

Duration 120 mins (2 hours)

Answer all questions. Section A (worth 40 marks) are a series of questions with short answers. Section B (worth 60 marks) involve longer questions

Question	Mark	<i>Arithmetic checked</i>	<i>Double Marked</i>
Total:			

Page 1 of 16

Section A

Question A 1

- (a) Explain why over expressive machines are likely to generalise poorly when the number of training examples is small. (5 marks)

5

- (b) Explain why CNNs capture the structure of typical image datasets. (5 marks)

5

- (c) Explain the major ways (i) gradient descent (ii) Newton's method and (iii) quasi-Newton methods differ in terms of the information they use. (5 marks)

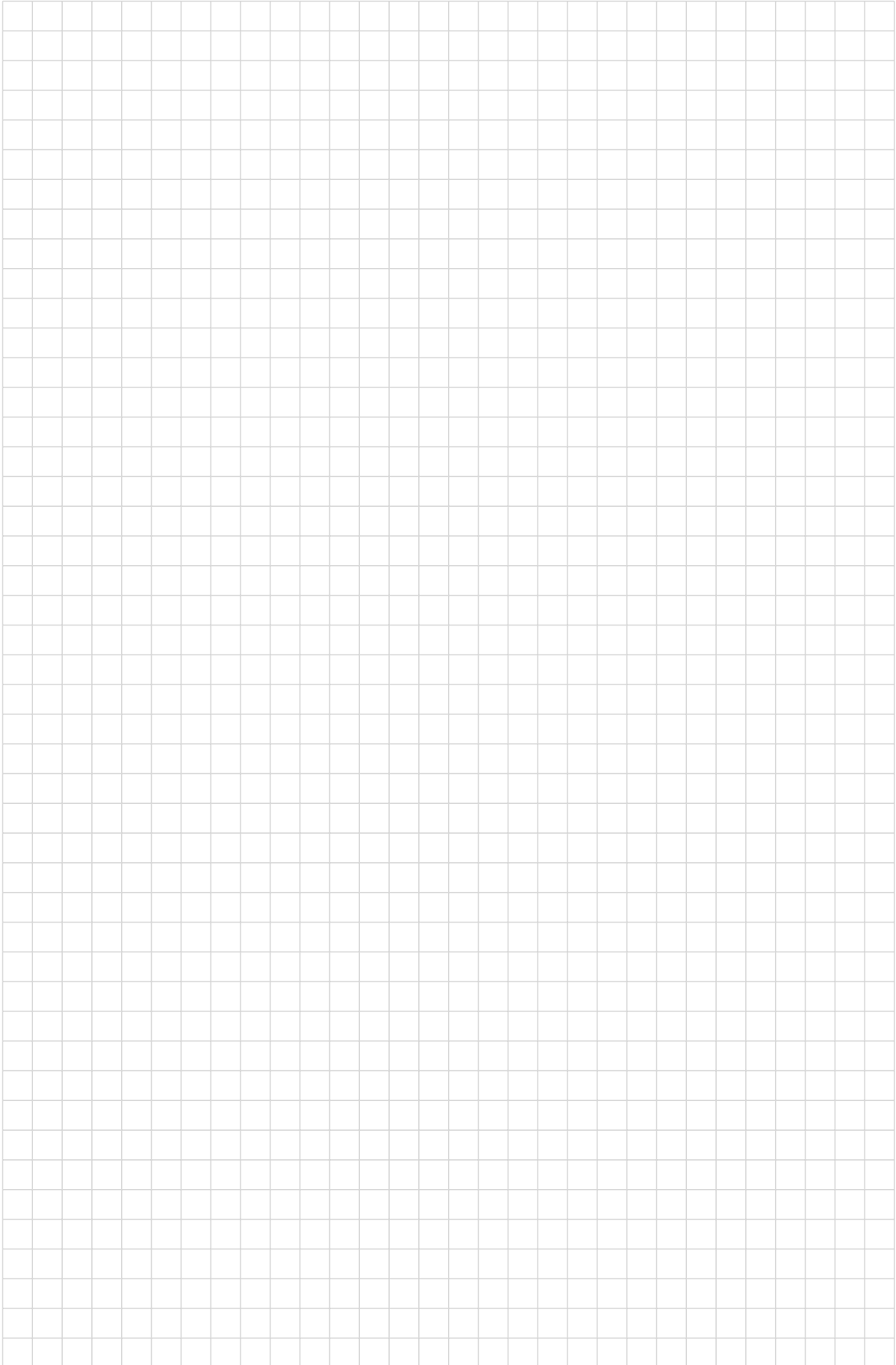
i

ii

iii

5

Additional space. Do not use unless necessary. Clearly mark corresponding question.



- (d) In stochastic gradient descent (SGD) explain what are mini-batches and their possible advantages and disadvantages. (5 marks)

5

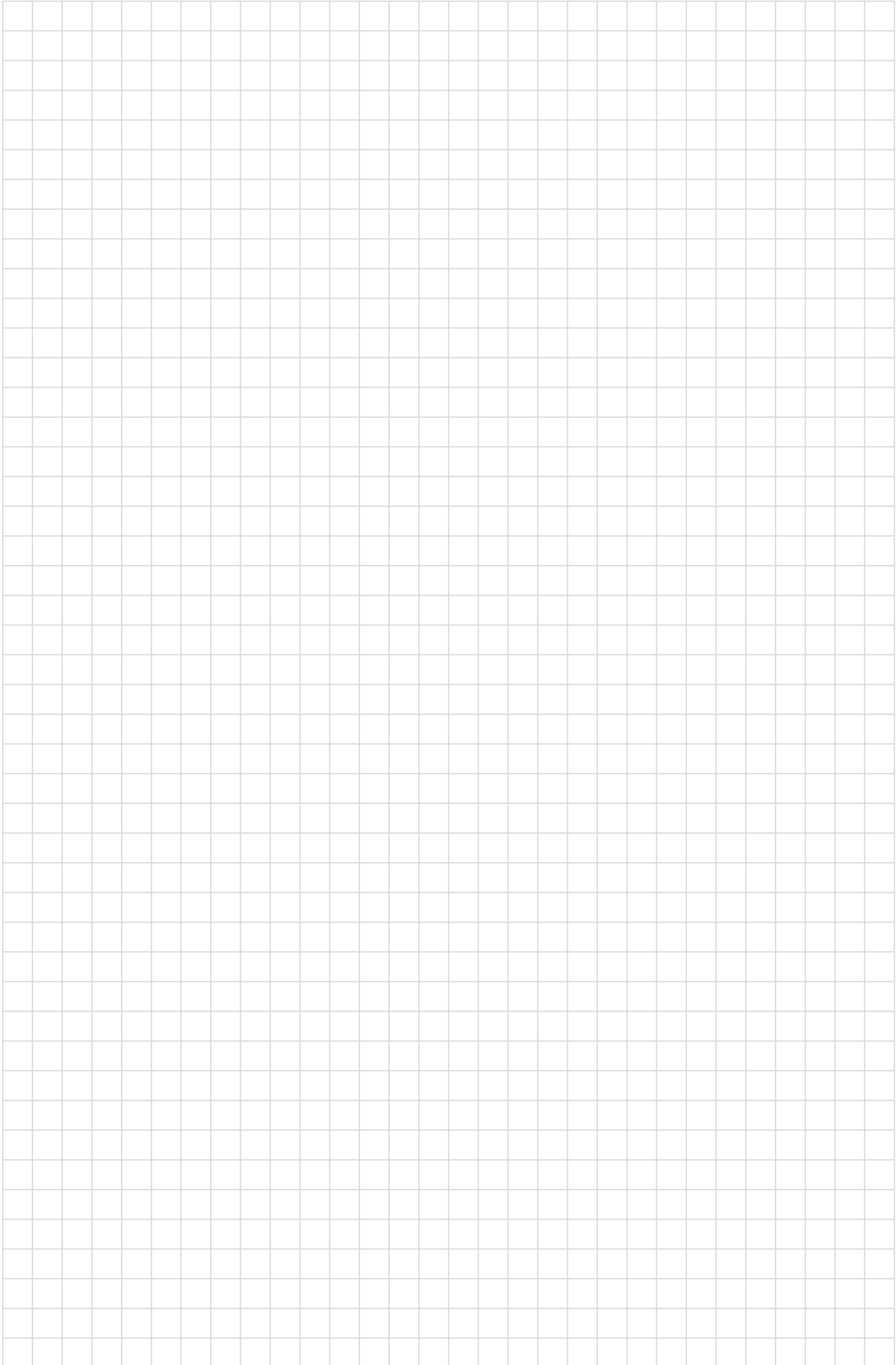
- (e) Describe the Karush-Kuhn-Tucker (KKT) conditions for constrained optimisation. (5 marks)

5

- (f) Show that the Dirichlet distribution given by $\text{Dir}(\mathbf{p}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\prod_{i=1}^d \Gamma(\alpha_i)} \prod_{i=1}^d p_i^{\alpha_i-1}$, where $\mathbf{p} = (p_1, p_2, \dots, p_d)$, $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_d)$ and $\alpha_0 = \sum_{i=1}^d \alpha_i$ is a conjugate prior to the multinomial likelihood $\text{Binom}(\mathbf{k}|n, \mathbf{p}) = n! \prod_{i=1}^d \frac{p_i^{k_i}}{k_i!}$, where \mathbf{k} is a vector of counts (k_1, k_2, \dots, k_d) with $\sum_{i=1}^d k_i = n$. Derive update equations for the parameters $\boldsymbol{\alpha}'$ of the posterior distribution after observing counts \mathbf{k} . (5 marks)

5

Additional space. Do not use unless necessary. Clearly mark corresponding question.



(g) Prove that the set of positive semi-definite matrices is a convex set. (5 marks)

5

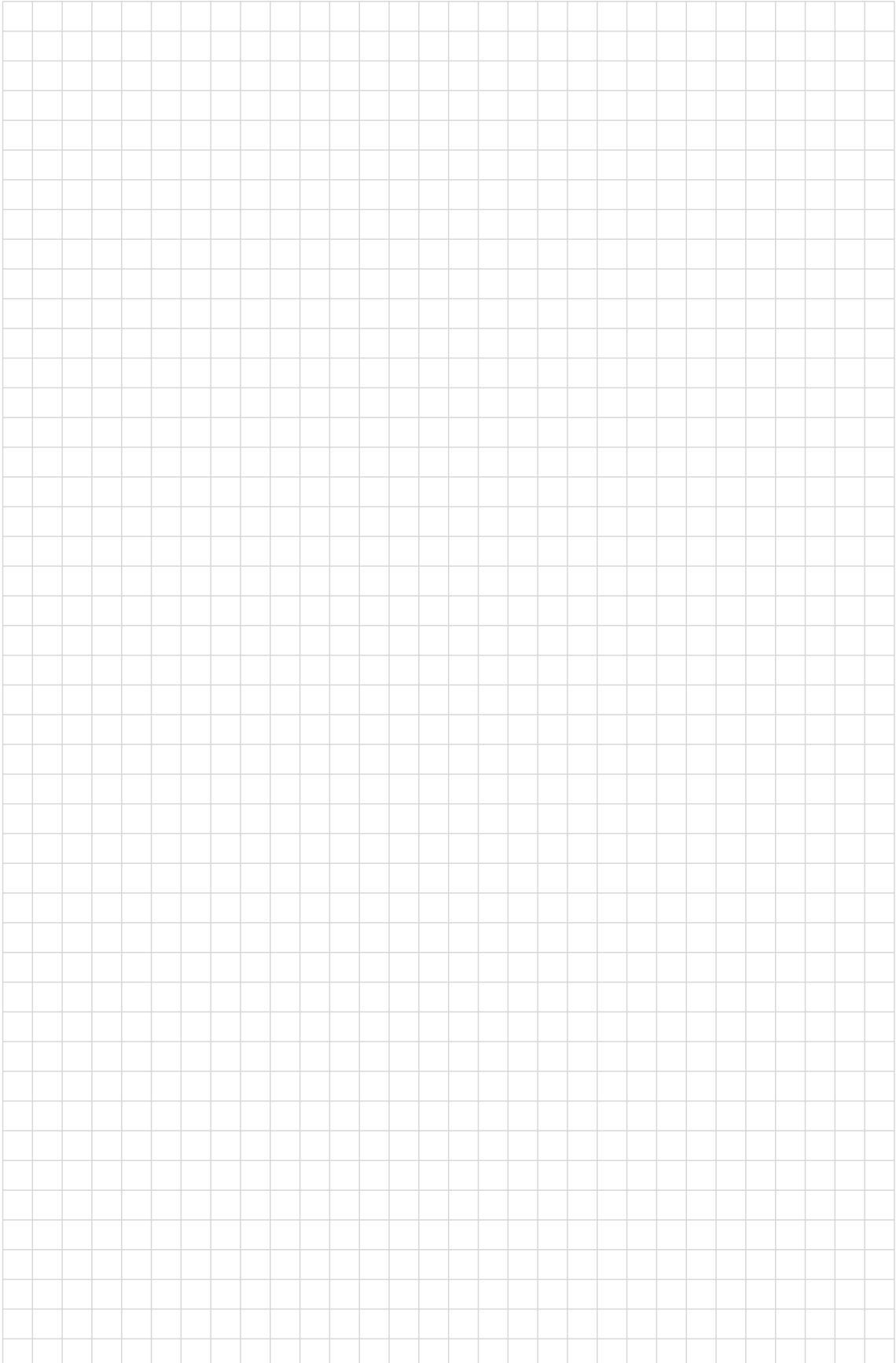
(h) Explain what are the hyper-parameters of a Gaussian process and why they are relatively easy to learn. (5 marks)

5

End of question A1

(a) $\frac{\quad}{5}$	(b) $\frac{\quad}{5}$	(c) $\frac{\quad}{5}$	(d) $\frac{\quad}{5}$	(e) $\frac{\quad}{5}$	(f) $\frac{\quad}{5}$	(g) $\frac{\quad}{5}$	(h) $\frac{\quad}{5}$	Total $\frac{\quad}{40}$
-----------------------	-----------------------	-----------------------	-----------------------	-----------------------	-----------------------	-----------------------	-----------------------	--------------------------

Additional space. Do not use unless necessary. Clearly mark corresponding question.



Section B

Question B 2

(a) If $\{X_i | i = 1, 2, m\}$ is a set of correlated random variables such that

$$\langle X_i \rangle = \mu \quad \langle (X_i - \mu)(X_i - j) \rangle = \begin{cases} \sigma^2 & \text{if } i = j \\ \rho \sigma^2 & \text{if } i \neq j \end{cases}$$

show

$$\left\langle \left(\frac{1}{n} \sum_{i=1}^n X_i - \mu \right)^2 \right\rangle = \rho \sigma^2 + \frac{(1-\rho) \sigma^2}{n}$$

(10 marks)

[illegible]

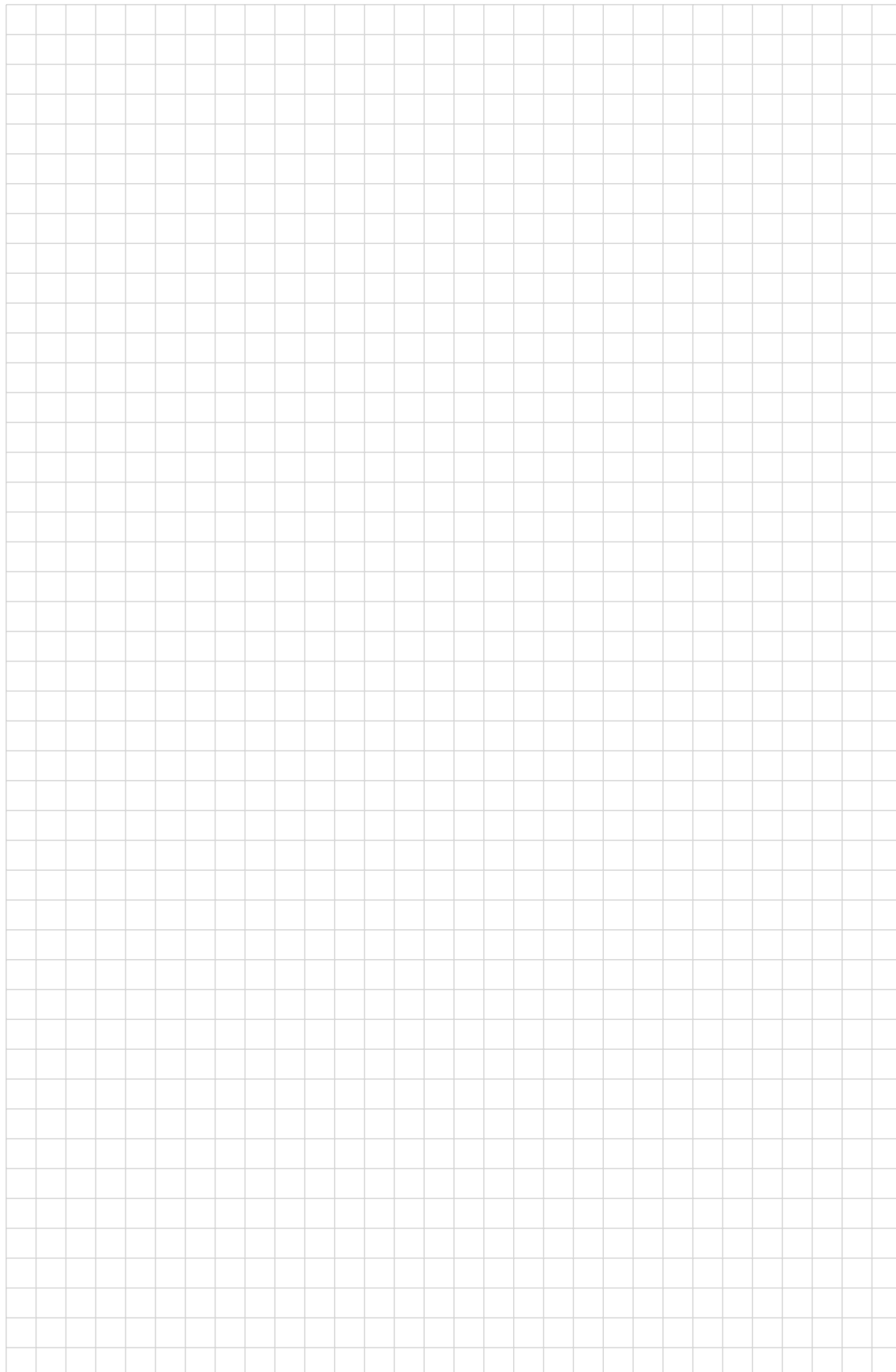
- (b) Using the result derived in part (a) explain why ensembling many machines can reduce the variance in the bias-variance dilemma if the machine predictions are not heavily correlated. Use this to explain the success of random forest. (10 marks)

$\frac{10}{10}$

End of question B2

(a) $\frac{10}{10}$	(b) $\frac{10}{10}$	Total $\frac{20}{20}$
---------------------	---------------------	-----------------------

Additional space. Do not use unless necessary. Clearly mark corresponding question.



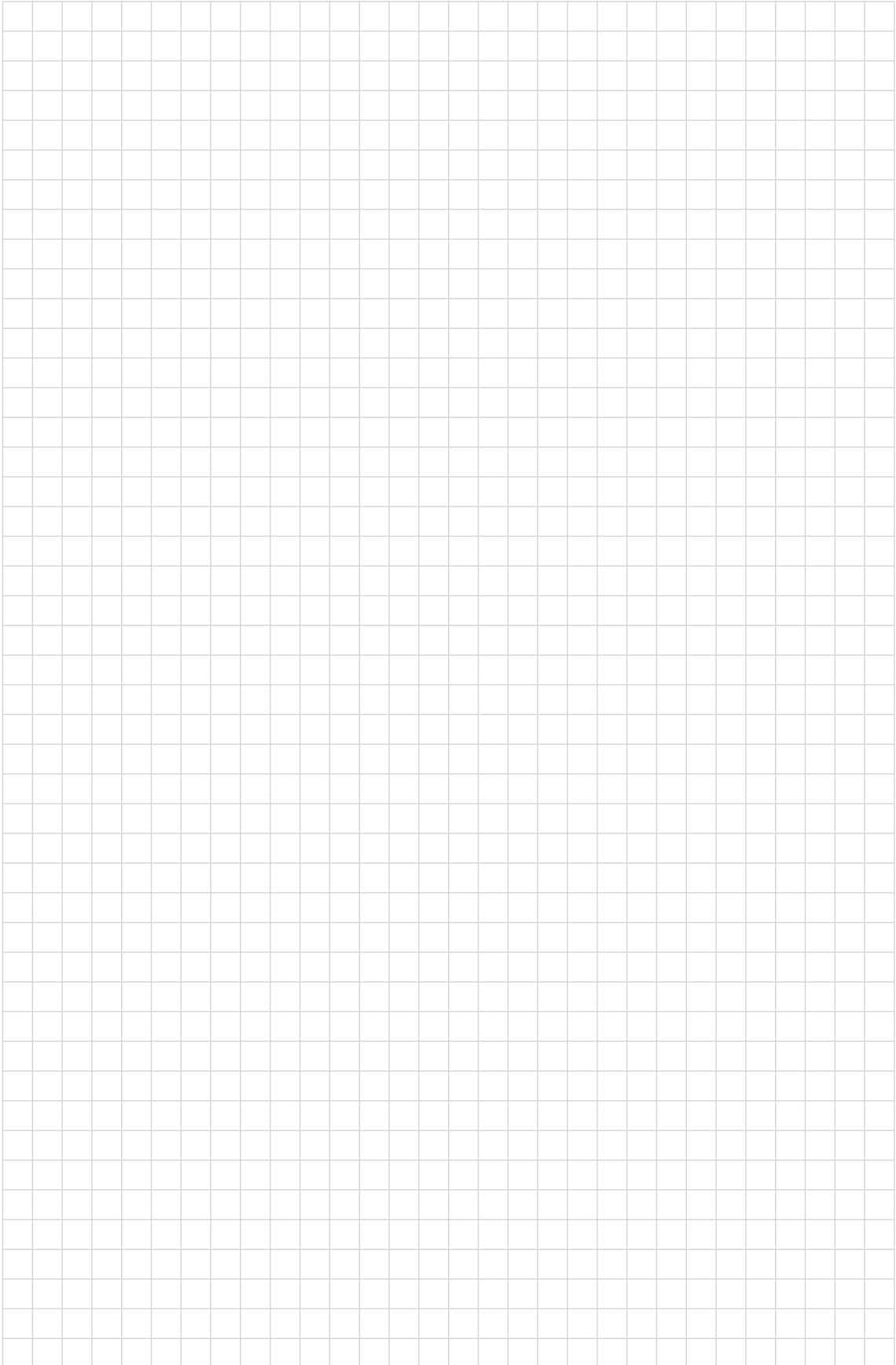
- (b) Use the result that you derive to explain how adding the L_2 regulariser $\eta \|w\|^2$ improves the conditioning of the solution and is likely to improve generalisation.
(5 marks)

End of question B3

5

(a) $\frac{\quad}{15}$	(b) $\frac{\quad}{5}$	Total $\frac{\quad}{20}$
------------------------	-----------------------	--------------------------

Additional space. Do not use unless necessary. Clearly mark corresponding question.



Question B 4

- (a) Write a Lagrangian for the linear programming problem of choosing x to minimise $c^\top x$, subject to the constraints $\mathbf{M}x = b$. Show that the Lagrangian can be rewritten to obtain the dual problem where the roles of the variables x and the Lagrange multipliers are exchanged. Write down the dual problem as a maximisation problem plus a new set of constraints. (5 marks)

[illegible]

- (b) Describe the Wasserstein distance as a linear programming problem and describe the dual problem. Describe how this is used in the Wasserstein GAN. (15 marks)

This image shows a single sheet of white paper with horizontal ruling lines. The lines are evenly spaced and run across the width of the page. There is no text or other markings on the paper.

15

End of question B4

(a) $\frac{\quad}{5}$ (b) $\frac{\quad}{15}$ Total $\frac{\quad}{20}$

Additional space. Do not use unless necessary. Clearly mark corresponding question.

