**Matrix Analysis and Applications (Autumn 2021)** **Final Project**

# 综合课题研究

*Lecturer: Feng Chen* chenfeng@mail.tsinghua.edu.cn

*TA: Tianren Zhang, Yizhou Jiang, Chongkai Gao* zhang-tr19,jyz20,gck20@mails.tsinghua.edu.cn

## 1 基本要求

**A.** 本次大作业采用分组形式进行。每 1 ∼ 2 名同学可以自由组合为一个小组，共同协作讨论完成文献调研、仿真实验、展示 PPT 以及附录文档，并派一名代表将所有相关内容（**PPT、附录文档、仿真程序**等）一起压缩打包 (以 "*姓名 1_ 姓名 2_ 学号 1_ 学号 2*" 形式进行压缩文件命名) 上传网络学堂，同时在 "作业内容" 中再次注明小组成员。其他小组成员也要提交作业，但无需重复上传附件，只需在 "作业内容" 中注明小组成员，并写明由哪位成员负责提交附件。

**B.** 作业成果以展示 PPT 为主，其内容包含作业中各项要求所对应的技术路线、理论推导、算法设计、实现方法、数值实验结果等内容；各项要求间的详略分配可自行安排，整体篇幅以支撑 10 ∼ 15 分钟的课堂展示为宜。附录文档主要包括作业中所涉及，但不宜以 PPT 形式展示的诸细节，如详细推导过程、算法设计思路、实验数据整理、代码自述文件、关键参考文献等，具体形式和内容可根据需要自行安排。

**C.** 展示内容应充分说明问题的解决思路，以及理论或算法的特点，结合自己的理解给出相应结论。技术方案应当兼顾时间复杂度与准确性，对具有创新性的想法和内容，将给予一定加分。仿真实验程序应包含适当的注释以保证可读性。

**D.** 对于超过 1 人完成的作业，展示 PPT 中应在末页注明各成员贡献比例（如 1 : 0.9），并说明每位成员负责的工作内容。没有分工说明或说明不充分的报告将被予以一定程度的扣分。

**E.** 作业应独立完成。一经发现抄袭现象，抄袭与被抄袭者的成绩均以 0 分计。实验代码或仿真程序的主体部分应自己编写，直接调用网上公开的程序代码应在代码文件的相应位置进行注释，并在报告中对出处进行引用。使用过多的公开代码会影响报告得分。

**F.** 作业截止时间为**第 14 周周日（12 月 19 日）23:59**。所有小组需在最终报告截止前提交所有相关资料，没有按时完成的小组将被扣分。所有小组都应当为课堂展示做好准备，我们会根据完成情况选择一部分小组进行课堂展示。

**评分标准：** 根据作业的完成度、工作量、展示度、创新性等综合评分。大作业满分 100，对于在以上任意一个方面有突出表现的作业在评分时会酌情予以加分，但总分不超过 100。

## 2 课题研究内容：CSSP

CSSP（Column Subset Selection Problem）是矩阵近似（matrix approximation）领域的经典问题之一，其建模了一个有约束低秩近似问题，所关注的是如何从一个矩阵的列向量集合中找出一个最有代表性的 $k$ 元子集来作为对该矩阵的压缩表征。CSSP 在机器学习、科学计算、信号处理等诸多领域都有广泛的应用。在 Frobenius 范数下，CSSP 有如下定义：

**CSSP:** 对于给定的矩阵 $\boldsymbol{A} \in \mathbb{R}^{m\times n}$，从其列向量集合 $\{\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_n\}$ 中选择一个 $k$ 元子集 $S$，从而最小化

$$\mathrm{Er}_{\boldsymbol{A}}(S) := \|\boldsymbol{A} - \boldsymbol{P}_S\boldsymbol{A}\|_F^2 .$$

其中 $\|\cdot\|_F$ 是 Frobenius 范数，$\boldsymbol{P}_S$ 是从 $\boldsymbol{A}$ 的列空间到 $\mathrm{span}\,(S)$ 的投影矩阵。

直观来看，CSSP 要求从 $m \times n$ 矩阵 $\boldsymbol{A}$ 中选出 $k$ 列，使其尽可能近似“张成” $\boldsymbol{A}$ 中所有的列。作为组合优化问题，CSSP 显然可以用遍历法得到精确解，但这种方法在高维情况下的时间复杂度过大（$C_n^k$ 量级）；对于 CSSP，文献中有许多适用于高阶矩阵的近似解法，亦存在相当数量的理论研究。

本课题研究内容共有 4 个具体要求：

**要求 1.** 实现通过遍历精确求解 CSSP 问题的算法（针对低维矩阵），并给出相应的列向量集合 $S$，投影矩阵 $\boldsymbol{P}_S$，以及拟合误差 $\mathrm{Er}_{\boldsymbol{A}}(S)$。

**要求 2.** 实现一种近似求解 CSSP 问题的算法（同时适用于低维 + 高维矩阵），并分析该算法的复杂度。

**要求 3.** 分析讨论近似算法导致的精度损失，并在低维矩阵上进行实验验证。

**要求 4.** 自选指标对比 CSSP 和其他的矩阵近似方法（如 SVD）的近似效果，并讨论不同方法的优劣。

其中，编程部分可以基于任何语言（例如：MATLAB，Python，C++，R 等）完成，但必须提供如下接口：支持随机生成/人为指定任意数量、任意尺寸的实矩阵以及人为指定列向量子集大小 $k$，并能输出 *CSSP* 问题的求解结果和误差。

## 初始参考文献

[1] Derezinski, Michal, Rajiv Khanna, and Michael W. Mahoney. “Improved guarantees and a multiple-descent curve for Column Subset Selection and the Nystrom method.” *Advances in Neural Information Processing Systems* 33 (2020).

[2] Deshpande, Amit, et al. “Matrix approximation and projective clustering via volume sampling.” *Theory of Computing* 2.1 (2006): 225-247.

[3] Boutsidis, Christos, Michael W. Mahoney, and Petros Drineas. "An improved approximation algorithm for the column subset selection problem." *Proceedings of the twentieth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics, 2009.

[4] Tropp, Joel A. "Column subset selection, matrix factorization, and eigenvalue optimization." *Proceedings of the twentieth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics, 2009.

[5] Broadbent, Mary E., et al. "Subset selection algorithms: Randomized vs. deterministic." *SIAM undergraduate research online* 3.01 (2010).

[6] Altschuler, Jason, et al. "Greedy column subset selection: New bounds and distributed algorithms." *International conference on machine learning*. PMLR, 2016.

[7] Gu, Ming, and Stanley C. Eisenstat. "Efficient algorithms for computing a strong rank-revealing QR factorization." *SIAM Journal on Scientific Computing* 17.4 (1996): 848-869.