IMPERIAL COLLEGE LONDON

MSc COURSE IN COMPUTATIONAL METHODS IN ECOLOGY AND EVOLUTION

MULTIPLE CHOICE EXAM

For Internal Students of Imperial College of Science, Technology and Medicine

Exam Date: Monday, 29th March 2021, 10:00 – 12:00

Length of Exam: 2 HOURS

Instructions:

- You will be expected to answer 30 multiple choice questions in this exam.
- A question may have more than one correct answer and this should be indicated clearly in the questions wording.
- There will be exactly 5 options to choose from in each question
- There is no negative marking
- Please **follow the instructions for the remote assessment carefully** .
- This is an open book exam: you may refer to teaching, revision materials and online resources but you **must not confer with any other individual during the examination**.
- You will have 30 minutes after the end of the exam to prepare and upload answer files.

Continues on next page

Ecological modelling

Question 1

In the year 400 the human population had a size of 190 million people. In the year 1200 the population size was 360 million. In the year 1700 the human population had a size of 610 million, and in the year 1900 a size of 1625 million. Based on these data you can calculate the doubling times of the human population.

Which **one** of the following statements is true?

A) The doubling time of the human population after 1650 was shorter than before 1500. Therefore the human population was growing slower after 1650.

B) The doubling time of the human population after 1650 was longer than before 1500. Therefore the human population was growing slower after 1650.

C) The doubling time of the human population after 1650 was shorter than before 1500. Therefore the human population was growing faster after 1650.

D) The doubling time of the human population after 1650 was longer than before 1500. Therefore the human population was growing faster after 1650.

E) The doubling time of the human population after 1650 was the same as before 1500. Therefore the human population was growing at the same rate after 1650 as it did before 1500

Continues on next page

Question 2

Malaria is transmitted through mosquitos. On average a person gets bitten by 2 different mosquitoes per day. Upon biting an infected person there is a 10% chance that the person bitten becomes infectious. The average duration of the infection for humans is 10 months. What is the basic reproductive number of malaria?

A) 200

B) 60

C) 20

D) 6

E) 2

Continues on next page

Question 3

In a Hopf bifurcation an equilibrium can lose its stability and give rise to an limit cycle. Which **one** statement about the eigenvalues is true for an equilibrium at a Hopf bifurcation.

A) There is one eigenvalue with value zero

B) There is one complex eigenvalue with zero imaginary part

C) There is one complex eigenvalue with zero real part

D) There are two complex eigenvalues, both with zero imaginary part

E) There are two complex eigenvalues, both with zero real part

Continues on next page

Evolutionary modelling

Question 4

Which **one** of the following is **not** a consequence of genetic drift?

A) It changes the mean allele frequency over time.

B) It changes the variance of allele frequency over time.

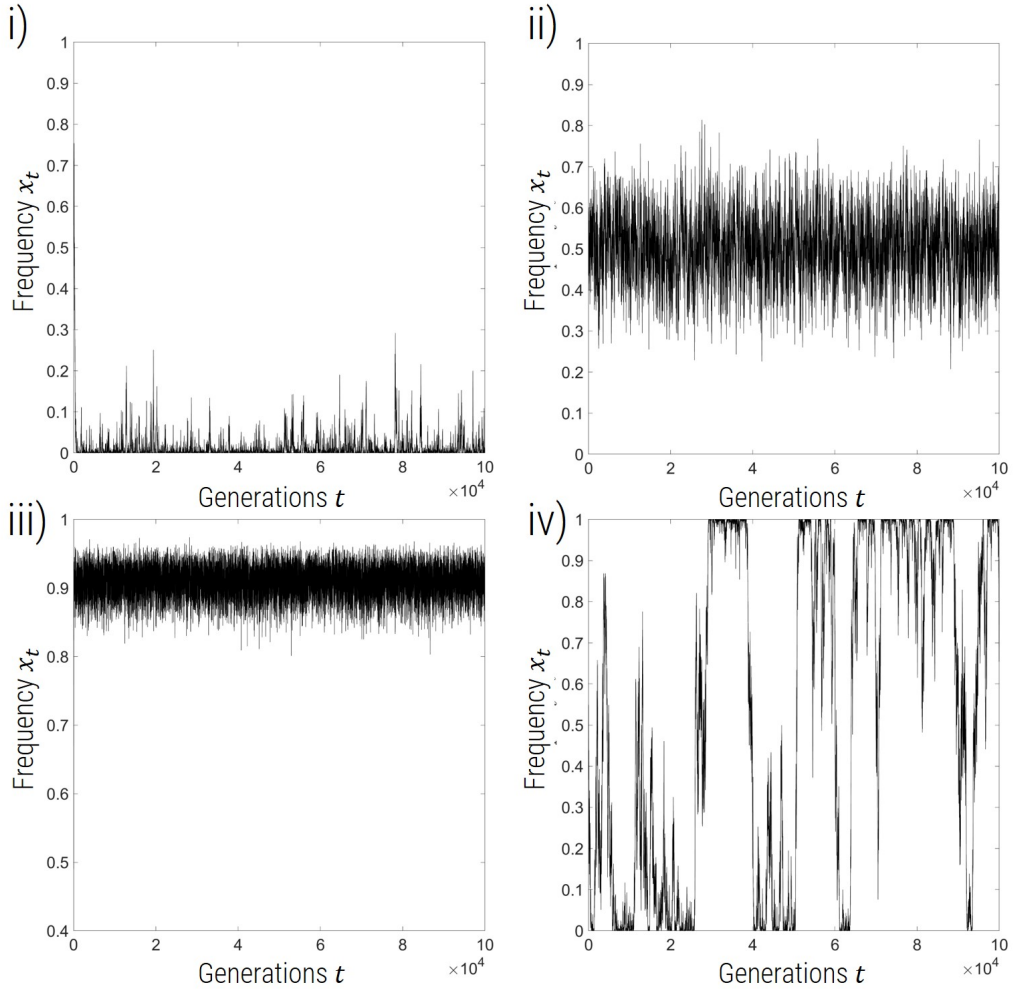
C) It reduces heterozygosity over time.

D) It ultimately drives an allele towards fixation or extinction.

E) None of the above.

Continues on next page

Question 5



Assume an idealised Wright-Fisher population of N individuals, the graphs above show the evolution of the frequency of a mutant allele A, whose frequency is x_t at generation t, where the mutation rate between wild type allele a and mutant is symmetric and equal to μ. In addition, given a selection coefficient s for mutant vs wild type (i.e. fitness$(a) = w$, fitness$(A) = w(1+s)$), which **one** of the following options correctly describes the different regimes of evolutionary dynamics for each graph

A) $\begin{bmatrix} \text{i)} & N\mu \gg 1, s > 0, N|s| \gg 1 & \text{ii)} & N\mu \gg 1, N|s| \ll 1 \\ \text{iii)} & N\mu \ll 1, s < 0, N|s| \gg 1 & \text{iv)} & N\mu \ll 1, N|s| \ll 1 \end{bmatrix}$

B) $\begin{bmatrix} \text{i)} & N\mu \gg 1, s > 0, N|s| \gg 1 & \text{ii)} & N\mu \ll 1, N|s| \ll 1 \\ \text{iii)} & N\mu \ll 1, s < 0, N|s| \gg 1 & \text{iv)} & N\mu \gg 1, N|s| \ll 1 \end{bmatrix}$

C) $\begin{bmatrix} \text{i)} & N\mu \ll 1, s < 0, N|s| \gg 1 & \text{ii)} & N\mu \gg 1, s > 0, N|s| \gg 1 \\ \text{iii)} & N\mu \ll 1, N|s| \ll 1 & \text{iv)} & N\mu \gg 1, N|s| \ll 1 \end{bmatrix}$

D) $\begin{bmatrix} \text{i)} & N\mu \ll 1, s < 0, N|s| \gg 1 & \text{ii)} & N\mu \gg 1, N|s| \ll 1 \\ \text{iii)} & N\mu \gg 1, s > 0, N|s| \gg 1 & \text{iv)} & N\mu \ll 1, N|s| \ll 1 \end{bmatrix}$

E) $\begin{bmatrix} \text{i)} & N\mu \ll 1, s < 0, N|s| \gg 1 & \text{ii)} & N\mu \gg 1, s > 0, N|s| \gg 1 \\ \text{iii)} & N\mu \ll 1, N|s| \ll 1 & \text{iv)} & N\mu \gg 1, N|s| \ll 1 \end{bmatrix}$

Continues on next page

Question 6

A small group of n_0 individuals migrate to an idealised Wright-Fisher population of N individuals, carrying an allele A that has a selective advantage $s > 0$, compared to the wild type allele a. Assume that the wild type allele is fixed before the arrival of the migrants, that there are no further emigrations or immigrations and that $n_0 \ll N$. Which **one** of the following conditions would ensure that the migrant allele goes to fixation with probability near 1.

A) $n_0 = \frac{1}{2s}$

B) $n_0 \gg \frac{1}{2s}$

C) $n_0 > \frac{1}{2s}$

D) $n_0 < \frac{1}{2s}$

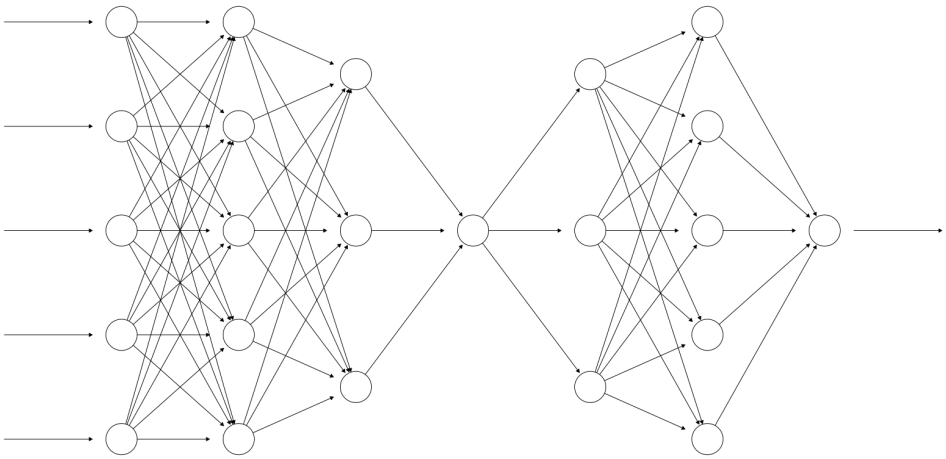
E) $n_0 \ll \frac{1}{2s}$

Continues on next page

Machine learning

Question 7

The diagram below depicts an artificial neural network. Select the **one** true statement about the network.



A) This is an autoencoder.

B) This network could be specified using the following R code:

```
library(neuralnet)
y <- rnorm(1000)
x <- rnorm(1000)
model <- neuralnet(y ~ x, hidden=c(5,3,1,3,5), data=data)
```

C) This network could be specified using the following R code:

```
library(keras); install_keras()
model <- keras_model_sequential()
model %>%
  layer_dense(units = 5) %>%
  layer_dense(units = 3) %>%
  layer_dense(units = 1) %>%
  layer_dense(units = 3) %>%
  layer_dense(units = 5)
```

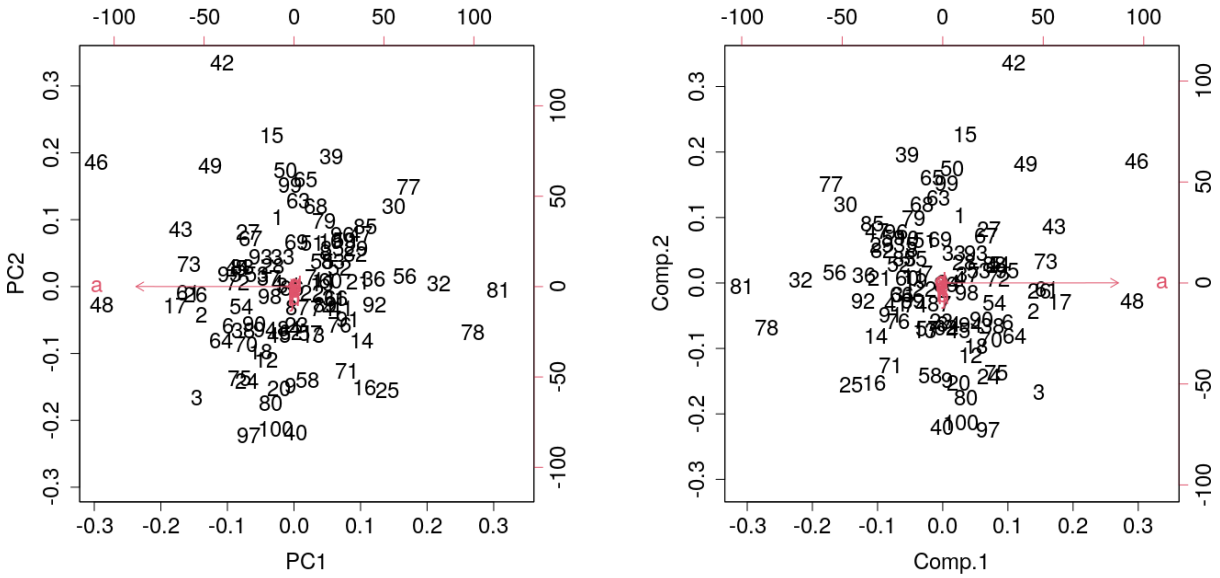
D) The figure above lacks bias neurons, and so does not represent an artificial neural network.

E) None of the above

Continues on next page

Question 8

The figure below shows a biplot from a PCA (left) and a PCoA (right) of ten continuous variables (named after the first ten letters of the alphabet). Choose the **one** correct statement from the options below.



A) The differences between these biplots are unimportant. The rescaling of a PCoA by the largest eigenvalue (i.e., variable a) drives the apparent difference (note the re-scaled upper-most axes).

B) Differences in the variances of the variables are causing the differences between these plots. The PCoA must have been performed on the covariance or correlation matrix of the data.

C) There is no way that the PCA and PCoA have been conducted on the same data. The data are in completely different parts of component space (note the numbered rows of data and their positions).

D) The differences reflect the inability of the the PCA to map variation in variable a into a Euclidean space.

E) None of the above.

Continues on next page

Question 9

Below is a summary of a dataset that you wish to analyse. Select the **one** answer whose code comments accurately describe what is going on in the R code. Note that the package calls (e.g., library(e1071)) are all correct; there are no trick answers!

```
> head(data)
           y          x           z location
1 -0.8791793 -0.8791793 -0.09777731        b
2  2.0509622  2.0509622  0.54911953        c
3  0.3938882  0.3938882 -0.16198026        c
4  1.0002481  1.0002481 -1.30370694        b
5  0.5497598  0.5497598 -0.03998964        a
6  0.5894560  0.5894560  0.99029963        a
> str(data)
'data.frame':^^I1000 obs. of  4 variables:
 $ y       : num  -0.879 2.051 0.394 1 0.55 ...
 $ x       : num  -0.879 2.051 0.394 1 0.55 ...
 $ z       : num  -0.0978 0.5491 -0.162 -1.3037 -0.04 ...
 $ location: chr  "b" "c" "c" "b" ...
```

A)
```
# Load random forest library
library(randomForest)
# Correctly run a bagged regression tree
model <- randomForest(y~., data=data, mtry=ncol(data)-1)
# Validate model predictions with independent data
cor(data$y, predict(model))
```

B)
```
# Load LAR library
library(lars)
# Correctly run a least angle regression
model <- lars(cbind(data$x,data$z), data$y, type="lar", normalize=FALSE)
# Check coefficients and arc length
coef(model)
```

C)
```
# Load SVM library
library(e1071)
# Correctly run a SVM
model <- svm(y~., data=data, type="eps-regression")
# Examine model parameters
summary(model)
```

D)
```
# Correctly run PCA
model <- prcomp(data, scale=TRUE)
# Examine PCA biplot
biplot(data)
```

E) None of the above

Continues on next page

Biological data and C

Question 10

Consider the following lines of code

```
void *ptr;
char *message = "It was a bright cold day in April\n";

ptr = message;
```

Which **two** of the following program statements is correctly described by its associated comment?

A) `*ptr; // Evaluates to a string variable`

B) `(char *)ptr; // Evaluates to a string variable, but theres no guarantee in this case.`

C) `*(char *)ptr; // Evaluates to the character 'I'.`

D) `(char *)ptr; // Evaluates to the character 'I'.`

E) `*(char *)ptr; // In this case, it is guaranteed to evaluate to a string variable.`

Continues on next page

Question 11

The following code has program errors the compiler wont catch.

```
struct node {
    int         index;
    void        *data;
    struct node *next;
};

/**
 A function to create memory for a single node
 */
struct node *create_node(void)
{
    struct node *newnd = malloc(sizeof(struct node *));

    return newnd;
}

struct node ** build_linked_list(int nelems)
{
    int i = 0;
    struct node **ndlist;

    ndlist = malloc(nelems * sizeof(struct node *));

    for (i = 0; i < nelems-1; ++i) {
        ndlist[i]->next = ndlist[i + 1];
    }
    ndlist[i]->next = NULL;

    return ndlist;
}
```

What are the **two** errors?

A) The allocation is never checked for NULL return from `malloc()`.

B) The call to `malloc` should use `sizeof(struct node)`.

C) The memory allocated to `ndlist` by `malloc` will be of the incorrect size.

D) After the for-loop, the NULL initialization will be out of bounds.

E) The linked list is declared as both an array and a linked list. But it should only be one of these.

Continues on next page

Bayesian statistics

Question 12

Which **two** of the following statements are true on the use of Bayesian statistics for data analysis?

A) It is only used when a simulator is available to generate data points to compare against observations.

B) It requires a sampling model as in a likelihoodist approach.

C) It never uses the observed data to define prior distributions.

D) It can use the observed data to define prior distributions.

E) It conditions on the parameters and integrates over the data.

Continues on next page

Question 13

Which **one** of the following statements is true on prior distributions?

A) Non-informative priors can be modelled as uniform distributions.

B) Non-informative priors are always modelled as uniform distributions.

C) Informative priors can not be modelled as uniform distributions.

D) Without informative priors it is not possible to use Bayesian statistics.

E) With non-informative priors, the maximum a posteriori point estimate of a parameter is equal to its p-value.

Continues on next page

Question 14

Which **two** of the following statements are true on Approximate Bayesian Computation (ABC)?

A) ABC is only useful when the likelihood function is defined.

B) ABC is useful when the likelihood function can not be defined or evaluated.

C) ABC suffers from the curse-of-dimensionality when the number of summary statistics is significantly lower than the dimensionality of the data.

D) In the discrete case, the ABC rejection algorithm requires a tolerance to compare simulated with observed data points.

E) Regression-based ABC potentially allows for a larger acceptance rate.

Continues on next page

Maximum Likelihood

Question 15

Let $X \sim$ Uniform$(a,\ b)$, with pdf $f_X(x) = \frac{1}{b-a}$, $a \leq x \leq b$. What is $E\left(X^3\right)$?

A) $\dfrac{(b+a)^3}{4}$

B) $\dfrac{(b^2+a^2)(b+a)}{4}$

C) $\dfrac{(b+a)(b-a)}{4}$

D) $\dfrac{b^3+ba+a^3}{4}$

E) $\dfrac{b^3+ba+a^3}{2}$

Continues on next page

Question 16

Which **one** of the following statements about likelihood functions and maximum likelihood estimators are TRUE?

(i) Likelihood is a function of the parameters.

(ii) Both the log-likelihood and original likelihood are maximised under the same parameter value.

(iii) Maximum likelihood estimators are unbiased.

(iv) Maximum likelihood estimators are asymptotically normal, when sample size is large.

A) (i) and (ii) only.

B) (i), (ii), and (iii) only.

C) (i), (iii), and (iv) only.

D) (ii), (iii), and (iv) only.

E) All of the above.

Continues on next page

Question 17

Let θ be the parameter (univariate) of interest, $\widehat{\theta}$ be the MLE, and $\ell(\theta)$ be the log-likelihood function. Which **one** of the following statements regarding confidence interval (C.I.) estimation are TRUE?

(i) The 95% C.I. for θ is the collection of parameter values whose log-likelihood values are within -1.92 units from the maximum (i.e. $\ell(\widehat{\theta})$).

(ii) The 95% C.I. can also be inferred by assuming normality when sample size is large. Under this case, the 95% C.I. is approximately $\widehat{\theta} \pm 1.96\sqrt{Var(\widehat{\theta})}$.

(iii) $Var(\widehat{\theta})$ can be approximated from the second derivative of the log-likelihood function, evaluated at $\widehat{\theta}$.

(iv) $Var(\widehat{\theta})$ can be found by setting $\ell'(\theta) = 0$.

A) (i) and (ii) only.

B) (i), (ii), and (iii) only.

C) (i), (ii), and (iv) only.

D) None of the above.

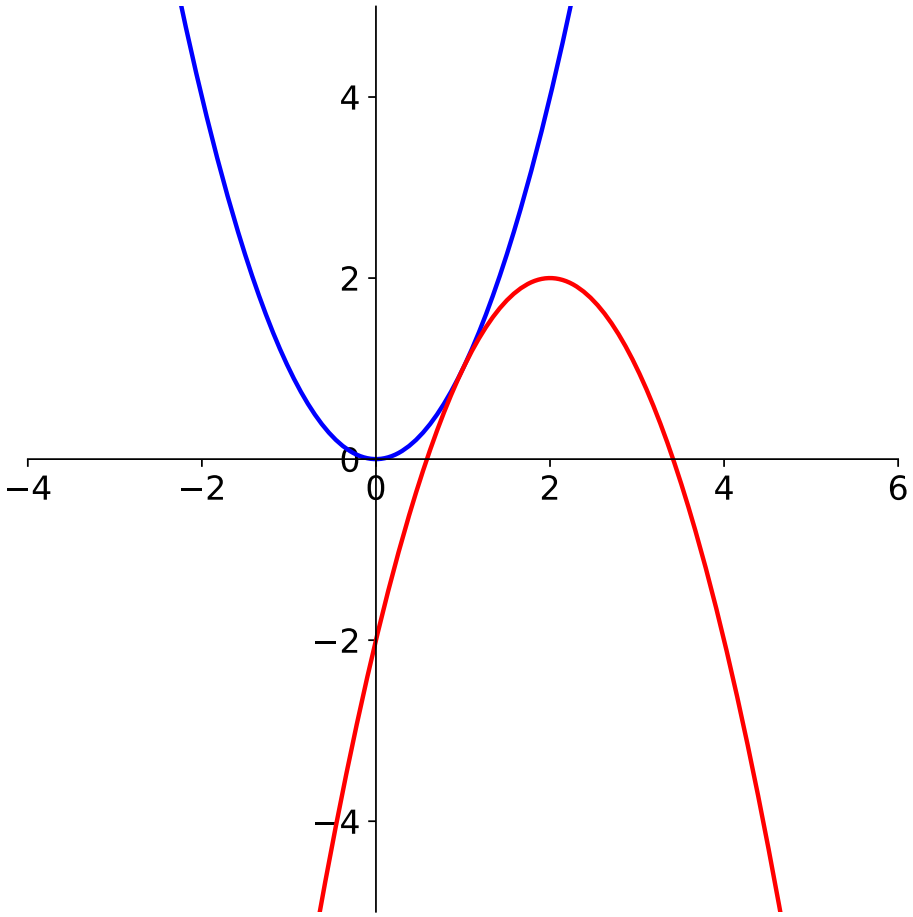
E) All of the above.

Continues on next page

Maths

Question 18

If the function $f(x) = x^2$ corresponds to the blue curve, what function corresponds to the red curve?



A) $g(x) = (x-2)^2 + 2$

B) $g(x) = (x+2)^2 - 2$

C) $g(x) = -(x-2)^2 + 2$

D) $g(x) = -(x+2)^2 + 2$

E) $g(x) = -(x-2)^2 - 2$

Continues on next page

Question 19

For positive constants a and k, what is the linear approximation of the function $f(R) = \dfrac{aR}{k+R}$ for R close to k?

A) $\dfrac{a}{k}R$

B) $\dfrac{a}{2} + \frac{a}{4k}(R-k)$

C) $\dfrac{a}{2} + \frac{a}{4}(R-k)$

D) $\dfrac{a}{2} + \frac{a}{4k}R$

E) $-\dfrac{a}{8k^2}(R-k)^2$

Continues on next page

Question 20

Assume that a population is divided into two age classes and that 50% of the individuals age 0 survive until the end of the next breeding season. Assume further that individuals age 0 have an average of 0.5 individual offspring and individuals age 1 have an average of 4 individual offspring. What is the projection matrix for this age-structured model?

A) $\begin{pmatrix} 0 & 4 \\ 50 & 0.5 \end{pmatrix}$

B) $\begin{pmatrix} 0 & 4 \\ 0.5 & 0.5 \end{pmatrix}$

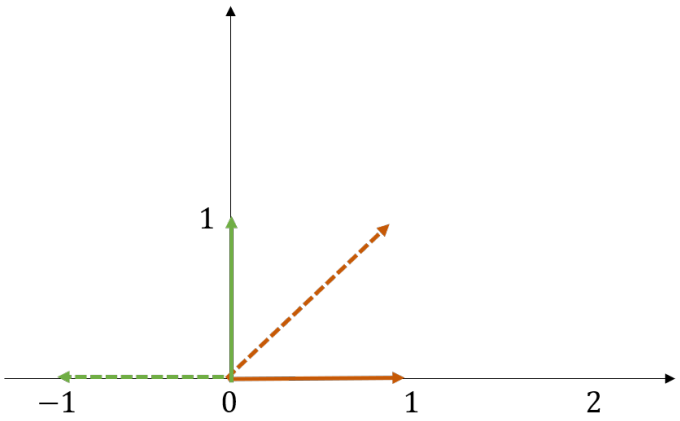
C) $\begin{pmatrix} 0.5 & 4 \\ 50 & 0 \end{pmatrix}$

D) $\begin{pmatrix} 0.5 & 0 \\ 50 & 4 \end{pmatrix}$

E) $\begin{pmatrix} 0.5 & 4 \\ 0.5 & 0 \end{pmatrix}$

Continues on next page

Question 21



A matrix $\boldsymbol{A}$ acts on the unit vectors $\underline{e}_1$ (red solid arrow) and $\underline{e}_2$ (solid green vector) to produce vectors $\underline{a}_1 = \boldsymbol{A}\underline{e}_1$ (dashed red arrow) and $\underline{a}_2 = \boldsymbol{A}\underline{e}_2$ (dashed green arrow), respectively. Which of the choices below correctly describes the matrix $\boldsymbol{A}$, it's determinant $|\boldsymbol{A}|$ and eigenvalues λ:

A) $\boldsymbol{A} = \begin{pmatrix} 1 & -1 \\ 1 & 0 \end{pmatrix}, \quad |\boldsymbol{A}| = -1, \quad \lambda = \frac{1}{2} \pm \frac{i\sqrt{3}}{2}$

B) $\boldsymbol{A} = \begin{pmatrix} -1 & -1 \\ 1 & 0 \end{pmatrix}, \quad |\boldsymbol{A}| = -1, \quad \lambda = -\frac{1}{2} \pm \frac{i\sqrt{3}}{2}$

C) $\boldsymbol{A} = \begin{pmatrix} 1 & -1 \\ 1 & 0 \end{pmatrix}, \quad |\boldsymbol{A}| = 1, \quad \lambda = \frac{1}{2} \pm \frac{i\sqrt{3}}{2}$

D) $\boldsymbol{A} = \begin{pmatrix} 1 & -1 \\ 1 & 0 \end{pmatrix}, \quad |\boldsymbol{A}| = 1, \quad \lambda = -\frac{1}{2} \pm \frac{i\sqrt{3}}{2}$

E) $\boldsymbol{A} = \begin{pmatrix} -1 & -1 \\ 1 & 0 \end{pmatrix}, \quad |\boldsymbol{A}| = 1, \quad \lambda = -\frac{1}{2} \pm \frac{i\sqrt{3}}{2}$

Continues on next page

Question 22

A differential equation describing the dynamics of a variable x is given by

$$\frac{\mathrm{d}x}{\mathrm{d}t} = f(x)$$

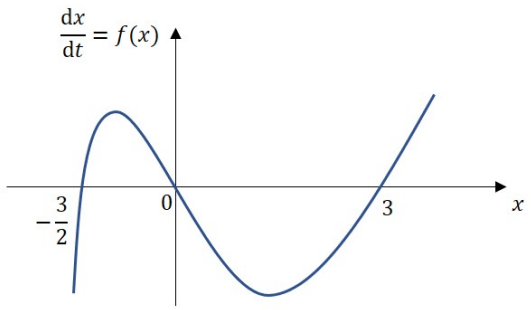
.

For different initial values of $x(0) = x_0$ the following occur:

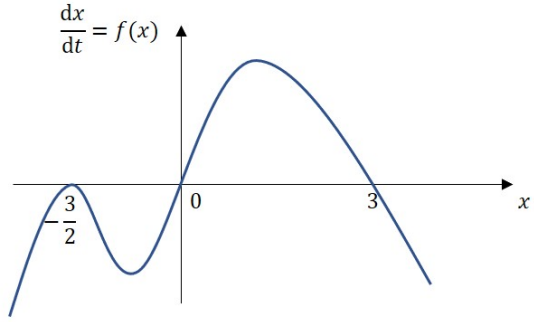
- $x_0 < -3/2$: x increases and stops asymptotically at $x = -3/2$ as $t \to \infty$
- $-3/2 < x_0 < 0$: x decreases and stops asymptotically at $x = -3/2$ as $t \to \infty$
- $0 < x_0 < 3$: x increases and stops asymptotically at $x = 3$ as $t \to \infty$
- $x_0 > 3$: x decreases and stops asymptotically at $x = 3$ as $t \to \infty$

Which of the following phase-portraits is consistent with the above events.

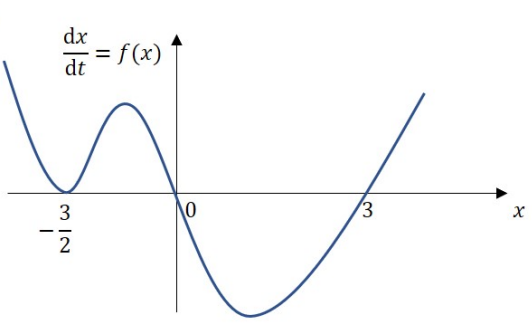
a)



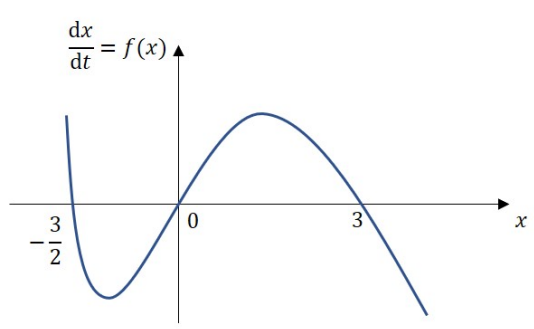
b)



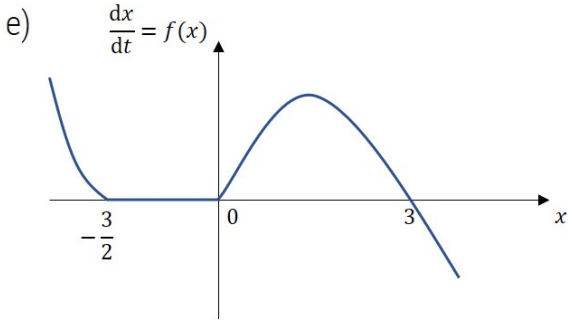
c)



d)



e)



Continues on next page

Genomics

Question 23

Which **three** of the following statements are true regarding the VCF file format?

A) It is a space-delimited csv file that stores information about gene sequence variations

B) It can contain information on individual genotype in a position of the sequenced genome

C) It results after mapping raw reads to a reference sequence.

D) It results after calling variants from mapped reads.

E) It is useful for visualizing information of sequencing data.

Continues on next page

Question 24

Which **two** of the following statements are true regarding the infinite sites model?

A) It allows us to distinguish between ancestral and derived alleles.

B) It assumes that sequences are short enough so that the probability of recurrent mutations is low

C) It assumes that sequences are infinitely long so that the probability of recurrent mutations is moderate.

D) It is a valid assumption for species with short genomes and high mutation rate.

E) It is often used to model genetic variation in higher eukaryotes.

Continues on next page

Question 25

Which **three** of the following statements are true regarding coalescence theory with 2N individual gene copies?

A) The probability of finding the first common ancestor four generations ago is $[1 - 1/(2N)]^4 * [1/(2N)]$

B) The probability of finding the first common ancestor four generations ago is $[1 - 1/(2N)] * [1 - 1/(2N)] * [1 - 1/(2N)] * [1/(2N)]$

C) If we make the assumption of an infinitely large population, we can model coalescent rates as a continuous distribution.

D) The probability of two individual gene copies having the same parent in the previous generation is $2N$.

E) The probability of two individual gene copies not having the same parent in the previous generation is $(2N - 1)/2N$.

Continues on next page

GIS

Question 26

The text below is the Well Known Text representation of a GIS coordinate system.

```
PROJCS["OSGB 1936 / British National Grid",
    GEOGCS["OSGB 1936",
        DATUM["OSGB_1936",
            SPHEROID["Airy 1830", 6377563.396, 299.3249646,
                     AUTHORITY["EPSG","7001"]],
            TOWGS84[446.448, -125.157, 542.06, 0.15,
                    0.247, 0.842, -20.489],
            AUTHORITY["EPSG","6277"]],
        PRIMEM["Greenwich",0,AUTHORITY["EPSG","8901"]],
        UNIT["degree", 0.0174532925199433, AUTHORITY["EPSG","9122"]],
        AUTHORITY["EPSG","4277"]],
    PROJECTION["Transverse_Mercator"],
    PARAMETER["latitude_of_origin",49],
    PARAMETER["central_meridian",-2],
    PARAMETER["scale_factor",0.9996012717],
    PARAMETER["false_easting",400000],
    PARAMETER["false_northing",-100000],
    UNIT["metre",1,AUTHORITY["EPSG","9001"]],
    AXIS["Easting",EAST],
    AXIS["Northing",NORTH],
    AUTHORITY["EPSG","27700"]]
```

Which **two** of the following statements are true:

A) This is a geographic coordinate system

B) It uses the WGS84 geographic datum

C) The units of the coordinate system are in metres

D) GPS point with a longitude of -2° will have an X value around 400000

E) A GPS point with a latitude of 49° will have an X value around 400000

Continues on next page

Question 27

Which **three** of the following statements about spatial modelling are correct?

A) Spatial autoregression requires a list defining cells that are neighbours

B) The y-axis values on a correlogram are typically smallest at short distances.

C) Data display stationarity if the spatial autocorrelation is the same in all directions.

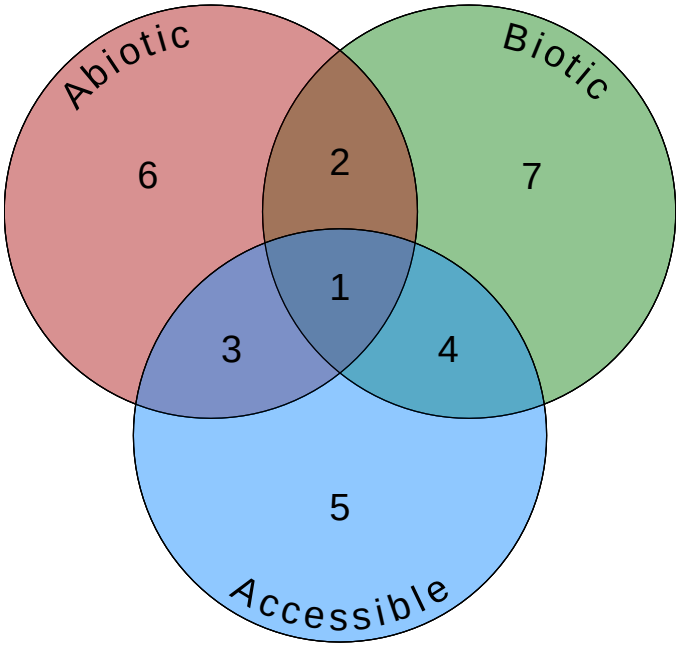
D) Spatial eigenvector filters can capture spatial autocorrelation that is not stationary or isotropic

E) Geographically weighted regression fits a model for every point in a spatial dataset.

Continues on next page

Question 28

The plot below shows a visualisation of Soberón & Petersons (2005) partitioning of species distribution into abiotic and biotic niches and accessible area. Which **three** of the following statements are true?



A) Sink populations are found exclusively in zone 5.

B) Zones 1 and 2 form the realised niche.

C) In a species distribution model, pseudo-absence or background points must be selected from locations in zones 6 and 7.

D) Species distribution models are likely to underpredict the suitability of zone 2 and overpredict the suitability of zones 3, 4 and 5.

E) Zone 1 is the species core geographic range.

Continues on next page

Generalised Linear Modelling

Question 29

Mixed models can be used to estimate the variance explained by a random effect. Using the output from a mixed model where student group (Group.Code) is fitted as a random effect and log overall grade for each individual as the response variable. Using the information from the output, approximate the percentage of variation explained by Group.Code and select **one** option below.

```
Linear mixed model fit by REML ['lmerMod']
Formula: log(Overall) ~ 1 + (1 | Group.Code)
   Data: biochem

REML criterion at convergence: 8.7

Scaled residuals:
    Min      1Q  Median      3Q     Max
-4.2408 -0.3754  0.2137  0.6686  1.3094

Random effects:
 Groups     Name        Variance Std.Dev.
 Group.Code (Intercept) 0.00352  0.05933
 Residual               0.05740  0.23959
Number of obs: 126, groups:  Group.Code, 23

Fixed effects:
            Estimate Std. Error t value
(Intercept) -0.45513    0.02477  -18.38
```

A) 5.78%

B) 6.13%

C) 0.06%

D) 19.84%

E) 0.20%

Continues on next page

Question 30

The z-value in a generalised linear model is similar to what metric in linear models? Select **one** correct option.

A) t-value

B) F-value

C) Slope coefficient

D) Intercepts coefficient

E) The adjusted R^2

End of paper