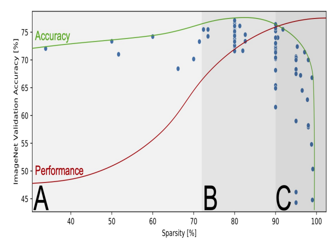
Sparsity, Low-rank, and other low-dimensional models in data science and machine learning.

Introduction

Machine learning, and other data science settings, are typically modelled using matrices as linear operators. Many of these tasks benefit from acting in very high dimensional setting with potentially millions (or many more) parameters, but to avoid computational challenges they can achieve the desired goal by acting on a low-dimensional subspace. For instance, deep learning repeatedly applies an entry wise non-linear activation to an affine transform, $h_{k+1} = \phi(A_k h_k + b)$ and seeks to learn the parameters in A_k for a desired task such as image classification. In this deep learning setting having the dimensions of A_k being large is beneficial, but not all entries in A_k need to be non-zero or independent. Exemplar low-dimensional subspaces for A_k are A_k having few non-zero entries or having a low rank.

The extremely large number of parameters typical in deep learning models is also raises fundamental questions about the model overfitting the data. Reducing the number of parameters in the network can have beneficial improvements in the networks performance on unseen data, generating even improving the accuracy of the network for moderate sparsity.



From "Sparsity in DL" by Hoefler et al.

In a very different setting, compressed sensing tells us that an unknown object X that is known to be low-dimensional (such as sparse or low-rank) can be measured more efficiently than by measuring each of entries. Modelling a matrix as the sum of a low-rank matrix plus a sparse matrix can allow allow for robust principal component analysis or segmentation of global features from moving objects.

(a) Original frames (b) Low-rank $\hat{L}$ (c) Sparse $\hat{S}$
From "Robust PCA" by Candes et al.

Project

The project starts by selecting a topic in data science and machine learning where sparsity, low-rank, or other low-dimensional structure will be explored. Exemplar cases include pruned models for deep learning or compressed sensing. Once the topic has been selected we will determine an aspect of the topic to be studied which focuses on the benefits of the low-dimensional structure. Projects typically benefit from, but need not, have computational components.

Prerequisites

There are no prerequisites beyond the core courses in Prelims and Part A. Students would benefit from having taken Part A Numerical Analysis and having a good understanding of linear algebra. Some project directions would benefit from concurrently taking Part B6.2 Optimisation for Data Science.

Reading

References for helping select a topic, and for their associated topics, are:

Pruning Deep Neural Networks from a Sparsity Perspective by Diao et al., https://arxiv.org/abs/2302.05601.

Sparsity in Deep Learning: Pruning and Growth for Efficient Inference and Training in Neural Networks by Hoefler et al., https://arxiv.org/abs/2102.00554.

Improved Projection Learning for Lower Dimensional Feature Maps by Price and Tanner, https://arxiv.org/abs/2210.15170.

SparseGPT: Massive Language Models Can be Accurately Pruned in One-Shot by Frantar and Alistarh, https://arxiv.org/abs/2301.00774.

Exact Matrix Completion via Convex Optimization by Candes and Recht, https://arxiv.org/abs/2301.00774.

Robust Principal Component Analysis? by Candes et al., https://arxiv.org/abs/0912.3599.

A Mathematical Introduction to Compressive Sensing by Foucart and Rauhut,