MSR CORE's Optimization Toolkit

Ishai Menache and Marco Molinaro

February 23, 2023

In this document we provide an overview of the optimization tools that have been central in our research.

1 Integer Programming

Integer Programming (IP) is one of the most important tools in optimization whose impact spans a variety of applications, such as train scheduling in Europe, planning the MLB season in the U.S., and numerous supply-chain optimizations.

Our research provides fundamental understanding of the main algorithmic components involved in solving IPs, such as branching [27, 28] and cutting planes [7, 20, 16, 21, 8, 25, 11, 29, 32, 31, 9, 13, 33, 14, 12], which are crucial for enhancing the solution of complex and large scale problems. In particular, we recently showed that Branch-and-Bound (the standard algorithm for solving Integer Programs) solves random knapsack problems in polynomial time [26]. Surprisingly, although the Branch-and-Bound algorithm has been widely applied in practice over several decades, this is the first theoretical proof of its effectiveness. See [54, 15, 30, 10] for our additional theoretical work in this area.

Based on our deep understanding of IP, our group is also conducting research on accelerating IP methods using Machine Learning (ML). As an example, we are currently exploring the use of ML for speeding up the sequential solution of related Integer Programming instances, e.g., where the input data does not differ by much across instances.

2 Stochastic Optimization

In many applications, there is inherent uncertainty, e.g., about future demands, prices, etc., which needs to be taken into account for appropriate decision-making. Multi-stage stochastic optimization is one important tool for tackling such problems, especially when there is some a-priori knowledge about the distribution of uncertain elements. Unfortunately, off-the-shelf stochastic optimization solutions suffer from the curse of dimensionality and cannot handle larger problem instances.

In our research, we develop new and more tractable ways of incorporating uncertainty, as well as new computational methods for handling stochastic problems at scale. Some examples are:

- By exploiting theoretical properties of the problem, we are able to perform end-to-end optimization of server deployment at Microsoft datacenters, taking into account fine-grained practical constraints, as well as multiple risk measures [40].
- We propose a new lightweight method based on simulating future realizations of uncertainty. Not only do we provide theoretical support for this method, but we have also successfully employed it in production for optimizing infrastructure placement at Microsoft

datacenters; this is expected to reduce stranded power across Microsoft datacenters, which would translate into annual savings of hundreds of million dollars [24].

3 ML-Augmented Optimization

Machine Learning-based predictions have become an important component in decision-making. However, in many real scenarios, the predictions are fairly noisy. One fundamental research question here is how to judiciously utilize noisy predictions and obtain outcomes that are better than ignoring them. In our research, we design provably robust algorithms that incorporate noisy predictions. Our work in this area has led to novel VM allocation algorithms which utilize VM lifetime information and lead to higher packing efficiency [17, 6]. We are working on deploying these algorithms within Azure's VM allocator.

4 Online Algorithms

Solving complex problems under uncertainty often requires designing strategies that perform provably well even without predictions. Online algorithms target *sequential* decisions under such conditions.

Our research on online algorithms includes foundational theoretical aspects, where we obtained state-of-the-art results for some of the most classical problems in the area (e.g., Load-Balancing and Online Resource Allocation [35, 34, 39]). An important general direction herein is designing "best-of-both-worlds" algorithms, which automatically adapt to the uncertainty nature of the problem, while performing significantly better than worst-case guarantees whenever the uncertainty is more "benign". We obtained several state-of-the art result in this area [47, 38, 48, 1].

Another large body of our work studies fundamental resource allocation problems that arise in different cloud settings, including big-data analytics systems [36, 41, 19, 3], database-as-a-service [46, 51], VM allocation [17], WAN routing [37], systems for ML [5] and cloud security [4]. See [50, 34, 18, 49, 2, 52, 44] for some of our additional work in this area.

5 Large-scale Heuristics

Due to the scale of some real-life problems, their large solution space precludes the use of classic optimization techniques such as Integer Programming. In our research, we design novel large-scale heuristics to tackle such problems. As an example, we designed a hybrid approach that uses Adaptive Large Neighborhood Search augmented with Linear Programming to optimize large food-service operations [43]. We have also developed similar techniques for performing multi-itinerary optimization [22, 23], which are currently used in production as part of Bing Maps APIs. We also designed large-scale optimizers for Dynamics'365 Field and Retail offerings, where we combine LP solvers and our own combinatorial methods for trimming the huge search space.

6 Optimization via ML

Traditionally, Machine Learning has been largely applied to provide and refine the input for optimization models, which are then responsible for making decisions, an approach referred to as *predict-then-optimize*.

However, motivated by the advances in ML's scope and scale, our group has been researching the possibility of using ML *directly* for performing end-to-end decision-making, bypassing the need for separate ML and optimization components. This may lead to improved solution quality given the more direct coupling between data and decisions. As an example, in recent work [53] we apply stochastic gradient descent to solve a WAN routing problem with demand uncertainty. We prove the asymptotic optimality of our scheme, and show empirically that it outperforms the current state-of-the-art (which relies on a predict-then-optimize), in both accuracy and execution time.

We have also been examining, both theoretically and empirically, the use of other direct schemes for resource management, including deep RL [42, 55] and bandits [45]. Finally, in on-going work, we are investigating how GNNs and LLMs can be used to solve challenging combinatorial problems.

References

[1] C. Argue, A. Gupta, M. Molinaro, and S. Singla. *Robust Secretary and Prophet Algorithms for Packing Integer Programs*, pages 1273–1297. 2022.

[2] C. J. Argue, A. Gupta, and M. Molinaro. Lipschitz selectors may not yield competitive algorithms for convex body chasing. *Discrete and Computational Geometry*, 2023.

[3] Y. Azar, I. Kalp-Shaltiel, B. Lucier, I. Menache, J. Naor, and J. Yaniv. Truthful online scheduling with commitments. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, pages 715–732, 2015.

[4] Y. Azar, S. Kamara, I. Menache, M. Raykova, and B. Shepard. Co-location-resistant clouds. In *Proceedings of the 6th Edition of the ACM Workshop on Cloud Computing Security*, pages 9–20, 2014.

[5] M. Babaioff, R. Lempel, B. Lucier, I. Menache, A. Slivkins, and S. C.-w. Wong. Truthful online scheduling of cloud workloads under uncertainty. In *Proceedings of the ACM Web Conference 2022*, pages 151–161, 2022.

[6] H. Barbalho, P. Kovaleski, B. Li, L. Marshall, M. Molinaro, A. Pan, E. Cortez, M. Leao, H. Patwari, Z. Tang, T. V. C. Santos, L. R. Gonçalves, D. Dion, T. Moscibroda, and I. Menache. Virtual machine allocation with lifetime predictions. 2023. Submitted.

[7] A. Basu, G. Cornuéjols, and M. Molinaro. A probabilistic analysis of the strength of the split and triangle closures. In *IPCO*, pages 27–38, 2011.

[8] A. Basu, R. Hildebrand, M. Köppe, and M. Molinaro. A (k+1)-Slope Theorem for the k-Dimensional Infinite Group Relaxation. *ArXiv e-prints*, 2011.

[9] A. Basu, R. Hildebrand, and M. Molinaro. Minimal cut-generating functions are nearly extreme. *Mathematical Programming*, 1–2(172), 2018.

[10] A. Basu, H. Jiang, P. Kerger, and M. Molinaro. Information complexity of mixed-integer convex optimization. In *IPCO*, 2023.

[11] A. Basu and M. Molinaro. Characterization of the split closure via geometric lifting. *Eur. J. Oper. Res.*, 243(3):745–751, 2015.

[12] G. Blekherman, S. S. Dey, M. Molinaro, and S. Sun. Sparse PSD approximation of the PSD cone. *Math. Program.*, 191(2):981–1004, 2022.

[13] M. Bodur, A. Del Pia, S. S. Dey, M. Molinaro, and S. Pokutta. Aggregation-based cutting-planes for packing and covering integer programs. *Mathematical Programming*, 171(1–2), 2018.

[14] M. Bodur, A. D. Pia, S. S. Dey, and M. S. Molinaro. Lower bounds on the lattice-free rank for packing and covering integer programs. *SIAM J. Optim.*, 29(1):55–76, 2019.

[15] N. Boland, S. S. Dey, T. Kalinowski, M. Molinaro, and F. Rigterink. Bounding the gap between the mccormick relaxation and the convex hull for bilinear functions. *Math. Program.*, 162(1–2):523–535, mar 2017.

[16] P. Bonami, M. Conforti, G. Cornuéjols, M. Molinaro, and G. Zambelli. Cutting planes from two-term disjunctions. *Oper. Res. Lett.*, 41(5):442–444, 2013.

[17] N. Buchbinder, Y. Fairstein, K. Mellou, I. Menache, and J. S. Naor. Online virtual machine allocation with lifetime and load predictions. In L. Huang, A. Gandhi, N. Kiyavash, and J. Wang, editors, *SIGMETRICS '21: ACM SIGMETRICS / International Conference on Measurement and Modeling of Computer Systems, Virtual Event, China, June 14-18, 2021*, pages 9–10. ACM, 2021.

[18] N. Buchbinder, A. Gupta, M. Molinaro, and J. S. Naor. k-servers with a smile: Online algorithms via projections. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2019, San Diego, California, USA, January 6-9, 2019*, pages 98–116, 2019.

[19] N. Buchbinder, N. Jain, and I. Menache. Online job-migration for reducing the electricity bill in the cloud. In *NETWORKING 2011: 10th International IFIP TC 6 Networking Conference, Valencia, Spain, May 9-13, 2011, Proceedings, Part I 10*, pages 172–185. Springer, 2011.

[20] G. Cornuéjols, T. Kis, and M. Molinaro. Lifting gomory cuts with bounded variables. *Operations Research Letters*, 41:142–146, 2013.

[21] G. Cornuéjols and M. Molinaro. A 3-slope theorem for the infinite relaxation in the plane. *Mathematical Programming*, pages 1–23. 10.1007/s10107-012-0562-7.

[22] A. Cristian, L. Marshall, M. Negrea, F. Stoichescu, P. Cao, and I. Menache. Multi-itinerary optimization as cloud service. In *ACM SIGSPATIAL 2019*, November 2019.

[23] A. Cristian, L. Marshall, M. Negrea, F. Stoichescu, P. Cao, and I. Menache. Multi-itinerary optimization as cloud service. *Communications of the ACM*, 64(11):121–129, 2021.

[24] K. Cummings, K. Mellou, I. Menache, and M. Molinaro. Rack placement with uncertain demands. 2023. In preparation.

[25] S. Dash, O. Günlük, and M. Molinaro. On the relative strength of different generalizations of split cuts. *Discrete Optimization*, 16:36–50, 2015.

[26] S. S. Dey, Y. Dubey, and M. Molinaro. Branch-and-bound solves random binary ips in poly-time. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 579–591. SIAM, 2021.

[27] S. S. Dey, Y. Dubey, and M. Molinaro. Lower bounds on the size of general branch-and-bound trees. *Mathematical Programming*, 2023. To appear.

[28] S. S. Dey, Y. Dubey, M. Molinaro, and P. Shah. A theoretical and computational analysis of full strong-branching. *Mathematical Programming*, page To appear, 2023.

[29] S. S. Dey, A. Iroume, and M. Molinaro. Some lower bounds on sparse outer approximations of polytopes. *Operations Research Letters*, 43(3):323–328, 2015.

[30] S. S. Dey, A. Iroume, M. Molinaro, and D. Salvagnin. Improving the randomization step in feasibility pump. *SIAM Journal on Optimization*, 28(1):355–378, 2018.

[31] S. S. Dey and M. Molinaro. Theoretical challenges towards cutting-plane selection. *Math. Program.*, 170(1):237–266, 2018.

[32] S. S. Dey, M. Molinaro, and Q. Wang. Approximating polyhedra with sparse inequalities. *Mathematical Programming*, pages 1–24, 2015.

[33] S. S. Dey, M. Molinaro, and Q. Wang. Analysis of sparse cutting planes for sparse milps with applications to stochastic milps. *Mathematics of Operations Research*, 43(1):304–332, 2018.

[34] A. Gupta, R. Mehta, and M. Molinaro. Maximizing Profit with Convex Costs in the Random-order Model. In I. Chatzigiannakis, C. Kaklamanis, D. Marx, and D. Sannella, editors, *45th International Colloquium on Automata, Languages, and Programming (ICALP 2018)*, volume 107 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 71:1–71:14, Dagstuhl, Germany, 2018. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[35] A. Gupta and M. Molinaro. How the experts algorithm can help solve lps online. *Math. Oper. Res.*, 41(4):1404–1431, 2016.

[36] S. A. Jyothi, C. Curino, I. Menache, S. M. Narayanamurthy, A. Tumanov, J. Yaniv, R. Mavlyutov, I. Goiri, S. Krishnan, J. Kulkarni, et al. Morpheus: Towards automated slos for enterprise clusters. In *OSDI*, pages 117–134, 2016.

[37] S. Kandula, I. Menache, R. Schwartz, and S. R. Babbula. Calendaring for wide area networks. In *Proceedings of the 2014 ACM conference on SIGCOMM*, pages 515–526, 2014.

[38] T. Kesselheim and M. Molinaro. Knapsack Secretary with Bursty Adversary. In A. Czumaj, A. Dawar, and E. Merelli, editors, *47th International Colloquium on Automata, Languages, and Programming (ICALP 2020)*, volume 168 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 72:1–72:15, Dagstuhl, Germany, 2020. Schloss Dagstuhl–Leibniz-Zentrum für Informatik.

[39] T. Kesselheim, M. Molinaro, and S. Singla. Online and bandit algorithms beyond lp norms. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2023*, 2023.

[40] R. P. Liu, K. Mellou, T. Coffee, B. Li, J. Pathuri, X. Gong, D. Simchi-Levi, and I. Menache. Fast and exact cloud server deployment under demand uncertainty. 2023. In preparation.

[41] B. Lucier, I. Menache, J. Naor, and J. Yaniv. Efficient online scheduling for deadline-sensitive jobs. In *Proceedings of the twenty-fifth annual ACM symposium on Parallelism in algorithms and architectures*, pages 305–314, 2013.

[42] H. Mao, M. Alizadeh, I. Menache, and S. Kandula. Resource management with deep reinforcement learning. In *Proceedings of the 15th ACM workshop on hot topics in networks*, pages 50–56, 2016.

[43] K. Mellou, L. Marshall, K. Chintalapudi, P. Jaillet, and I. Menache. Optimizing onsite food services at scale. In *ACM SIGSPATIAL 2020*, November 2020.

[44] K. Mellou, M. Molinaro, and R. Zhou. Online demand scheduling with failovers. 2023. https://arxiv.org/abs/2209.00710.

[45] I. Menache, O. Shamir, and N. Jain. On-demand, spot, or both: Dynamic resource allocation for executing batch jobs in the cloud. In *11th International Conference on Autonomic Computing ({ICAC} 14)*, pages 177–187, 2014.

[46] I. Menache and M. Singh. Online caching with convex costs. In *Proceedings of the 27th ACM symposium on Parallelism in Algorithms and Architectures*, pages 46–54, 2015.

[47] M. Molinaro. Online and random-order load balancing simultaneously. In *Proceedings of the 2017 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1638–1650, 2017.

[48] M. Molinaro. Robust algorithms for online convex problems via primal-dual. In D. Marx, editor, *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 2078–2092. SIAM, 2021.

[49] M. Molinaro. Strong convexity of feasible sets in off-line and online optimization. *Mathematics of Operations Research*, 2023.

[50] M. Molinaro and R. Ravi. The geometry of online packing linear programs. *Mathematics of Operations Research*, 39(1):46–59, 2014.

[51] V. Narasayya, I. Menache, M. Singh, F. Li, M. Syamala, and S. Chaudhuri. Sharing buffer pool memory in multi-tenant relational database-as-a-service. *Proceedings of the VLDB Endowment*, 8(7):726–737, 2015.

[52] S. Perez-Salazar, I. Menache, M. Singh, and A. Toriello. Dynamic resource allocation in the cloud with near-optimal efficiency. *Operations Research*, 70(4):2517–2537, 2022.

[53] Y. Perry, F. V. Frujeri, C. Hoch, S. Kandula, I. Menache, M. Schapira, and A. Tamar. Dote: Rethinking (predictive) wan traffic engineering. In *NSDI*, 2023.

[54] A. D. Pia, S. S. Dey, and M. Molinaro. Mixed-integer quadratic programming is in np. *Math. Program.*, 162(1–2):225–240, mar 2017.

[55] S. R. Sinclair, F. V. Frujeri, C.-A. Cheng, L. Marshall, H. Barbalho, J. Li, J. Neville, I. Menache, and A. Swaminathan. Hindsight learning for mdps with exogenous inputs. 2023. Submitted.