专栏：数据驱动的优化
Data-Driven Optimization

客座编辑

陈卫（1968– ），男，博士，微软亚洲研究院首席研究员，中国科学院计算技术研究所客座研究员，中国计算机学会理论计算机科学专业委员会常务委员、大数据专家委员会委员，IEEE Fellow，入选斯坦福大学全球前2%顶尖科学家榜单。主要研究方向为在线学习和优化、社交和信息网络、网络博弈论和经济学、分布式计算、容错等，在社交网络影响力传播和最大化以及组合在线学习方向做出了很多颇有影响力的工作，该方面论文被引次数已逾一万次。在信息和影响力传播方面，2013年合著一本英文专著，2020年独立撰写一本中文专著。担任《大数据》等多个学术期刊的编委，并在多个学术会议中担任过技术委员会主席和委员。

导读

优化是计算机科学和运筹学领域的一个分支，它研究在不同场景不同模型下达到最优解的方法，在计算机工程和工业工程等领域有广泛的应用。传统的优化基于给定的模型及其参数的输入。这些模型和参数通常是通过从领域知识中获得的经验及对以往数据收集的结果进行分析获得的，这属于机器学习的范畴，即机器学习从收集的大量数据中总结出数据尊崇的模型和对应的参数设置。现有的从数据到优化结果的流程基本上先用机器学习学出模型和对应的参数，然后将模型和参数输入一个基于模型的优化算法得到优化结果。该流程有“分而治之”的好处：机器学习和优化有不同的技术，传统上也是两个不同的计算机科学分支，由不同的领域专家对它们进行研究。机器学习着重于从数据中提取和抽象出模型，优化的任务是从学得的模型中找到最优解。

但在大数据和人工智能时代，这样的分工可能会带来从数据端到优化端整体性能的损失。Balkanski等人最近就指出有些优化问题从采样数据到模型的学习过程是可行的，从模型到优化的过程也是可行的，但从采样数据到优化的端到端的目标却是不可行的[①]。这样的结果看似反直观，但它表达了机器学习和优化两个子任务潜在的不匹配问题。在大数据和人工智能的大背景下，很多应用需要不断地收集实时数据，优化的结果需要基于这些实时数据，模型只是其中的一个过渡部分。我们把这样的端到端的优化称为数据驱动的优化。数据驱动的优化在理论和应用上都带来了新的挑战。本专栏请到了3组学者从理论和实践的不同角度对数据驱动的优化加以阐述。

在《基于样本的优化》一文中，张智杰等人详细介绍了基于样本的优化框架，以及Balkanski等人[①]在这个优化框架下给出的学习和优化不匹配导致的框架的局限性；然后介绍了突破这种局限性的几个方案，其中包括作者提出的基于结构化采样的优化方案，即利用数据中的结构化信息将学习和优化方案匹配，从而实现能达到良好优化结果的端到端优化算法。

孔芳等人撰写的《基于优化反馈的组合在线学习》较全面地总结了组合在线学习的研究方向。这一方向可以被看作对线性单向的从数据到优化流程的有效改进。组合在线学习的关键步骤是加入了从优化结果到数据采样的反馈步骤，从而将单向流程变成带反馈的闭环。通过反复地从数据到学习到优化，再将优化结果返回，用于指导下一轮的数据采样，最终达到良好的优化效果。组合在线学习是将组合优化和在线学习很好结合的结果。文章总结了这个方向的基本框架和主要理论成果，对该方向的研究和应用很有帮助。

王金予等人在《强化学习在资源优化领域的应用》中介绍了他们将强化学习应用于资源优化领域的若干实例。这些应用的共同特点是都有大量数据，因此要基于大量数据进行优化。文章系统地介绍了如何对这些资源优化问题进行建模，如何进行智能体设计等，从而帮助读者学习如何通过数据驱动的方式进行资源优化。

数据驱动的优化是在大数据和人工智能时代做优化和决策的大趋势。它需要将数据采样、机器学习和优化有机地结合。本专栏的3篇文章肯定不能概括这个领域的所有方向，但希望它们作为一个引子，能激励有兴趣的研究者和实践者进一步深入地探索这一方向，并在这一方向得到更丰硕的成果。

① BALKANSKI E, RUBINSTEIN A, SINGER Y. The limitations of optimization from samples[C]// Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing. New York: ACM Press, 2017: 1016-1027.