L361: MPhil/Part III Project Guidelines

2024-2025

For Part III/MPhil students, L361 comprises four practicals and one final research project. The final practical slot is also used to discuss the early results of the final project. The Part II students receive separate guidelines.

The **final project** will be expected to be a well-presented, full-scale research paper. This hands-on project will be assessed based on the *report*, *source code*, *related documentation*, and a brief 8-minute *pre-recorded talk* summarising key project elements (any slides used are also submitted as part of the project). The final project is open-ended, depending on the exact nature of the research. The contents and format of a scientific work generally include the *motivations* and starting *assumptions* of the research; the *goal* of the research (e.g., improving the performance of a federated model in a cross-silo scenario); a *literature review* contextualizing the previous work and justifying the existence of a research gap, the *methods* applied to fill the research gap (e.g., a new aggregation algorithm), the *experimental setup* used to test the effectiveness of the proposed methods with clear standards and criteria for success (equivalent to the research hypotheses of the lab report); the *results* and their *interpretation* (including sources of error); any *conclusions*.

As the project report is a major form of assessment for the module, it is important that you invest a significant amount of time in writing and refining the report, paying attention to detail and presentation.

All project reports **must** use the NeurIPS template.

1 Contents

For L361, we require that the project report follow this structure concerning both section and overall report length. Marks will be deducted for failing to observe the final project's overall page limits, excluding references and appendices. The recommended lengths for each section provide a lower and upper bound; choosing the upper bound for every section will likely result in you going over the page limit. Thus, care should be given to how the focus is split depending on the nature of your work.

Page Limit The final project has a page limit of **9 pages**, including figures.

1.1 Title; author name/names; date

Two authors are expected at least.

1.2 Final Project: Abstract

The abstract should contain a high-level summary of the context of the work, the identified research gap, the proposed solution, and the solution's effectiveness based on the experimental results. It should be no longer than 140 words.

1.3 Introduction

Frames the report by providing (succinctly) context, motivations, approach, and results. (1-2 paragraphs)

1.4 Background

A literature review of work done around the topic of the final project with particular care towards identifying strengths and weaknesses of previous work. If a new FL method is proposed, the background **must** identify the research gap it is trying to fill and provide evidence for its relevance towards the wider FL field. If the final project consists of a novel combination of previously existing techniques, argumentation must be provided for why their interaction is worthy of exploration. (1-1.5 pages)

1.5 Methods

This section should provide an overview of the method that is being proposed/experimented with, together with relevant formulas and potential pseudocode. (1-1.5 pages)

1.6 Experimental setup and methodology

An exploration of the goals, hypotheses, experimental setup, procedure used in the experiment, and details of any steps taken to mitigate potential error or problems. Clear baselines must be chosen as comparison points (e.g., `FedAvg`, `FedProx`, `FedOPT` etc.) for the tested algorithm and that the choice of the dataset and evaluated architectures be properly justified. **Crucially**, the report **must** attempt to model **data** or **systems** heterogeneity. In the case of data heterogeneity, clear quantification of the nature and degree of heterogeneity should be provided with the distributions of labels/features/client dataset sizes present in the appendix for every tested combination of dataset and partitioning. For systems heterogeneity (if it can be simulated or obtained), details of hardware used (if using real heterogeneous devices) or the assumptions of hardware characteristics simulated should be provided. Reproducibility should be explicitly considered via proper library versioning and seeding, although details for reproducing experiments should be delegated to a GitHub repository

or the appendix. A separate paragraph should be included in the appendix to mention at a high level what steps have been taken to ensure reproducibility, similarly to the ICLR guidelines. (1-2 pages)

1.7 Results and discussion

The results obtained, graphs illustrating those results, and exploration and interpretation of the results, including important artifacts, the validity of the results, and conclusions they lead us to. Tables and figures required to explain the results should be present in all reports; in most cases, graphs will be expected. This is the body of the report. Explicit mention of whether the proposed method/combination of previously existing methods fulfilled its expected goals or validated the experimental hypotheses should be provided. For both successes and failures, the quality of argumentation for the likely cause of the outcome will be used to determine marking. (2-3 pages, including figures)

1.8 Conclusion

A summation of the results and thoughts on potential future directions. (1-2 paragraphs)

1.9 References

References to the most relevant scientific literature. (not counted towards page limit)

1.10 Appendices

Additional material that supplements prior sections - e.g., it might be desirable to reference content such as scripts to perform experiments in explaining material in the body, additional data tables, or more detailed illustrations of an experimental setup. As mentioned above, details on reproducibility and data distributions used should go here. (There is no length limit, but moderation is encouraged when including additional results)

1.11 Recap

Abstract (1 paragraph, 140 words), Intro (1-2 paragraphs), Background (1-1.5 pages), Methods (1-1.5 pages), Experimental setup and methodology (1-2 pages), Results and discussion (2-3 pages), Conclusion (1-2 paragraphs). Choose appropriate section lengths to stay within the 9-page limit.

1.12 Caveats

It is important that both the abstract, introduction, and conclusion describe the hypotheses/goals, experimental approach, and results - which will feel redundant but is the nature of the format.

Appendices should be included only where they improve understanding of the body of the report. Only pages within the page limit of final project will be assessed; if the appendices are too long in terms of text or, in exceptional circumstances, in terms of figures, they will not be read.

2 Style and presentation

Reports must be clearly written, spell-checked, and formatted to make them easy for the reader to follow, including concerning figure labels. Given length limitations, they will, of necessity, be high-level presentations of your experiments and cannot explore every detail — but instead, they should focus on important and interesting results and how they relate to effects observed in the labs or topics discussed in lectures/wider FL literature.

Particular attention should be paid to graphs and tables that will present the results: axes must be labeled, scales should be selected with care to avoid misunderstanding, and if, for example, there are clear artifacts of interest, then an additional graph may be appropriate to explore those in greater detail. All graphs must be described in the text's body and have a suitable (but brief) descriptive caption. It will be important to include error bars or other error information and explain when confidence intervals have been used. Thought should be put into stacking graphs vertically (for comparing results at the same x), horizontally (for comparing y-values), or if two graphs should be merged. Two y-axes may also be appropriate depending on context (e.g., inner-product and cosine similarity).

For the sake of experimental reproducibility, at least three experiments are required to present a median, min, and max result with other methods becoming available (e.g., interquartile range, mean and standard deviation, mean and confidence interval) as the number of experiments increases. If you are unable to run enough experiments to asses error, prefer changing the approach (e.g., move to only single-round analysis and simulate multiple client selection situations) rather than reporting single numbers without any context.

All graphs must be vector-based rather than raster images and must be prepared such that they are clear even if printed in black and white (so use different line styles/fillings). It may be appropriate to use diagramming packages such as tikz, code rendering via the listings package, and additional tools such as matplotlib, **R**, and graphviz to analyse and present results.

Students are cautioned that many of these tools are complex and subtle and, when used incautiously, tend to consume all available time. If you run into difficulties, seek help from the course instructor or one of the teaching assistants - and when in doubt, avoid exciting-sounding features in LaTeX!

3 Assessment

Below the expectations: extremely poor (or incomplete) experimental procedure or writeup that might include an incoherent description of the work, improper experimental design incapable of testing the experimental hypotheses/proposed method, poor experimental practice that leads to incorrect results, failure to discuss potential sources of error, lack of comprehension of the explored/proposed FL method, and/or poor data analysis that draws incorrect conclusions despite clear evidence to the contrary. This marking range will also be used if insufficient original thought is put into the argumentation.

Matching expectations: adequately performed experimental procedure and writeup, but with a few (but not many) of the following problems: (1) the experimental approach will have been roughly right, but failed to avoid potential sources of error, used inadequate procedures to manage variance, failed to pursue an important behavior or effect; (2) the writeup will have drawn reasonable conclusions, but failed to make proper use of statistics to prove the reproducibility of effects or failed to investigate a surprising effect or result; or (3) graphs will present useful results but are unclear or disagree with the experimental analysis.

Exceeding expectations: most or all of the following hold: a superior writing style and clarity; strong experimental procedure and error analysis, in which surprising results or artifacts are adequately illustrated via graphs and explained in the text; the effects of heterogeneity are properly considered, measured, and analyzed concerning useful partitions (e.g., very heterogeneous, heterogeneous, homogeneous); analysis looking at the impacts of a given FL algorithm at a given round is provided together with analysis consider cross-round trends and strong or even new insights into performance are gained.

4 Plagiarism

Excessive similarity of graphs, text, and analysis between non-collaborators may be penalized as plagiarism.

5 Credit

Thanks to Professor Robert N. M. Watson for his lab report guide used in the Advanced Operating Systems course.