$, and $p_c = 0.7$. What is the probability that $P(x) \in [0.15, 0.5]$?
What is

$$P\left(\left|\log \frac{P(x)}{0.2}\right| > 0.05\right)?$$

A1.21 (β, π) (**Entropies and Expectations**) Given a probability table $p(x, y)$ specified as a matrix and respective domains of two discrete random variables $x \in X$ and $y \in Y$, write programs to calculate

- (a) Expectations $\langle x \rangle$, $\langle y \rangle$, $\langle y|x \rangle$, $\langle x|y \rangle$, $\text{Cov}[x, y]$
(b) Joint Entropy

$$H[x, y] = -\langle \log p(x, y) \rangle_{p(x, y)}$$

- (c) Marginal Entropies

$$\begin{aligned} H[x] &= -\langle \log p(x) \rangle_{p(x)} \\ H[y] &= -\langle \log p(y) \rangle_{p(y)} \end{aligned}$$

- (d) Conditional Entropies

$$\begin{aligned} H[y|x] &= -\langle \log p(y|x) \rangle_{p(x, y)} \\ H[x|y] &= -\langle \log p(x|y) \rangle_{p(x, y)} \end{aligned}$$

- (e) Mutual Information

$$I(x, y) = H[x] - H[x|y] = KL(p(x, y) || p(x)p(y))$$

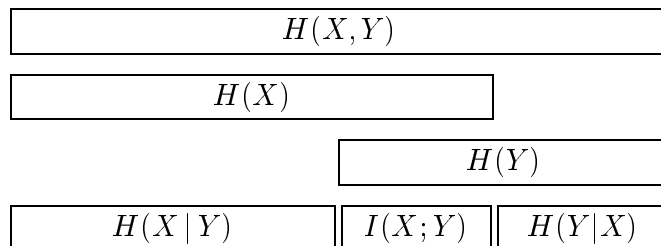
[Hint: Your program should correctly handle the limit case $0 \log 0 = 0$.]

- (f) Test your program for the following joint probability table

$p(x, y)$	$y = -1$	$y = 0$	$y = 5$
$x = 1$	0.3	0.3	0
$x = 2$	0.1	0.2	0.1

[Hint: Here, $X = \{1, 2\}$ and $Y = \{-1, 0, 5\}$.]

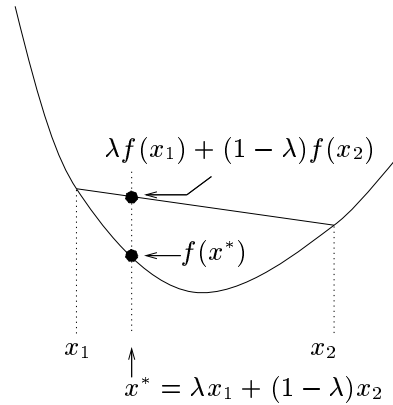
- (g) Verify the following picture



A1.22 (α) (**Jensen**) A function $f(x)$ is convex if

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

for $\lambda \in [0, 1]$. A function $f(x)$ is concave when $-f(x)$ is convex.



- (a) Specify if the following functions are convex, concave, both or none on the positive real numbers:

$$x^2, x^3, \log x, x \log x, e^{-x}, \log(\Gamma(x))$$

- (b) The celebrated Jensen's inequality states that for a convex function $f(x)$

$$\langle f(x) \rangle \geq f(\langle x \rangle)$$

By applying Jensen's inequality with $f(x) = \ln(x)$ show that the arithmetic mean of a set of real numbers is never less than their geometric mean.

[Hint: In Jensen's, the direction of inequality is reversed for a concave function. For x_1, x_2, x_3 , the arithmetic mean is $(x_1 + x_2 + x_3)/3$ and the geometric mean is $(x_1 x_2 x_3)^{1/3}$.]

- A1.23 (α) (**Bounds on Entropy**) Prove the assertion that $H(X) \leq \log(|\mathcal{A}_X|)$ with equality iff $p_i = 1/|\mathcal{A}_X|$ for all i . ($|\mathcal{A}_X|$ denotes the number of elements in the set \mathcal{A}_X .)

[Hint: Jensen involves both a random variable and a function, and you have quite a lot of freedom in choosing these; think about whether your chosen function f should be convex or concave.]

- A1.24 (γ) (**Differential Entropy**) Given a continuous real valued random variable x with density $p(x)$, the *differential entropy* is defined by

$$h(X) = - \int p(x) \log p(x)$$

Calculate the differential entropy of a

- (a) Gaussian $\mathcal{N}(x; \mu, \Sigma)$
- (b) Gamma $\mathcal{G}(x; a, b)$
- (c) Beta $\mathcal{B}(x; \alpha, \beta)$
- (d) Give an example where $h(X) < 0$.

- A1.25 (α) (**Gibbs' inequality**) Prove that the relative entropy

$$D_{\text{KL}}(P||Q) = \sum_x P(x) \log P(x)/Q(x)$$

satisfies $D_{\text{KL}}(P||Q) \geq 0$ with equality only if $P = Q$.

A1.26 (β) (**KL Divergence**) The KL (Kullback-Leibler) divergence is defined as

$$KL(P||Q) = \int p(x) \log p(x)/q(x)$$

(a) Let $p(x) = \mathcal{N}(x; 0, 1)$. Find an expression for $KL(\mu, \Sigma)$ when $q(x) = \mathcal{N}(x; \mu, \Sigma)$.

A1.27 (α) (**Twelve Balls and a Balance**) You are given 12 balls, all equal in weight except for one that is either heavier or lighter. You are also given a two-pan balance (=terazi) to use. In each use of the balance you may put any number of the 12 balls on the left pan, and the same number on the right pan, and push a button to initiate the weighing; there are three possible outcomes: either the weights are equal, or the balls on the left are heavier, or the balls on the left are lighter. Your task is to design a strategy to determine which is the odd ball *and* whether it is heavier or lighter than the others *in as few uses of the balance as possible*.

While thinking about this problem, you may find it helpful to consider the following questions:

- (a) How can one measure *information*?
- (b) When you have identified the odd ball and whether it is heavy or light, how much information have you gained?
- (c) Once you have designed a strategy, draw a tree showing, for each of the possible outcomes of a weighing, what weighing you perform next. At each node in the tree, how much information have the outcomes so far given you, and how much information remains to be gained?
- (d) How much information is gained when you learn (i) the state of a flipped coin; (ii) the states of two flipped coins; (iii) the outcome when a four-sided die is rolled?
- (e) How much information is gained on the first step of the weighing problem if 6 balls are weighed against the other 6? How much is gained if 4 are weighed against 4 on the first step, leaving out 4 balls?