Highly efficient non-Euclidean optimization and MCMC methods for deep matrix completion

A Data Management Plan created using DMPonline.be

Creators: Susan Ghaderi, Aldona Niemiro-Sznajder, n.n. n.n.

Affiliation: KU Leuven (KUL)

Funder: Fonds voor Wetenschappelijk Onderzoek - Research Foundation Flanders (FWO)

Template: FWO DMP (Flemish Standard DMP)

Principal Investigator: Susan Ghaderi, n.n. n.n.

Data Manager: Aldona Niemiro-Sznajder

Grant number / URL: 12AK924N

ID: 204091

Start date: 01-10-2023

End date: 30-09-2026

Project abstract:

The matrix completion (MC) problem is of interest in many applications, especially in bioinformatics, chemoinformatics, and systems biology. The existing methodologies for dealing with MC problems are based on the linear dependency between the data entries that factorize the matrix into two (or more) low-rank matrices. Moreover, due to the sparsity and scarcity of the entries of the data matrix, MC suffers from unavoidable large uncertainty.

The main goal of this proposal is to design and develop neural network MC algorithms leveraging MCMC sampling techniques, to consider the nonlinear dependency and uncertainty between entries of such incomplete data. As such, this project will handle the deep matrix completion problems with both
optimization and MCMC sampling techniques. On the ground that the deep matrix completion cost function is typically nonsmooth and nonconvex with non-Lipschitz gradients, we will tailor context-specific non-Euclidean optimization algorithms for such highly nonlinear problems. In addition, following the Bayesian paradigm, I will emply non-Euclidean ptoximal smoothing techniques to design gradient-based MCMC methods.
I will finally apply the developed algorithms to single-cell heterogeneity, gene prioritization, and prediction of pharmacological activities problems.

Last modified: 31-01-2024

Highly efficient non-Euclidean optimization and MCMC methods for deep matrix completion
DPIA

DPIA

Have you performed a DPIA for the personal data processing activities for this project?

- No

Highly efficient non-Euclidean optimization and MCMC methods for deep matrix completion
GDPR

GDPR

Have you registered personal data processing activities for this project?

- No

Highly efficient non-Euclidean optimization and MCMC methods for deep matrix completion
Application DMP

Questionnaire

Describe the datatypes (surveys, sequences, manuscripts, objects …) the research will collect and/or generate and /or (re)use. (use up to 700 characters)

- Generate new data
- Reuse existing data

Specify in which way the following provisions are in place in order to preserve the data during and at least 5 years after the end of the research? Motivate your answer. (use up to 700 characters)

1. Designation of responsible person (If already designated, please fill in his/her name.)
2. Storage capacity/repository
    - during the research
    - after the research

What's the reason why you wish to deviate from the principle of preservation of data and of the minimum preservation term of 5 years? (max. 700 characters)

No

Are there issues concerning research data indicated in the ethics questionnaire of this application form? Which specific security measures do those data require? (use up to 700 characters)

No

Which other issues related to the data management are relevant to mention? (use up to 700 characters)

No

Highly efficient non-Euclidean optimization and MCMC methods for deep matrix completion
FWO DMP (Flemish Standard DMP)

1. Research Data Summary

List and describe all datasets or research materials that you plan to generate/collect or reuse during your research project. For each dataset or data type (observational, experimental etc.), provide a short name & description (sufficient for yourself to know what data it is about), indicate whether the data are newly generated/collected or reused, digital or physical, also indicate the type of the data (the kind of content), its technical format (file extension), and an estimate of the upper limit of the volume of the data.

WP1 - Non-Euclidean methods for deep matrix optimization. To develop Non-Euclidean proximal alternating algorithms. To develop Stochastic non-Euclidean proximal algorithms. To develop Fast non-Euclidean line search algorithms. To develop Non-Euclidean proximal trust-region algorithms.

Type of data	Format	Volume	How created
publication	.pdf	1-10MB	the pdf files generated by latex for submission to journals and conferences
Algorithm iterates, and all other data outputs used for plotting	.dat	0-10MB	Collected algorithm iterates on toy example problems using Python
plots	.png,.pdf	0-10MB	plots generated from the iterates using latex (tikz)
codes and demos	.py, .mat	0-1MB	Python, Matlab
datasets used in simulation	.csv	1-5MB	Toy examples, Python packages and publicly available datasets used in experiments

WP2 -Alternating and simultaneous stochastic Bayesian matrix completion. To develop Bayesian (un)adjusted over/under-damped Langevin methods. To develop Bayesian (un)adjusted higher-order Langevin methods. To develop Bayesian (un)adjusted Hamiltonian MCMC methods. To develop Bayesian (un)adjusted MCMC methods with relatively smooth potential.

Type of data	Format	Volume	How created
publication	.pdf	1-10MB	the pdf files generated by latex for submission to journals and conferences
Algorithm iterates, and all other data outputs used for plotting	.dat	0-10MB	Collected algorithm iterates on toy example problems using Python
plots	.png, .pdf	0-10MB	plots generated from the iterates using latex (tikz)
codes and demos	.py, .mat	0-1MB	Python, Matlab
datasets used in simulation	.csv	1-5MB	Toy examples, Python packages and publicly available datasets used in experiments

If you reuse existing data, please specify the source, preferably by using a persistent identifier (e.g. DOI, Handle, URL etc.) per dataset or data type:

1. https://doi.org/10.1016/j.jpba.2021.114218

Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? Describe these issues in the comment section. Please refer to specific datasets or data types when appropriate.

- No

Will you process personal data? If so, briefly describe the kind of personal data you will use in the comment section. Please refer to specific

datasets or data types when appropriate.

- No

Does your work have potential for commercial valorization (e.g. tech transfer, for example spin-offs, commercial exploitation, …)? If so, please comment per dataset or data type where appropriate.

- No

Do existing 3rd party agreements restrict exploitation or dissemination of the data you (re)use (e.g. Material/Data transfer agreements/ research collaboration agreements)? If so, please explain in the comment section to what data they relate and what restrictions are in place.

- No

Are there any other legal issues, such as intellectual property rights and ownership, to be managed related to the data you (re)use? If so, please explain in the comment section to what data they relate and which restrictions will be asserted.

- No

2. Documentation and Metadata

Clearly describe what approach will be followed to capture the accompanying information necessary to keep data understandable and usable, for yourself and others, now and in the future (e.g., in terms of documentation levels and types required, procedures used, Electronic Lab Notebooks, README.txt files, Codebook.tsv etc. where this information is recorded).

1- All the codes will involve standard naming convention in the relevant programming language (eg. Julia and Python). For example, there will be a readme.txt file, project.toml and licsence files as required by standard package registration tools. All the packages will have an scr forlder (the main code), a test folder (unit tests), demos folder, and documentation.

2- The datasets used will consist of publicly available datasets that are typically in the .csv format. There are readme files describing the number of features and number of data pointsfor each dataset. When a part of a given dataset is used, the truncated or preprocessed dataset will be accompanied by a readme.txt file explaining exactly where the data is taken from and how to load it.

Will a metadata standard be used to make it easier to find and reuse the data? If so, please specify (where appropriate per dataset or data type) which metadata standard will be used. If not, please specify (where appropriate per dataset or data type) which metadata will be created to make the data easier to find and reuse.

- Yes

As we are mainly dealing with data from numerical simulations, we will use a dedicated metadata standard for this purpose such as the EngMeta standard.

3. Data storage & back-up during the research project

Where will the data be stored?

1. A time-stamped version of the data will be stored on the STADIUS DATASET Server, which is our research unit's central storage server. Copies can be made and kept on personal devices.
2. To allow for efficient collaboration with researchers from other research groups, the data will additionally be stored on GitHub (subject to Github regulations).

How will the data be backed up?

All data on the STADIUS DATASET Server are backed up daily and replicated to an off-site storage system housed in the ICTS data center.

Is there currently sufficient storage & backup capacity during the project? If yes, specify concisely.
If no or insufficient storage or backup capacities are available, then explain how this will be taken care of.

- Yes

The STADIUS DATASET Server has a total capacity of 14.88 TB. The capacity of the dataset server is monitored daily by the ESAT system admins.

How will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?

The access of the data on the STADIUS DATASET Server is regulated by an access control list (ACL) that grants:
- read-write access to the project owner and the FWO fellow
- read-only access to specific users
The ACL is managed by the project owner (Panagiotis Patrinos). Client computers can access the data using:
- SMB2 (or higher) from specific IP ranges
- NFSv4 from specific (IT managed) systems

What are the expected costs for data storage and backup during the research project? How will these costs be covered?

All data storage and back up will be performed on the STADIUS DATASET Server in pre-existing storage facilities of the Department of Electrical Engineering (ESAT) without the need of purchasing new infrastructures.

4. Data preservation after the end of the research project

Which data will be retained for at least five years (or longer, in agreement with other retention policies that are applicable) after the end of the project? In case some data cannot be preserved, clearly state the reasons for this (e.g. legal or contractual restrictions, storage/budget issues, institutional policies...).

All research data detailed previously will be retained for at least 5 years, conform the FWO data preservation policy.

Where will these data be archived (stored and curated for the long-term)?

The data will be stored on the university's central servers (with automatic back-up procedures) for at least 10 years, conform the KU Leuven RDM policy.

What are the expected costs for data preservation during the expected retention period? How will these costs be covered?

All data hosted on the STADIUS DATASET Server will be stored for a long term in pre-existing storage facilities of the Department of Electrical Engineering (ESAT) without the need of purchasing new infrastructures. Similarly, the data hosted on GitHub will be stored for a long term.

5. Data sharing and reuse

Will the data (or part of the data) be made available for reuse after/during the project? In the comment section please explain per dataset or data type which data will be made available.

- Yes, in an Open Access repository

If access is restricted, please specify who will be able to access the data and under what conditions.

1. The research articles submitted to journals and conference proceedings will be made publicly available on ArXiv e-prints archive under the perpetual, non-exclusive license. Therefore, it will be available to anyone for any purpose, provided that they give appropriate credit to the creators.
2. The source code will be released publicly on GitHub under the MIT license, along with the corresponding datasets. Hence, it will be available for everyone to use, change, and distribute the

software, only requiring preservation of copyright and license notices.

Are there any factors that restrict or prevent the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)? Please explain in the comment section per dataset or data type where appropriate.

- No

Where will the data be made available? If already known, please provide a repository per dataset or data type.

- In an Open Access repository

1. The research articles submitted to journals and conference proceedings will be made publicly available on ArXiv e-prints archive.
2. The source code will be released publicly on GitHub under the MIT license, along with the corresponding datasets.

When will the data be made available?

- Upon publication of the research results

Upon publication of new research results, the corresponding research article will be made publicly available on ArXiv and the corresponding source code and datasets will be released publicly on GitHub

Which data usage licenses are you going to provide? If none, please explain why.

No

Do you intend to add a PID/DOI/accession number to your dataset(s)? If already available, you have the option to provide it in the comment section.

- No

What are the expected costs for data sharing? How will these costs be covered?

Since free services like GitHub and ArXiv are used to distribute the research data, there are no expected costs.

6. Responsibilities

Who will manage data documentation and metadata during the research project?

The FWO Fellow will be responsible for the data documentation and the metadata.

Who will manage data storage and backup during the research project?

he FWO Fellow will be responsible for the data storage. The system administrators and data manager of the research division are responsible for the back up during and after the project.

Who will manage data preservation and sharing?

The FWO Fellow will be responsible for ensuring data preservation and reuse.

Who will update and implement this DMP?

The project owner (Yves Moreau) bears the end responsibility of updating & implementing this DMP