TRINITY COLLEGE DUBLIN

School of Computer Science and Statistics

Final Assignment 2023-24 CS7CS2 Optimisation for Machine Learning

- **You must do this assignment entirely yourself - you must not discuss or collaborate on the assignment with other students in any way, you must write answers in your own words and write code entirely yourself. If you use any online or other external content in your report you should take care to cite the source (that includes use of code assistants, chat GPT and the like). It is mandatory to complete the declaration that the work is entirely your own and you have not collaborated with anyone - the declaration form is available on Blackboard. All submissions will be checked for plagiarism.**
- **Reports must be typed and submitted as a separate pdf on Blackboard (not as part of a zip file).**
- **Include the source of code written for the assignment as an appendix in your submitted pdf report (the code itself as plain text, not a screenshot, so the plagiarism checker can run on it). Failure to do this will lead to a 50% reduction in your mark. Also include a separate zip file containing the executable code and any data files needed. Programs should be running code written in Python. Keep code brief and clean with meaningful variable names etc.**
- **Important: Your primary aim is to articulate that you understand what you're doing - not just running a program and quoting numbers it outputs. Generally most of the credit is given for the explanation/analysis as opposed to the code/numerical answer.**
- **Reports should typically be about 5 pages, with 10 pages the upper limit (excluding appendix with code).**

Assignment

1. It is thought that the "noise" added to the gradient when using mini-batch SGD acts as a regulariser and so helps prevent overfitting. Your task is to write a short report critically evaluating the use of mini-batch SGD to reduce overfitting and improve generalisation performance.

   Choose a neural network model and dataset to study e.g. a convolutional net for image processing, a transformer for text. The choice is up to you but it's important to justify why your choice is appropriate. Remember that it is overfitting, and its prevention, that is of interest so your choice of model and data should reflect that.

   Be sure to split the data into training, test and validation sets (cross-validation does not measure generalisation behaviour adequately). Also bear in mind the random nature of SGD, which means that you will probably need to do multiple runs and look at both the average behaviour and the fluctuations about the average from run to run.

   Investigate the role of mini-batch size on overfitting (remember that SGD becomes gradient descent the batch size equals the full dataset). Think carefully about the best way to do that. You will need to evaluate the degree of overfitting and then assess how that changes with the mini-batch size. Try and measure the amount of gradient "noise" added by the mini-batch SGD (use multiple SGD runs to do this).

If you hold the mini-batch size constant and add different levels of Gaussian noise to the gradients, how does that affect over-fitting compared to varying the mini-batch size?

Investigate the interplay between step size and mini-batch size. For a constant mini-batch size try using a step size schedule (i.e. reducing the step size as the #epochs grows) to obtain a well-trained model. Compare this with the behaviour when you hold the step size constant and use a mini-batch schedule (i.e. reducing the batch size as the #epochs grows). What about if you vary mini-batch size and step size together?

So far you've used constant-step size (perhaps with a manual schedule). Now investigate the use adaptive approaches such as adam. Does that increase over-fitting? Does it change how best to choose the mini-batch size to control over-fitting?

85 marks: indicative breakdown (i) methodology 30 marks, (ii) evaluation and critical discussion 45 marks, (iii) report organisation and presentation 10 marks.

2. (a) Explain how line search can be used in gradient descent. Give one advantage and one disadvantage. Should line search be used with stochastic gradient descent? [5 marks]

   (b) Describe how a change of variables can be used to enforce a constraint on the decision variables, illustrating with an example. [5 marks]

   (c) Explain how a penalty can be added to the cost function so as to enforce a constraint on the decision variables. Illustrate with a brief example. [5 marks]