AFRL-RI-RS-TR-2021-146

COMPOSABLE ROBUST STRUCTURED DATA INFERENCE

CORNELL UNIVERSITY

SEPTEMBER 2021

FINAL TECHNICAL REPORT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

STINFO COPY

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE

■ **AIR FORCE MATERIEL COMMAND** ■ **UNITED STATES AIR FORCE** ■ **ROME, NY 13441**

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09. This report is available to the general public, including foreign nations. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RI-RS-TR-2021-146 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE CHIEF ENGINEER:

/ S /
MICHAEL J. MANNO
Work Unit Manager

/ S /
SCOTT D. PATRICK
Deputy Chief,
Intelligence Systems Division
Information Directorate

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
SEPTEMBER 2021	FINAL TECHNICAL REPORT	MAR 2017 – MAR 2021

4. TITLE AND SUBTITLE

COMPOSABLE ROBUST STRUCTURED DATA INFERENCE

5a. CONTRACT NUMBER
FA8750-17-2-0101

5b. GRANT NUMBER
N/A

5c. PROGRAM ELEMENT NUMBER
62702E

6. AUTHOR(S)

Madeleine Udell

5d. PROJECT NUMBER
D3ME

5e. TASK NUMBER
00

5f. WORK UNIT NUMBER
02

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Cornell University
136 Hoy Road
Ithaca NY 14853

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Research Laboratory/RIED
525 Brooks Road
Rome NY 13441-4505

DARPO/I2O
675 North Randolph St
Arlington VA 22203-2114

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/RI

11. SPONSOR/MONITOR'S REPORT NUMBER
AFRL-RI-RS-TR-2021-146

12. DISTRIBUTION AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited. This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Messy data — heterogeneous values, missing entries, and large errors — presents a major obstacle to automated data-driven discovery of models. Data cleaning is the first step in any data processing pipeline, and has serious consequences for the results of any subsequent analysis. Yet this step is generally performed using ad-hoc methods. This effort seeks to cleanse the data set, and build a structured data interface to reduce noise from data sets, to deliver a production of clean data sets, and leverage model selection and automated techniques.

15. SUBJECT TERMS

Data Modeling Primitives, automated data model selection, machine learning.

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			**MICHAEL J. MANNO**
U	U	U	UU	20	19b. TELEPHONE NUMBER (*Include area code*) **N/A**

Table of Contents

1 Summary

Messy data — heterogeneous values, missing entries, and large errors — presents a major obstacle to automated data-driven discovery of models. Data cleaning is the first step in any data processing pipeline, and has serious consequences for the results of any subsequent analysis. Yet this step is generally performed using ad-hoc methods.

The simplest approach to coping with missing data is also the most common: many researchers simply ignore any datum with a missing element. In settings where most data is present, this practice results in decreased statistical power; in settings where most data is missing, this practice is disastrous, and renders the data useless. When the data is numeric, another common strategy is to impute with the column or row mean. However, this approach results in nonsense when values are Boolean, ordinal, or nominal: what is the mean of "apples" and "oranges"? Methods for inference in general data tables rely either on highly stylized models or on bespoke algorithms for particular domains. Furthermore, existing algorithms either scale poorly to large problems or are not provably globally convergent on general data sets. No existing approach is available to exploit network or spatiotemporal structure to infer missing heterogeneous data.

Cleaning data by learning from data. In this project, we developed composable modeling primitives for robust data inference by exploiting structure in the data set. These primitives for structured data inference accept data with noisy, uncertain, or missing values, and produce clean, complete data sets. Thus, this primitive automatically extend any other modeling primitive to work on highly incomplete, noisy data. We also developed a novel AutoML idea to use primitives for missing value imputation to predict the quality of (and recommend) machine learning methods for new datasets.

2 Introduction

Big datasets are everywhere: in science, in health, in commerce, and in government, data is becoming easier and cheaper to collect. Yet extracting value from this data is a challenge. Every step in the standard machine learning pipeline requires human intervention, and often substantial computational costs: cleaning the data, identifying useful features, choosing a machine learning model, optimizing the model, and validating the model. Meanwhile the steps that transcend the pipeline are critical and difficult to reason about: collecting the right data, interpreting the model, and understanding how the model might change the world for the better or worse.

In this project, we developed new tools to accelerate and automate optimization and machine learning and new frameworks to help researchers understand their data, choose measurements, interpret their model, and understand potential benefits or harms of their

model. To achieve these goals, we sought to discover hidden mathematical structure in the data, algorithms, and procedures that humans use to make their decisions. By learning and exploiting these structures, we can improve, accelerate, and even automate machine learning (including data cleaning, feature extraction, and algorithm selection) and optimization. Automation frees humans from data cleaning and parameter tuning to concentrate on the important questions that algorithms can never answer: are we solving the right problems, and do we have the right data?

Low-dimensional structure can provide the key to automation and speed. The insight that undergirds this research is the observation that measurements of a complex object, suchas a patient in a hospital, respondent on a survey, or even a machine learning dataset, can be well described as simple functions of an underlying low-dimensional latent vector. When wecan identify these low-dimensional latent vectors and these simple measurement functions,we can de-noise and impute entries in messy datasets, reduce the dimensionality of feature vectors, recommend better machine learning algorithms, and speed up optimization and model validation.

On this project, we have developed methods to extract low-dimensional structure from big, messy datasets and to use low-dimensional structure to accelerate large scale optimization and automate machine learning.

Low dimensional structure is everywhere! One continuing theme in this research isthe importance of low-dimensional structure for learning from large and sparse data. We've developed several scalable, parallel, distributed, and low-memory methods to find low rank approximations with provable guarantees (1; 2; 3), that handle matrices with billions of entries on a laptop. These methods can solve problems from a broad range of applications, including medicine (4), social science survey data (5), revenue management (6; 7; 8), environmental monitoring (9; 10; 11), electronics (12), and machine learning (13; 14; 15), whichwe discuss in more detail below. We developed variants of low rank models that work with very weak information, such as assortment choices (6; 7), and that work for sparse data (16).

3 Methods Assumptions, and Procedures.

Why low rank? In each of these applications, we found low rank models performed well — indeed, suspiciously well — for filling in missing data (imputation). Why do low rank models work so well? In (17), we prove that under a very general class of generative model, the rank of an -approximation to an $m \times n$ matrix grows as $O(\log(m+n)/\epsilon^2)$. Hence large enough, square-ish ($m \sim n$) datasets have low rank relative to the trivial bound $\min(m, n)$.

Beyond low rank. The bound on the epsilon-rank of a matrix makes one important, if straightforward, point: while many square matrices are approximately low rank, rectangular matrices almost never are. A long, skinny matrix (even with missing entries) presents much more information about each column than about each row. Low rank factorizations cannot exploit this structure, since row and column ranks are the same. Instead, we must develop

more sophisticated models that allow us to learn more about columns (for which we have much information) than about the rows (about which we know little). A resulting research theme is the search for nonlinear structures that explain real data well, such as polynomial models (18; 19) and copula models (20; 21; 22), and developing algorithms to handle huge, streaming data with these methods, with provable performance guarantees.

Automated machine learning. Algorithm selection and hyperparameter tuning are two of the most frustrating tasks in machine learning. Automated machine learning (AutoML) seeks to automate these tasks to enable widespread use of machine learning by non-experts. The lab's work in this area uses low-dimensional structure to accelerate AutoML. Our first such system, Oboe, views AutoML as a recommender system problem: for each new dataset, we must recommend a machine learning model (14) or pipeline (15). Oboe forms a matrix (or tensor) of the cross-validated errors of a large number of supervised learning models (algorithms together with hyperparameters) on a large number of datasets, and fits a low rank model to learn the low-dimensional representations for the models and datasets thatbest predict the cross-validated errors, among all bilinear models. To find promising models for a new dataset, Oboe runs a set of fast but informative algorithms on the new datasetand uses their cross-validated errors to infer the feature vector for the new dataset. Despite its simplicity, Oboe is among the top AutoML approaches.

We've also worked on other fresh ideas for AutoML: using text descriptions of the dataset and algorithm (23) and a graph neural network (24) to improve algorithm recommendations. The resulting AutoML system is competitive with other state-of-the-art AutoML systems, and recommends a pipeline in ($< 1s$); most others require an order of magnitude more time to return a sensible answer.

Experiment design. Even as data becomes abundant, good data is still not cheap: consider medicine (expensive lab tests or invasive procedures); biology (lab experiments); scientific computing (supercomputer time for simulations); or surveys (more questions). Which data is worth collecting? AutoML provides an ideal testbed to develop new methods for information-directed sampling: which information (here, pipelines evaluated on a dataset) is most important to ensure accurate predictions and sensible decisions? In (14; 15; 25),we develop simple methods to guide information acquisition using classic ideas from experiment design; while (26) develops tools for tensor recovery under (unknown) biased samplingpatterns.

Sketching and matrix approximation. Low rank matrix approximations form a fundamental primitive for an exceedingly wide variety of algorithms. Streaming matrix approximations go further: they can approximate a matrix so large it cannot be stored in memory. Instead, typically these methods collect information about a matrix by computing a random linear function of the matrix, often called a random sketch. The sketch is then used to recover an approximate factorization of the matrix. We've designed several sketching methods

for low rank matrix (27; 28; 9) and tensor (10) approximation, and more efficient sketches based on tensor products (29).

One advantage of the sketching approach is that the total memory required is proportional to the size of the final factorization, rather than the original matrix. This property has important implications for many basic computational problems, including cross-validation (30), dynamic optimization (31), semidefinite programming (32; 33), and solving linear systems (ongoing work).

Scalable semidefinite programming. Semidefinite programming (SDP) is a powerful framework from convex optimization that has potential for data science applications — if algorithms for SDP could handle large scale data! This work has developed new provably correct algorithms for large SDP by economizing on both the storage and the arithmetic costs. We have developed two different approaches to this problem: one based on sketching (32; 34; 33), and the other on complementarity (35). Numerical evidence shows that these methods are effective for a range of applications, including relaxations of MaxCut, abstract phase retrieval, and quadratic assignment, and can handle SDP instances where the matrix variable has over 10^{13} entries. Moreover, the methods handle nearly every useful SDP: the conditions required for these methods to succeed are generic (36) and are necessary for any SDP solver to yield a useful solution given noisy problem data (37).

4 Results

This effort, over the course of the D3M program, provided research and development for incomplete or missing data in tabular datasets, by researching the expectation maximization methods for missing entries in a matrix, with an arbitrary marginal. Using automated machine learning, the marginal distribution, and validating the use of low rank models for missing value imputation, was learned. The team contributed this to the D3M program with the Pyglrm_d3m python package for modeling and fitting generalized low rank models (GLRMs). The primitive was made available for other performers to use, and was compatible with D3M's primitive interface.

We also performed experiments in machine learning, large scale low rank optimization, and tensor factorization, to address datasets not in tabular form.

In addition to development, over the course of this project, we produced, 10 journal publications, 8 refereed conference proceedings, 5 refereedworkshop papers, 29 presentations, and 7 open-source software packages.

5 Conclusions

This work demonstrates the promise of collaborative filtering approaches to AutoML. However, there is much more left to do. Future work is needed to adapt Oboe to different loss metrics, budget types, sparsely observed error matrices, and a wider range of machine learning algorithms. Adapting a collaborative filtering approach to search for good machine learning pipelines, rather than individual algorithms, presents a more substantial challenge. Development of new algorithms that do not require estimating the marginal, and understand when these methods should outperform as a function of data sparsity, and scaling these methods to larger datasets (especially, with more columns), constitutes important future work.

We also hope to see more approaches to the challenge of choosing hyper-hyperparameter settings subject to limited computation and data: meta-learning is generally data(set)-constrained. With continuing efforts by the AutoML community, we look forward to a world in which domain experts seeking to use machine learning can focus on data quality and problem formulation, rather than on tasks — such as algorithm selection and hyperparameter tuning — which are suitable for automation.

A Publications and presentations

In the pipeline

6. J. Fan, L. Ding, C. Yang, and M. Udell. Low-rank tensor recovery with euclidean- norm-induced schatten-p quasi-norm regularization. *Submitted*, 2020, 2012.03436

5. L. Ding and M. Udell. A strict complementarity approach to error bound and sensitivityof solution of conic programs. *Submitted*, 2020, 2012.00183

4. L. Ding, J. Fan, and M. Udell. *k*FW: A Frank-Wolfe style algorithm with stronger subproblem oracles. *Submitted*, 2020, 2006.16142

3. I. Drori, L. Liu, Q. Ma, J. Deykin, B. Kates, and M. Udell. Real-time AutoML. *Submitted*, 2020

2. C. Yang, J. Fan, Z. Wu, and M. Udell. Efficient AutoML pipeline search with matrix and tensor factorization, 2020, 2006.04216

1. L. Ding and M. Udell. On the regularity and conditioning of low rank semidefinite programs. *Submitted to SIOPT*, 2020, 2002.10673

Refereed Journal Articles

10. L. Ding, A. Yurtsever, V. Cevher, J. A. Tropp, and M. Udell. An optimal-storage approach to semidefinite programming using approximate complementarity. *Major Revision at SIOPT*, 2019, 1902.03373

9. J. Fan, C. Yang, and M. Udell. Robust non-linear matrix factorization for dictionary learning, denoising, and clustering. *Major Revision at IEEE Trans. Signal Processing (TSP)*, 2020, 2005.01317

8. R. Muthukumar, D. Kouri, and M. Udell. Randomized sketching algorithms for low memory dynamic optimization. *Under revision at SIOPT*, 2019

7. A. Yurtsever, J. A. Tropp, O. Fercoq, M. Udell, and V. Cevher. Scalable semidefinite programming. *Accepted at SIAM Journal on Mathematics of Data Science (SIMODS)*, 2019, 1912.02949

6. Y. Sun, Y. Guo, C. Luo, J. A. Tropp, and M. Udell. Low-rank Tucker approximation of a tensor from streaming data. *SIAM Journal on Mathematics of Data Science (SIMODS)*, 2020, 1904.10955. J. A. Tropp, A. Yurtsever, M. Udell, and V. Cevher. Streaming low-rank matrix approximation with an application to scientific simulation. *SIAM Scientific Computing (SISC)*, 41(4):A2430–A2463, 2019, 1902.08651

4. M. Udell and O. Toole. Optimal design of efficient rooftop photovoltaic arrays. *IN- FORMS Journal on Applied Analytics (Interfaces)*, 49(4):281–294, 2019

3. M. Udell and A. Townsend. Why are big data matrices approximately low rank? *SIAM Journal on Mathematics of Data Science (SIMODS)*, 1(1):144–160, 2019, 1705.07474

2. N. Kallus and M. Udell. Dynamic assortment personalization in high dimensions. *Operations Research*, 2019, 1610.05604

1. J. A. Tropp, A. Yurtsever, M. Udell, and V. Cevher. Practical sketching algorithms for low-rank matrix approximation. *SIAM Journal of Matrix Analysis and Applications (SIMAX)*, 38(4):1454–1485, 2017, 1609.00048

Refereed Conference Proceedings

8. J. Fan, L. Ding, Y. Chen, and M. Udell. Factor group-sparse regularization for efficient low-rank matrix recovery. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pages 5105–5115, 2019, 1911.05774

7. C. Yang, Y. Akimoto, D. W. Kim, and M. Udell. OBOE: Collaborative filtering for AutoML model selection. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, volume 25, pages 1173–1183. ACM, 2019, 1808.03233

6. J. Fan and M. Udell. Online high-rank matrix completion. In *Computer Vision and Pattern Recognition (CVPR)*, pages 8690–8698, 2019

5. J. Chen, N. Kallus, X. Mao, G. Svacha, and M. Udell. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *FAT*: Conference on Fairness, Accountability, and Transparency*, pages 339–348, 2019, 1811.11154

4. S. Zhou, S. Gupta, and M. Udell. Limited memory Kelley's method converges for composite convex and submodular objectives. In *Advances in Neural Information Processing Systems*, 2018, 1807.07531

3. N. Kallus, X. Mao, and M. Udell. Causal inference with noisy and missing covariates via matrix factorization. In *Advances in Neural Information Processing Systems*, 2018, 1806.00811
2. J. A. Tropp, A. Yurtsever, M. Udell, and V. Cevher. Fixed-rank approximation of a positive-semidefinite matrix from streaming data. In *Advances in Neural Information Processing Systems*, 2017, 1706.05736

1. A. Yurtsever, M. Udell, J. A. Tropp, and V. Cevher. Sketchy decisions: Convex low-rank matrix optimization with optimal storage. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1188–1196, 2017, 1702.06838

Refereed Workshops

5. I. Drori, L. Liu, S. Koorathota, N. Yi, J. Li, A. Moretti, J. Freire, and M. Udell. AutoML using metadata language embeddings. In *NeurIPS Workshop on Meta-Learning*, 2019, 1910.03698

4. Y. Zhang, K. Song, Y. Sun, S. Tan, and M. Udell. "Why should you trust my explanation?" understanding uncertainty in LIME explanations. In *ICML Workshop AI for Social Good*, 2019, 1904.12991

3. Y. Sun, Y. Guo, J. A. Tropp, and M. Udell. Tensor random projection for low memory dimension reduction. In *NeurIPS Workshop on Relational Representation Learning*, 2018

2. C. Yang, Y. Akimoto, D. W. Kim, and M. Udell. OBOE: Collaborative filtering for AutoML initialization (workshop version). *NeurIPS Workshop on Automated Machine Learning*, 2018, 1808.03233

1. M. Paradkar and M. Udell. Graph-regularized generalized low rank models. In *CVPR Workshop on Tensor Methods in Computer Vision*, 2017

Presentations

Applied Math Seminar, Princeton *2020*
Scalable Semidefinite Programming

Science on Tap, Ithaca, NY *2020*
Filling in Missing Data: Elections,_____, Healthcare.

Low-rank models winter school, Villars-Sur-Ollon, Switzerland *2020*
Low Rank Models for Missing Data and Optimization

Statistics and Computation, Alan Turing Institute, London *2020*
Big Data is Low Rank

Reunion Conference on Foundations of Data Science, Simons Institute *2019*
Missing Value Imputation for Mixed Data Through Gaussian Copula

NeurIPS, Vancouver
Factor Group-Sparse Regularization for Efficient Low-Rank Matrix Recovery *2019*

INFORMS, Seattle *2019* *2019*
Low Rank Tucker Approximation of a Tensor from Streaming Data

Knowledge Discovery and Data Mining (KDD), Anchorage *2019* *2019*
Oboe: Collaborative Filtering for AutoML Initialization

JuliaCon, Baltimore *2019* *2019*
Keynote: Big Data is Low Rank using LowRankModels

Applied Math Seminar, UC Boulder *2019* *2019*
Optimal-Storage Semidefinite Programming using Approximate Complementarity

Learning for Dynamics and Control (L4DC), MIT *2019* *2019*
Oboe: Collaborative Filtering for AutoML Initialization (poster)

Machine Learning for Health (ML4H), Vector Institute, Toronto *2019* *2019*
Representation Learning, Patient Similarity, and Subtyping

Low Rank Optimization Workshop, Leipzig MPI for Mathematics in the Sciences *2019*
Low Rank Tucker Approximation of a Tensor from Streaming Data

Optimization and Statistical Learning, Les Houches *2019*
Optimal-Storage Semidefinite Programming using Approximate Complementarity

Women and Mathematics (WAM) Ambassador Program, Cornell University *2019*
Filling in Missing Data: Elections, Healthcare

CME 300, Stanford *2019*
Big Data is Low Rank

Women in Data Science, Stanford *2019*
Plenary: Big Data is Low Rank 100,000 conference attendees worldwide!

Johns Hopkins AMS seminar, Baltimore *2019*
Big Data is Low Rank

CAM Colloquium, Cornell University *2019*
Low Memory Convex Optimization

NeurIPS workshop on AI in financial services, Montreal *2018*
Moderated Industry Panel

NeurIPS workshop on AI in financial services, Montreal *2018*
Fairness under Unawareness

NeurIPS spotlight talk, Montreal *2018*
Limited Memory Kelley's Method Converges for Composite Convex and Submodular Optimization

Rutgers Optimization Seminar, New Brunswick *2018*
Low Memory Convex Optimization

Princeton Optimization Seminar, Princeton *2018 2018*
Low Memory Convex Optimization

UC Davis Mathematics of Data and Decisions Seminar, Davis *2018 2018*
Big Data is Low Rank

Georgia Tech OR Colloquium, Atlanta *2018 2018*
Big Data is Low Rank

Stanford Linear Algebra and Optimization Seminar, Stanford *2018 2018*
Low Memory Convex Optimization

ISMP, Bordeaux *2018 2018*
Sketchy Decisions: Convex Optimization with Optimal Storage

Ecole Polytechnique: Statistics Special Seminar, Paris *2018 2018*
Big Data is Low Rank

DARPA D3M Workshop, Arlington *2018*
Composable Robust Structured Data Inference: AutoML, Causal Inference, Big Data is Low Rank

AI in advancement, Cornell *2018*
Panel Discussion

Penn State OR Colloquium, State College, PA *2018*
Big Data is Low Rank

Cornell Engineering College Council, New York, *2017*
The New Educational Paradigm: Data Science

INFORMS, Houston *2017*
Optimal Design of Rooftop Photovoltaic Arrays

SIMONS Institute, Berkeley *2017*
Sketchy Decisions: Convex Optimization with Optimal Storage

MIT ORC Seminar, Cambridge, MA *2017*
Sketchy Decisions: Convex Optimization with Optimal Storage

Capital One Tech Talk, New York *2017*
Low Rank Models for Automatic Machine Learning and Interpretability

Schonfeld Quantitative Conference, New York *2017*
Convex Optimization Modeling

STRATA, New York *2017*
Generalized Low Rank Models

Two Sigma Tech Talk, New York *2017* *2017*
Generalized Low Rank Models

CATALYST Academy Field Session: Operations Research, Cornell *2017* *2017*
Outreach session to introduce URM high school students to the discipline of OR

CURIE Academy Field Session: Operations Research, Cornell *2017*
Outreach session to introduce female high school students to the discipline of OR

JuliaCon, Berkeley *2017* *2017*
Julia: the Type of Language for Mathematical Programming

LCCC workshop on Distributed Optimization (Invited), Lund *2017* *2017*
Sketchy Decisions: Convex Optimization with Optimal Storage

UW Optimization Seminar, Seattle *2017* *2017*
Sketchy Decisions: Convex Optimization with Optimal Storage

SIOPT, Vancouver *2017* *2017*
Sketchy Decisions: Convex Optimization with Optimal Storage

DARPA D3M Kickoff, Arlington *2017*

B Bibliography

Appendix B

[1] M. Udell, C. Horn, R. Zadeh, and S. Boyd. Generalized low rank models. *Foundations and Trends in Machine Learning*, 9(1), 2016, 1410.0342.

[2] D. Davis, B. Edmunds, and M. Udell. The sound of APALM clapping: Faster nonsmooth nonconvex optimization with stochastic asynchronous PALM. In *Advances in Neural Information Processing Systems*, 2016, 1606.02338.

[3] J. Fan, L. Ding, Y. Chen, and M. Udell. Factor group-sparse regularization for effi- cient low-rank matrix recovery. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pages 5105–5115, 2019, 1911.05774.

[4] A. Schuler, V. Liu, J. Wan, A. Callahan, M. Udell, D. Stark, and N. Shah. Discover- ing patient phenotypes using generalized low rank models. In *Pacific Symposium on Biocomputing (PSB)*, 2016.

[5] Sengupta, Nandana, M. Udell, N. Srebro, and J. Evans. Matrix factorization for missing value imputation and sparse data reconstruction. *Submitted*, 2017.

[6] N. Kallus and M. Udell. Learning preferences from assortment choices in a heterogeneous population. In *ICML Workshop on Computational Frameworks for Personalization*, 2016, 1509.05113.

[7] N. Kallus and M. Udell. Revealed preference at scale: Learning personalized preferences from assortment choices. In *The 2016 ACM Conference on Economics and Computation*, New York, NY, USA, 2016. ACM.

[8] N. Kallus and M. Udell. Dynamic assortment personalization in high dimensions. *Operations Research*, 2019, 1610.05604.

[9] J. A. Tropp, A. Yurtsever, M. Udell, and V. Cevher. Streaming low-rank matrix approximation with an application to scientific simulation. *SIAM Scientific Computing (SISC)*, 41(4):A2430–A2463, 2019, 1902.08651.

[10] Y. Sun, Y. Guo, C. Luo, J. A. Tropp, and M. Udell. Low-rank Tucker approximation of a tensor from streaming data. *SIAM Journal on Mathematics of Data Science (SIMODS)*, 2020, 1904.10951.

[11] E. A. Ricci, M. Udell, and R. A. Knepper. An information-theoretic approach to per- sistent environment monitoring through low rank model based planning and prediction, 2020, 2009.01168.

[12] E. Lee, M. Udell, and S. Wong. Factorization for analog-to-digital matrix multiplication.In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015.

[13] C. Yang, Y. Akimoto, D. W. Kim, and M. Udell. OBOE: Collaborative filtering for AutoML initialization (workshop version). *NeurIPS Workshop on Automated Machine Learning*, 2018, 1808.03233.

[14] C. Yang, Y. Akimoto, D. W. Kim, and M. Udell. OBOE: Collaborative filtering for AutoML model selection. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, volume 25, pages 1173–1183. ACM, 2019, 1808.03233.

[15] C. Yang, Z. Wu, J. Fan, and M. Udell. AutoML pipeline selection: Efficiently navigating the combinatorial space. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2020, 2006.04216.

[16] M. Paradkar and M. Udell. Graph-regularized generalized low rank models. In *CVPR Workshop on Tensor Methods in Computer Vision*, 2017.

[17] M. Udell and A. Townsend. Why are big data matrices approximately low rank? *SIAM Journal on Mathematics of Data Science (SIMODS)*, 1(1):144–160, 2019, 1705.07474.

[18] J. Fan and M. Udell. Online high-rank matrix completion. In *Computer Vision and Pattern Recognition (CVPR)*, pages 8690–8698, 2019.

[19] J. Fan, Y. Zhang, and M. Udell. Polynomial matrix completion for missing data imputation and transductive learning. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 3842–3849, 2020, 1912.06989.

[20] Y. Zhao and M. Udell. Missing value imputation for mixed data through gaussian copula.In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2020, 1910.12845.

[21] Y. Zhao, E. Landgrebe, E. Shekhtman, and M. Udell. Online missing value imputation and correlation change detection for mixed-type data via gaussian copula. *Submitted*, 2020, 2009.12326.

[22] Y. Zhao and M. Udell. Matrix completion with quantified uncertainty through low rank gaussian copula. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020, 2006.10829.

[23] I. Drori, L. Liu, S. Koorathota, N. Yi, J. Li, A. Moretti, J. Freire, and M. Udell. AutoML using metadata language embeddings. In *NeurIPS Workshop on Meta-Learning*, 2019, 1910.03698.

[24] I. Drori, L. Liu, Q. Ma, J. Deykin, B. Kates, and M. Udell. Real-time AutoML. *Sub- mitted*, 2020.

[25] C. Yang, J. Fan, Z. Wu, and M. Udell. Efficient AutoML pipeline search with matrix and tensor factorization, 2020, 2006.04216.

[26] C. Yang, L. Ding, Z. Wu, and M. Udell. Tenips: Inverse propensity sampling for tensor completion. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2021, 2101.00323.

[27] J. A. Tropp, A. Yurtsever, M. Udell, and V. Cevher. Fixed-rank approximation of a positive-semidefinite matrix from streaming data. In *Advances in Neural Information Processing Systems*, 2017, 1706.05736.

[28] J. A. Tropp, A. Yurtsever, M. Udell, and V. Cevher. More practical sketching algorithms for low-rank matrix approximation. Technical Report 2018-01, California Institute of Technology, Pasadena, California, 2018.

[29] Y. Sun, Y. Guo, J. A. Tropp, and M. Udell. Tensor random projection for low memory dimension reduction. In *NeurIPS Workshop on Relational Representation Learning*, 2018.

[30] W. Stephenson, M. Udell, and T. Broderick. Approximate cross-validation with low- rank data in high dimensions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020, 2008.10547.

[31] R. Muthukumar, D. Kouri, and M. Udell. Randomized sketching algorithms for low memory dynamic optimization. *Under revision at SIOPT*, 2019.

[32] A. Yurtsever, M. Udell, J. A. Tropp, and V. Cevher. Sketchy decisions: Convex low- rank matrix optimization with optimal storage. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1188–1196, 2017, 1702.06838.

[33] A. Yurtsever, J. A. Tropp, O. Fercoq, M. Udell, and V. Cevher. Scalable semidefinite programming. *Accepted at SIAM Journal on Mathematics of Data Science (SIMODS)*, 2019, 1912.02949.

[34] L. Ding and M. Udell. Frank-Wolfe style algorithms for large scale optimization. In *Large-Scale and Distributed Optimization*. Springer, 2018.

[35] L. Ding, A. Yurtsever, V. Cevher, J. A. Tropp, and M. Udell. An optimal-storage approach to semidefinite programming using approximate complementarity. *Major Re- vision at SIOPT*, 2019, 1902.03373.

[36] L. Ding and M. Udell. On the regularity and conditioning of low rank semidefinite programs. *Submitted to SIOPT*, 2020, 2002.10673.

[37] L. Ding and M. Udell. A strict complementarity approach to error bound and sensitivityof solution of conic programs. *Submitted*, 2020, 2012.00183.

[38] J. Fan, L. Ding, C. Yang, and M. Udell. Low-rank tensor recovery with euclidean-norm-induced schatten-p quasi-norm regularization. *Submitted*, 2020, 2012.03436.

[39] L. Ding, J. Fan, and M. Udell. *k*FW: A Frank-Wolfe style algorithm with stronger subproblem oracles. *Submitted*, 2020, 2006.16142.

[40] J. Fan, C. Yang, and M. Udell. Robust non-linear matrix factorization for dictionary learning, denoising, and clustering. *Major Revision at IEEE Trans. Signal Processing (TSP)*, 2020, 2005.01317.

[41] M. Udell and O. Toole. Optimal design of efficient rooftop photovoltaic arrays. *IN-FORMS Journal on Applied Analytics (Interfaces)*, 49(4):281–294, 2019.

[42] J. A. Tropp, A. Yurtsever, M. Udell, and V. Cevher. Practical sketching algorithms for low-rank matrix approximation. *SIAM Journal of Matrix Analysis and Applications (SIMAX)*, 38(4):1454–1485, 2017, 1609.00048.

[43] J. Chen, N. Kallus, X. Mao, G. Svacha, and M. Udell. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *FAT*: Conference on Fairness, Accountability, and Transparency*, pages 339–348, 2019, 1811.11154.

[44] S. Zhou, S. Gupta, and M. Udell. Limited memory Kelley's method converges for composite convex and submodular objectives. In *Advances in Neural Information Processing Systems*, 2018, 1807.07531.

[45] N. Kallus, X. Mao, and M. Udell. Causal inference with noisy and missing covariates via matrix factorization. In *Advances in Neural Information Processing Systems*, 2018, 1806.00811.

[46] Y. Zhang, K. Song, Y. Sun, S. Tan, and M. Udell. "Why should you trust my explanation?" understanding uncertainty in LIME explanations. In *ICML Workshop AI for Social Good*, 2019, 1904.12991.

List of symbols, abbreviations, and acronyms

AutoML	Automatic Machine Learning
SDP	Semidefinite programming