論文 / 著書情報
Article / Book Information

題目(和文)	スパース行列に対する行列分解の統計力学的解析
Title(English)	Statistical mechanics analysis of low-rank matrix factorization for sparse matrices
著者(和文)	野口千尋
Author(English)	Chihiro Noguchi
出典(和文)	学位:博士(理学), 学位授与機関:東京工業大学, 報告番号:甲第11671号, 授与年月日:2020年12月31日, 学位の種別:課程博士, 審査員:渡邊 澄夫,金森 敬文,高安 美佐子,三好 直人,山下 真,樺島 祥介
Citation(English)	Degree:Doctor (Science), Conferring organization: Tokyo Institute of Technology, Report number:甲第11671号, Conferred date:2020/12/31, Degree Type:Course doctor, Examiner:,,,,,
学位種別(和文)	博士論文
Category(English)	Doctoral Thesis
種別(和文)	論文要旨
Type(English)	Summary

（博士課程）
Doctoral Program

論 文 要 旨

THESIS SUMMARY

系・コース： Department of, Graduate major in	数理・計算科学 系 数理・計算科学 コース	申請学位（専攻分野）： Academic Degree Requested	博士 Doctor of （ 理学 ）
学生氏名： Student's Name	野口 千尋	指導教員（主）： Academic Supervisor(main)	渡辺澄夫
		指導教員（副）： Academic Supervisor(sub)	樺島祥介

要旨（英文 800 語程度）
Thesis Summary（approx.800 English Words ）

Low-rank matrix factorization is the basis of a real data analysis in fields such as machine learning and high-dimensional statistics. This is because the real data often contains much smaller amounts of principal information than the original data size. In this thesis, we address issues of low rank matrix factorization utilizing notions and knowledge of statistical mechanics. One direction is development of practically efficient algorithms for matrix factorization. In real-world problems, data are often given as considerably large matrices, and thus the practical algorithms are required to be performable with low computational costs in parallel and/or distributed manners. The other direction is theoretical analysis of possibilities and limitations of matrix factorization. Specifically, we focus on the community detection from the perspective of the effect of the complex structures contained in real data on the algorithm performance.

In the first direction, we focus on matrix completion. It is the task to recover missing entries of the matrix, whose entries are given only partially. As a comprehensive example, we consider a user-movie matrix $Y \in \mathbb{R}^{N\times M}$, where N and M denote the number of users and movies, respectively, and an entry of Y, y_{ij}, denotes rating from user i movie j. Users normally evaluate only a small fraction of movies, and thus most entries of Y are missing. Under the aforementioned types of setting, the primary objective of matrix completion involves predicting missing entries. In matrix completion problem, the observed data is often given with a particularly large size. Therefore, the practical algorithm is required to work fast in parallel and/or distributed manner. For answering such demands, we propose a new algorithm, which we call cavity-based matrix factorization (CBMF) and approximate cavity-based matrix factorization (ACBMF), borrowing an idea from the cavity method of statistical mechanics. Our algorithms outperform existing algorithms, alternating least square (ALS) and stochastic gradient decent (SGD), which are shown by numerical experiments. This outperformance presumably originates from avoiding extra computation and weakening the self-feedback effect via cavity treatment. Compared to the existing algorithms in terms of computational cost, our algorithms are superior to ALS, while they are competitive with SGD. However, it is known that SGD is highly sensitive to the learning rate in its update equations and required its delicate scheduling. By contrast, our algorithms allow us to avoid such troublesome.

In the second direction, we focus on spectral clustering. It is the algorithm to find community structures in a graph. The graph here indicates a basic data structure that consists of nodes and edges. A node represents a data point and an edge represents relationship between a pair of nodes. The goal of the community detection is to identify communities whose nodes are more densely connected in a graph. This study particularly concentrates on the community detection from the perspective of the effect of the complex structures contained in real data on the algorithm performance. Spectral clustering, which is based on the low-rank matrix factorization, is a popular method for community detection. Although this algorithm exhibits excellent performance when graphs are generated from a simple random graph model, complex structures underlying in real-world data often deteriorate the detection performance. It is, therefore, of importance to clarify the mechanism of the deterioration. To this end, we focus particularly on an overlapping structure of communities, and investigate the effect of the overlapping structure on the performance of spectral clustering by employing the replica method. The detectability performance of the spectral clustering is closely related to the eigenvalues and eigenvectors

of a matrix associated with a graph structure, which is called the Laplacian matrix. When a graph contains two clearly identified communities, its isolated eigenvalue has relevant information of the community structure, while the other bulk of eigenvalues emerge from the randomness of the graph. Therefore, by investigating the relation between them, we can evaluate the performance of the spectral clustering. Our analysis evaluates it by focusing on the model parameters related to the size and density of the overlapping structure. As a result, we found that the size of the overlapping structure has a prominent effect of making the isolated eigenvalue smaller. In contrast, increasing the density of the overlapping structure makes the width of bulk of eigenvalues larger, while the isolate eigenvalue remains almost the same.