Practical Design of Performant Recommender Systems using Large-scale Linear Programming-based Global Inference

Practical Design of Performant Recommender Systems using Large-scale Linear Programming-based Global Inference

Aman Gupta
S. Sathiya Keerthi
amagupta@linkedin.com
keselvaraj@linkedin.com
LinkedIn
Sunnyvale, California, USA

Ayan Acharya
Miao Cheng
acharya@linkedin.com
miacheng@linkedin.com
LinkedIn
Sunnyvale, California, USA

Borja Ocejo Elizondo
bocejo@linkedin.com
LinkedIn
Sunnyvale, California, USA

Rohan Ramanath*
ron.ramanath@gmail.com
Chico AI

Rahul Mazumder
rmazumder@linkedin.com
LinkedIn
Sunnyvale, California, USA

Kinjal Basu*
kinjal@alumni.stanford.edu
Aliveo AI

J. Kenneth Tay
ktay@linkedin.com
LinkedIn
Sunnyvale, California, USA

Rupesh Gupta
rugupta@linkedin.com
LinkedIn
Sunnyvale, California, USA

ABSTRACT

Several key problems in web-scale recommender systems, such as optimal matching and allocation, can be formulated as large-scale linear programs (LPs) [4, 1]. These LPs take predictions from ML models such as probabilities of click, like, etc. as inputs and optimize recommendations made to users. In recent years, there has been an explosion in the research and development of large-scale recommender systems, but effective optimization of business objectives using the output of those systems remains a challenge. Although LPs can help optimize such business objectives, and algorithms for solving LPs have existed since the 1950s [5, 8], generic LP solvers cannot handle the scale of these problems. At LinkedIn, we have developed algorithms that can solve LPs of various forms with trillions of variables in a Spark-based library called "DuaLip" [7], a novel distributed solver that solves a perturbation of the LP problem at scale via gradient-based algorithms on the smooth dual of the perturbed LP. DuaLip has been deployed in production at LinkedIn and powers several very large-scale recommender systems. DuaLip is open-sourced and extensible in terms of features and algorithms.

In this first-of-its-kind tutorial, we will motivate the application of LPs to improve recommender systems, cover the theory of key LP algorithms [8, 6], and introduce DuaLip (https://github.com/linkedin/DuaLip), a highly performant Spark-based library that solves extreme-scale LPs for a large variety of recommender system problems. We will describe practical successes of large-scale LP in the industry [3, 2, 9], followed by a hands-on exercise to run DuaLip.

ACM Reference Format:
Aman Gupta, S. Sathiya Keerthi, Ayan Acharya, Miao Cheng, Borja Ocejo Elizondo, Rohan Ramanath, Rahul Mazumder, Kinjal Basu, J. Kenneth Tay, and Rupesh Gupta. 2023. Practical Design of Performant Recommender Systems using Large-scale Linear Programming-based Global Inference. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23), August 6–10, 2023, Long Beach, CA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3580305.3599183

TUTORIAL OUTLINE

(1) **Intro to the tutorial (10 minutes)**
(2) **LPs in Recommender Systems (40 minutes)**
  (a) How LP formulations arise in recommender systems
  (b) Discuss one MOO and one matching problem
(3) **Algorithms and scalable implementation (25 minutes)**
  (a) Discuss difficulties with solving large-scale LPs
  (b) Discuss dual decomposition approach
  (c) Discussion of alternative solutions
(4) **Coffee break (15 minutes)**
(5) **Introduction to DuaLip and its usage in 2 applications (10 minutes)**
  (a) High-level overview of the properties of DuaLip
(6) **Hands-on DuaLip (70 minutes) [2 speakers + 2 devs]**
  (a) Installation of DuaLip + 2 use cases

TUTORS' BIO

Keerthi Selvaraj is a Principal Staff Researcher at LinkedIn where he works on huge scale LPs and Deep Net optimization projects. Prior to LinkedIn, he was a Distinguished researcher at Criteo Research, working on fundamental and applied research problems in computational advertising. Before that, he spent time at Microsoft. Even earlier, he was with the Machine Learning Group of Yahoo!

*Work done while the authors were at LinkedIn.

KDD '23, August 6–10, 2023, Long Beach, CA, USA

ACM ISBN 979-8-4007-0103-0/23/08.
https://doi.org/10.1145/3580305.3599183

Research. Keerthi also worked for 11 years at the Indian Institute of Science, Bangalore, and for 5 years at the National University of Singapore. His research focused on the development of practical algorithms for areas such as machine learning, robotics, computer graphics and optimal control. Overall, he has published 100+ papers in leading journals and conferences.

Rahul Mazumder is an Associate Professor (with tenure) at MIT Sloan School of Management, affiliated with the Operations Research Center, Center for Statistics and Data Science. He is also a LinkedIn consultant. His research interests are in statistics, ML and large-scale optimization. He is a recipient of the Donald P. Gaver, Jr. Early Career Award for Excellence in Operations Research, INFORMS Optimization Society Young Researchers Prize, Office of Naval Research Young Investigator Award. Student co-authors of his papers have received: SIGKDD Best Student Paper Award '22, INFORMS Optimization Society Best Student Paper Award '15, and other paper awards from INFORMS Computing Society '20, Mixed Integer Programming Workshop '18, '21, MIT Operations Research Center '20.

Aman Gupta is the manager of the AI foundations optimization team at LinkedIn, overseeing the integration of DuaLip with products. His work has had significant impact on products like Feed, Ads, Jobs and People You May Know (PYMK). Aman's interests lie in large-scale optimization and ML modeling. Prior to LinkedIn, Aman spent several years building scalable ML systems for Apple and Amazon. He holds degrees from Carnegie Mellon University, Pittsburgh and BITS, Pilani, He has published papers and conducted tutorials at top conferences like KDD and NeurIPS.

Kenneth Tay is a data scientist at LinkedIn. He is a contributor to the DuaLip project. Kenneth has experience in both the technical aspects of statistics and optimization, as well as the business aspects of applying such methods to client-facing problems (through data science consulting experience with the Singapore government). He earned his PhD in Statistics at Stanford University, advised by Robert Tibshirani.

Miao Cheng is a staff software engineer at LinkedIn, contributing to the development of extreme-scale parallelism with DuaLip for the PYMK recommender system at LinkedIn. Miao also helped onboard products within LinkedIn including premium and people search to use the DuaLip solver. Miao got her MS in Computer Science from University of Pennsylvania, and her Bachelor's degrees in Automation (EECS) and Economics from Tsinghua University.

Ayan Acharya is a senior software engineer at LinkedIn. He graduated from the Department of Elec. and Computer Engineering at the University of Texas at Austin, focusing on generative models and efficient inference algorithms for solving transfer learning problems prevalent in text document analysis, social networks, recommender systems. At LinkedIn, he has been involved with developing large-scale linear programming formulations for PYMK, Premium Promotions and Ads Autobidding and contributing towards DuaLip.

Borja Ocejo is a software engineer at LinkedIn, focusing on deep neural network optimization and large scale LPs applied to recommender systems. He earned his Master of Professional Studies degree in Information Science from Cornell University's College of Computing and Information Science.

Rohan Ramanath is an entrepreneur experienced in developing data products. He was a Sr. Staff Engineer for LinkedIn's Ads AI team. Rohan got his master's degree from Carnegie Mellon University. His research was published at top-tier conferences for AI and NLP. Rohan's work on extreme-scale linear programming solvers contributed a step-function improvement to marketplaces. Through personalization, his models contributed *xxx* million in realized revenue for LinkedIn Ads.

Kinjal Basu is an entrepreneur who most recently was a Sr. Staff Engineer at LinkedIn, primarily focusing on Responsible AI. His focus has ranged from developing prediction models for complex recommender systems powering Feed Ranking and PYMK to extreme large-scale optimization problems. He has been the chief architect for the AutoML library used internally by various teams. He has also worked towards developing accurate causal estimates in the presence of network interference. Before LinkedIn, Kinjal finished his PhD in the Department of Statistics at Stanford University advised by Prof. Art Owen. Prior to the doctoral program, he earned his B.Stat (Hons) and M.Stat from ISI Kolkata.

Rupesh Gupta is a Sr. Staff Engineer in the Search AI team at LinkedIn. He has over 10 years of industrial experience in search and recommender systems. In the past he has worked on real-time personalization, retention, feed relevance, email optimization and internal promotion optimization. He holds a master's degree from Purdue University, a master's degree from IIT Delhi, and a bachelor's degree from IIT Delhi.

REFERENCES

[1] Deepak Agarwal et al. 2015. Personalizing linkedin feed. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1651–1660.

[2] Eduardo M Azevedo and E Glen Weyl. 2016. Matching markets in the digital age. *Science*, 352, 6289, 1056–1057.

[3] Rupesh Gupta, Guangde Chen, and Shipeng Yu. 2019. Internal promotion optimization. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2358–2366.

[4] Rupesh Gupta, Guanfeng Liang, and Rómer Rosales. 2017. Optimizing email volume for sitewide engagement. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 1947–1955.

[5] Olvi L Mangasarian and RR Meyer. 1979. Nonlinear perturbation of linear programs. *SIAM Journal on Control and Optimization*, 17, 6, 745–752.

[6] Brendan O'donoghue, Eric Chu, Neal Parikh, and Stephen Boyd. 2016. Conic optimization via operator splitting and homogeneous self-dual embedding. *Journal of Optimization Theory and Applications*, 169, 1042–1068.

[7] Rohan Ramanath, S Sathiya Keerthi, Yao Pan, Konstantin Salomatin, and Kinjal Basu. 2022. Efficient vertex-oriented polytopic projection for web-scale applications. In *Proceedings of the AAAI Conference on Artificial Intelligence* number 4. Vol. 36, 3821–3829.

[8] Philip Wolfe. 1976. Finding the nearest point in a polytope. *Mathematical Programming*, 11, 128–149.

[9] Huanyang Zheng and Jie Wu. 2017. Online to offline business: urban taxi dispatching with passenger-driver matching stability. In *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 816–825.