1. (1 point) We have seen some data $\mathcal{D}$ which we try to represent using parameter θ. Which combination of "semantic" names for the distribution below is correct?

$$\underbrace{p(\theta|\mathcal{D})}_{A} = \frac{\overbrace{p(\mathcal{D}|\theta)}^{B}\,\overbrace{p(\theta)}^{C}}{\underbrace{p(\mathcal{D})}_{D}}$$

(a) {A,B,C,D} = {Prior, Evidence, Posterior, Likelihood}
(b) {A,B,C,D} = {Likelihood, Prior, Evidence, Posterior}
(c) {A,B,C,D} = {Posterior, Likelihood, Prior, Evidence}
(d) {A,B,C,D} = {Posterior, Joint, Posterior, Likelihood}
(e) {A,B,C,D} = {Evidence, Posterior, Likelihood, Prior}

2. (1 point) In Bayesian learning, the structure of the function we infer (linear regression, GP, neural network etc.) is part of which probability distribution,

(a) The prior
(b) The likelihood
(c) The posterior
(d) None of the above

3. (1 point) Given the probability table below what will the expected value of y be?

y_i	$p(y = y_i)$
0	0.1
1	0.2
2	0.5
3	0.1
4	0.1

(a) 2.0
(b) 0.5
(c) 0.2
(d) 1.9

4. (1 point) You have specified a model with a Hypergeometric likelihood function with known population size and you now want to derive the posterior distribution over the number of target members. The conjugate prior for the target members is a Beta-binomial distribution. Using the conjugate prior which functional form will the posterior have?

(a) Gaussian
(b) Beta-binomial
(c) Poission
(d) Cathegorical
(e) Dirichlet

5. (1 point) If there are multiple models which can explain the observed data equally well, at least one of the models must be based on wrong assumptions. Is this statement true?

   (a) Yes
   (b) No

6. Which of the following statements regarding the two random variables $\mathbf{x}$ and $\mathbf{y}$ is false,

   (a) If $\mathbf{x}$ and $\mathbf{y}$ are jointly Gaussian, $\mathbf{x}$ is a Gaussian and $\mathbf{y}$ is a Gaussian.
   (b) If $\mathbf{x}$ and $\mathbf{y}$ are jointly Gaussian, $p(\mathbf{x}|\mathbf{y})$ is a Gaussian.
   (c) If $\mathbf{x}$ is a Gaussian and $\mathbf{y}$ is a Gaussian, $\mathbf{x}$ and $\mathbf{y}$ are jointly Gaussian.
   (d) If $\mathbf{x}$ and $\mathbf{y}$ are part of a Gaussian process, $\mathbf{x}$ and $\mathbf{y}$ are jointly Gaussian.

7. (1 point) A Dirichlet process can be constructed with what is known as the "stick-breaking" construction as follows,

$$\beta'_0 = \text{Beta}(1, \alpha) \tag{1}$$
$$\beta'_k = \text{Beta}(1, \alpha), \forall k > 1 \tag{2}$$
$$\beta_k = \beta'_k \prod_{i=1}^{k} (1 - \beta'_i), \tag{3}$$

   where the set of sticks after K breaks is,

$$\boldsymbol{\beta} = \{\beta_0, \beta_1, \ldots, \beta_K\}.$$

   Using this construction will each "break" of the stick result in a set of *strictly* shorter sticks, eg. that stick i is always longer than stick $i+1$?

   (a) No
   (b) Yes

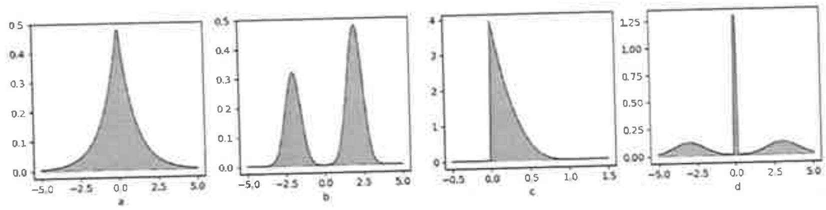
8. (1 point) When describing a set of data $\mathbf{y}$ using a conditional probability $p(\mathbf{y}|\theta)$ which of the following statements is **always** correct?

   (a) θ and $\mathbf{y}$ are in a causual relationship such that the former causes the latter
   (b) $\mathbf{y}$ can never be the cause of θ

(c) θ parametrises the distribution over $\mathbf{y}$

(d) Conditional distributions always correspond to causual relationships

9. (1 point) The Laplace approximation fits a Gaussian distribution around the mode of the posterior. Consider the following posteriors in the plot below, in which scenario will the Laplace approximation provide a good approximation to the true posterior?



(a) a

(b) b

(c) c

(d) d

10. (1 point) When writing a generative model we describe the process that the data has been generated. Considering a set of data $\mathbf{y}$ and generative model with the following joint distribution,

$$p(\mathbf{y}, f, x, \theta, \gamma) = p(\mathbf{y}|f)p(f|x, \theta)p(\theta)p(x|\gamma)p(\gamma)$$

which of the following statements is false for the model above?

(a) The factorisation on the right hand side above specifies a more restricted probability distribution compared to the general joint distribution,

$$p(\mathbf{y}, f, x, \theta, \gamma).$$

(b) In order to generate data from the model we need to have seen training data

(c) We can generate data from the model above using ancestral sampling, this means that we start sampling from the γ, and θ knowing γ we can sample x from $p(x|\gamma)$ and knowing θ and x we can sample from $p(f|x, \theta)$ and finally knowing f we can sample y.

(d) Using the rules of probability we can re-write the factorisation above

11. (1 point) When learning a set of parameters using a maximum likelihood formulation we will recover the parameters that maximises the likelihood of the observed data. Contrasting this with the posterior distribution of the parameters which of the following statements is true?

(a) Given sufficient data the maximum likelihood solution will **always** be the mode of the posterior distribution

(b) Given sufficient data the maximum likelihood solution will **never** be the mode of the posterior distribution

(c) Given sufficient data and a prior distribution with support everywhere (i.e. it is non-zero everywhere) the maximum likelihood solution will always be the mode of the posterior distribution

12. (1 point) In many sampling approaches we use a proposal distribution to draw samples from, which of the following statements is true for sampling using a proposal distribution?

(a) in the limit sampling will always converge to the true distribution independent of the form of the proposal distribution

(b) the sampling proceedure is only guaranteed to converge to the true distribution provided that the proposal distribution have support everywhere the true distribution has

(c) the speed of convergence for the sampling proceedure is independent of the form of the proposal distribution

13. (1 point) In Bayesian optimisation we aim to find the global extrema of a explicitly unknown function. Which of the following statements is true for Bayesian optimisation?

(a) We will always recover the global solution

(b) The time it takes to find the solution is independent of the choice of aquisition function

(c) The time it takes to find the solution is independent of the prior on the function

(d) With a determinstic aquisiton function each run of Bayesian optimisation on the same function will return the same result

14. (1 point) A Gaussian process is completely defined by its mean and covariance function. Which of the following statements is false for a Gaussian process prior?

(a) The mean function is always a constant function

(b) The covariance function describes how the function values co-vary along the input domain

(c) The covariance between two function values are completely defined by their input locations

(d) The Gaussian process prior is defined over an uncountable infinite index set

(e) The covariance matrix evaluated on any subset of the index set will always generate a positive definite covariance matrix

15. (1 point) Assuming that we are able to draw samples from a uniform distribution. We now want to transform these samples to a different distribution $p(\cdot)$ which function should we use to transform the samples

    (a) The cumulative distribution function
    (b) The inverse of the cummulative distribution function
    (c) The probability density function
    (d) The posterior distribution
    (e) The prior distribution

16. (1 point) We often say that computing the posterior is intractable due to the computation of which term in Bayes' rule?

    (a) The Likelihood
    (b) The Prior
    (c) The Evidence
    (d) The Joint Distribution

17. (1 point) In variational inference we use a distribution $q(\boldsymbol{\theta})$ to approximate an intractable posterior $p(\boldsymbol{\theta}|\mathbf{Y})$. In order to fit q to p we formluate a lower bound on which distribution?

    (a) The posterior $p(\boldsymbol{\theta}|\mathbf{Y})$
    (b) The prior $p(\boldsymbol{\theta})$
    (c) The likelihood $p(\mathbf{Y}|\boldsymbol{\theta})$
    (d) The joint distribution $p(\mathbf{Y}, \boldsymbol{\theta})$
    (e) The evidence $p(\mathbf{Y})$

18. (1 point) Considering mapping a distribution through a series of functions, what will happen with the "volume" of the distribution after the final function compared to the first. Which of the following statement is true?

    (a) it will shrink if at least one of the functions are non-monotonic
    (b) it will shrink only if all functions are non-monotonic
    (c) it will grow for each consequtive composition
    (d) it will stay the same independent of the functions in the composition

19. (1 point) Consider the marginal likelihood of a Gaussian process regression model,

$$p(\mathbf{Y}|\mathbf{X},\boldsymbol{\theta}) = \int p(\mathbf{Y}|\mathbf{f})p(\mathbf{f}|\mathbf{X},\boldsymbol{\theta})\mathrm{d}\mathbf{f}$$

If we want to maximise the marginal likelihood with respect to $\boldsymbol{\theta}$ it is equvivalent to minimise the negative logarithm of the marginal as,

$$\mathrm{argmax}_{\boldsymbol{\theta}} p(\mathbf{Y}|\mathbf{X},\boldsymbol{\theta}) = \mathrm{argmin}_{\boldsymbol{\theta}} - \log\left(p(\mathbf{Y}|\mathbf{X},\boldsymbol{\theta})\right) = \mathrm{argmin}_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}) \tag{4}$$

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} + \frac{1}{2}\log|\mathbf{K}| + \frac{N}{2}\log(2\pi) \tag{5}$$

Considering the terms above, the determinant will naturally encourage a solution with a slowly varying function. Which of the following statements about the determinant of a co-variance matrix is false?

(a) The maximum value of the determinant is for a diagonal co-variance matrix

(b) The minimum value of the determinant is for a rank-deficient co-variance matrix

(c) The minimum value of the determinant is when all instantiations of the function are independent

(d) The determinant of the covariance function is always positive

20. (1 point) Consider the factorisation of the joint distribution shown below, which of the following statments regarding the assumptions made is false?

$$p(a,b,c,d,e,f) = p(a|c,d)p(b|c,d)p(c|e)p(d|f)p(e)p(f)$$

(a) a and b are conditionally independent given c and d

(b) a and b are independent

(c) e and f are independent

(d) Given e, c is independent from f

End of Paper