分布式优化、学习理论与方法

陈为胜　著

科学出版社

分布式优化、学习理论与方法

陈为胜 著

科学出版社
北京

内 容 简 介

本书旨在介绍作者及其研究团队在分布式优化与学习理论方面的最新研究成果. 全书共 7 章, 第 1、2 章为绪论和相关数学基础; 第 3、4 章为连续时间和基于采样数据的分布式优化算法; 第 5、6 章分别为基于群体智能的分布式优化算法和分布式机器学习算法; 第 7 章为基于自适应神经网络输出反馈控制的分布式合作学习方案设计. 本书主要关注从分布式技术中总结出来的理论与方法方面的问题, 但相关研究结论可以为解决通信网络、电网、燃气网、交通网等相关的网络优化问题提供借鉴和指导.

本书适合通信、计算机、机器学习、数据处理、自动控制等相关专业的本科生、研究生、教师及相关工程技术人员学习或参考.

图书在版编目(CIP)数据

分布式优化、学习理论与方法/陈为胜著. —北京: 科学出版社, 2019.1
ISBN 978-7-03-059764-9

I. ①分… II. ①陈… III. ①机器学习-研究 IV. ①TP181

中国版本图书馆 CIP 数据核字(2018) 第 261786 号

责任编辑: 李 萍 孙翠勤 / 责任校对: 郭瑞芝
责任印制: 张 伟 / 封面设计: 迷底书装

科学出版社 出版
北京东黄城根北街 16 号
邮政编码: 100717
http://www.sciencep.com

北京中石油彩色印刷有限责任公司 印刷
科学出版社发行 各地新华书店经销
*
2019 年 1 月第 一 版 开本: 720 × 1000 B5
2019 年 1 月第一次印刷 印张: 12 1/2
字数: 252 000

定价: 90.00 元

(如有印装质量问题, 我社负责调换)

前 言

进入 21 世纪以后, 随着数字化技术与网络化技术的普及, 人类进入一个全新的时代. 其中, 以去中心化为代表的分布式技术已经成为这个时代的代表性的创新技术之一, 如大数据的分布式储存技术、飞行器分布式编队、区块链技术等. 这些新型技术相应地延伸出一些急需解决的新的基础理论问题, 这正是本书的研究出发点.

分布式优化问题是传统优化问题的进一步延伸, 是从当前工程实践中的分布式技术问题中抽象出来的, 如无线通信网络中的功率分配问题和电网中的最优调度问题等. 这些问题都涉及网络化系统的优化问题, 对应的最优解与网络拓扑、目标函数等都有关系. 与以往优化问题最大的不同是, 优化求解被具体到每个节点上, 而优化目标函数是全网的目标函数, 但每个节点却不知道全网的目标函数. 需要说明的是, 这与传统的并行优化和并行计算是不同的, 尽管有时候也有一些文献称并行优化为分布式优化.

分布式储存技术是大数据储存的典型技术之一, 这必然会导致大数据的分布式处理. 例如, 语音识别问题是把样本分布式储存后, 在每个处理器中进行小样本训练, 通过网络进行信息交互后, 最终达到和集中训练一样的效果. 这一问题是典型的分布式学习问题之一, 分布式学习问题是当前机器学习问题的进一步拓展, 它可以看作是分布式技术在人工智能领域的一个应用, 也可以看作是分布式优化的一个应用方向.

本书的研究成果得到多个研究机构的支持, 其中特别感谢国家自然科学基金面上项目 (61673308,61673014)、教育部新世纪优秀人才支持计划 (NCET-10-0665) 和西安电子科技大学华山学者支持计划. 感谢团队老师李靖、戴浩、房新鹏在统稿等方面所做的大量工作, 感谢博士生艾武、刘加云、高飞, 硕士生任鹏飞、花少勇、马建宏在相关内容的研究方面所做的工作, 感谢硕士生岳永凯、任广山、郑羽丰、刘孟辉在书稿整理方面所做的贡献, 还有其他对相关研究提出建设性意见的同行, 在

此一并致谢!

本书作者长期从事分布式优化、学习与控制理论和方法领域的研究工作, 本书是作者及其团队成员多年来在这一领域的相关研究成果的工作总结和提炼. 由于作者水平有限, 书中难免存在不足之处, 恳请广大读者批评指正.

作　者

2018 年 7 月于西安

符　号　表

$\mathbb{R}$	全体实数组成的集合
$\mathbb{R}_{\geqslant 0}$	全体非负实数组成的集合
$\mathbb{Z}_+$	全体正整数组成的集合
$\mathbb{N}$	全体自然数组成的集合
$\mathbb{R}^n$	全体 n 维实向量组成的集合
$\mathbb{R}^{m\times n}$	全体 $m\times n$ 维实矩阵组成的集合
1_n	分量都是 1 的 n 维向量
0_n	分量都是 0 的 n 维向量
I_n	n 阶单位矩阵
$A\otimes B$	矩阵 A 和 B 的克罗内克积
$x^{\mathrm{T}}y$	欧氏空间中的向量 x 与 y 的内积
$\|x\|$	向量 x 的欧氏范数, 即 $\sqrt{x^{\mathrm{T}}x}$
$\|A\|$	矩阵 A 的欧氏范数
$f(x)$	实函数
$\nabla f(x)$	函数 $f(x)$ 的梯度, $x\in\mathbb{R}^n$
$\nabla^2 f(x)$	函数 $f(x)$ 的 Hessian 矩阵, $x\in\mathbb{R}^n$
$\partial f(x)$	凸函数 $f(x)$ 的次梯度, $x\in\mathbb{R}^n$
inf	取下确界
$\forall$	任意
$\in$	属于
$\subset$	包含于
$\arg\min_X f(x)$	在集合 X 上的 $f(x)$ 的全局最小值点
$\sigma_{\min}(A)$	矩阵 A 的最小奇异值

$\rho(A)$	矩阵 A 的谱半径
$\lambda_{\min}(A)$	矩阵 A 的最小特征值
$\lambda_{\max}(A)$	矩阵 A 的最大特征值
$E(f)$	函数 f 的期望值
$\mathrm{conv}(c)$	集合 c 的凸包
$\mathrm{span}\{1_\mathrm{n}\}$	由 1_n 张成的零子空间

缩略语表

ADMM	alternating direction method of multipliers	交替方向乘子法
ATC	adapt-then-combine	先自适应后组合
CL	centralized learning	集中学习
CTA	combine-then-adapt	先组合后自适应
DAC	distributed average consensus	分布式平均一致性
DCA	distributed cooperative adaptation	分布式合作自适应
DCL	distributed cooperative learning	分布式合作学习
DCOP	distributed convex optimization problem	分布式凸优化问题
DL	decentralized learning	分散学习
D-PSO-ON	distributed particle swarm optimization over networks	基于网络的分布式粒子群优化算法
DSM	distributed sub-gradient method	分布式次梯度方法
FBF	fuzzy base function	模糊基函数
FLS	fuzzy logic system	模糊逻辑系统
FMF	fuzzy membership function	模糊隶属度函数
LASSO	least-absolute shrinkage and selection operator	最小绝对收缩和选择算子
LMS	least-mean square	最小均方
LSR	local silencing rule	局部终止规则
MAS	multi-agent system	多智能体系统
MCR	misclassification rate	误分类率
MSE	mean square error	均方误差

NCS	networked control system	网络化控制系统
NDCS	networked distributed control system	网络化分布式控制系统
NN	neural network	神经网络
P2P	peer-to-peer	点对点
PSO	particle swarm optimization	粒子群优化算法
RBF	radial basis function	径向基函数
RVFL	random vector functional-link	随机向量函数链接
SD-ET-ZGS	sampled data based event-triggered zero gradient sum	基于采样数据的事件驱动零梯度和
SNR	signal-to-noise ratio	信噪比
SVM	support vector machine	支持向量机
UES	uniformly exponential stability	一致指数稳定性
ULES	uniformly locally exponentially stable	一致局部指数稳定
UUB	uniformly ultimate boundedness	一致最终有界
WSN	wireless sensor network	无线传感器网络
ZGS	zero gradient sum	零梯度和

目　　录

第 1 章 绪 论

1.1 分布式优化理论

在对网络化多智能体系统的研究中, 分布式优化问题[1] 越来越受到广大学者的关注, 并逐渐成为一个新的极具挑战性的研究领域, 分布式优化问题的研究在机器学习[2-4]、计算机网络的资源配置[5] 和传感网络的源定位领域[6] 都有着广泛的应用, 如无线网络中节点的公平资源配置问题和大规模机器学习的问题, 最终都会转化为一个分布式优化问题. 分布式优化问题是在传统优化问题的基础上加入了网络环境, 这里的网络指的是多个处理节点构成的物理网络, 每个节点对应多智能体系统中的一个决策个体. 这些节点通常不是集中在一起, 而是在空间上分散在多个地点, 处在通信范围内的节点自然形成连接, 可以进行信息传输. 节点之间的通信通常采用异步方式, 并不要求系统同步, 因此在应用场景上更加广泛.

分布式优化问题是将整个网络系统复杂的大规模优化问题分布到多个节点上进行分布式优化计算. 这里的分布式是指各节点不需要知道全局信息, 而只需要根据获得的局部信息, 通过一定的协调机制和规则, 独立地进行各自的优化和决策, 最终实现整个网络系统的优化目标. 分布式优化问题的典型特征是将求解基于全局信息的大规模复杂问题的过程, 转化为求解基于局部信息的较小规模优化问题的过程. 这种“分而治之”的求解过程虽然降低了各节点要掌握全局信息的要求, 但也使得问题求解受到网络特征的影响更加明显, 同时也使得分布式优化比传统优化要复杂得多, 需要在整个系统的通信、计算、控制三个方面做整合和协调. 为了完成共同的目标, 节点之间需要通过网络进行信息传输, 这必然耗费大量的网络资源. 因此, 如何在实际应用中设计有效的通信机制以降低资源消耗，逐渐成为分布式技术研究的热点. 另外, 由网络引起的通信约束, 如连接拓扑的变化、对分布式优化算法的设计、收敛性分析及性能都带来了很大的挑战.

1.1.1　多智能体系统的分布式凸优化

在优化理论的发展过程中, 有一类特别有趣的问题是分布式多智能体优化, 这类问题产生的动机来自于无线传感网络中估计环境参数或解决某些类似温度和源定位. 例如, 含有 N 个传感器的无线传感网络的参数估计问题[6] 最终可以转化为一个分布式优化问题, 其中目标函数 $f(\theta)$ 具有下面的形式:

$$f(\theta)=\frac{1}{N}\sum_{i=1}^{N}f_i(\theta),$$

其中, θ 是被估计的未知参数; $f_i(\theta)$ 是仅依赖于传感器 i 的测量数据的局部目标函数, $i=1,2,\cdots,N$. 例如, $f(\theta)$ 和 $f_i(\theta)$ 可以分别取作下面平均值函数:

$$f(\theta)=\frac{1}{mN}\sum_{i=1}^{N}\sum_{j=1}^{m}(x_{ij}-\theta)^2, f_i(\theta)=\frac{1}{m}\sum_{j=1}^{m}(x_{ij}-\theta)^2,$$

其中, x_{ij} 是在第 i 传感器得到的第 j 个测量值.

注意到以上给出的分布优化问题的特点是其目标函数是各个节点局部目标函数的和[6-31], 具体描述为: N 个相互协作的个体组成一个动态网络化系统, 这个系统可以用一个网络来表示, 每个具有决策能力的个体对应网络中的一个节点, 每个节点都有自己专属的目标函数 $f_i:\mathbb{R}^n\to\mathbb{R}$, 但整个网络的目标函数 $F(x)$ 是 N 个局部目标函数的和, 即 $F(x)=\sum_{i=1}^{N}f_i(x)$. 因此, 所有节点的共同的任务是找到一个最优值 x^*, 即来求解下面的优化问题:

$$x^*=\arg\min_x\sum_{i=1}^{N}f_i(x). \tag{1.1}$$

注意到该问题中每个节点 i 仅知道自己的目标函数 f_i, 因此对每个节点来说, 都不可能单独计算出 $F(x)$ 的最优值 x^*, 这就需要节点之间通过网络与其他节点之间相互交流信息, 并利用自身的信息和接收到的邻近节点的信息调整自身的状态, 协作完成整个网络系统的目标.

如果局部目标函数都是凸的, 则 $F(x)$ 也是凸的, 这时称式 (1.1) 为分布式凸优化问题. 像式 (1.1) 这样的无约束、可分离、凸一致性优化问题, 在传感网络上的参数估计、集群和密度估计、能源来源定位等[6,32-34] 方面都有广泛的应用. 例

如, 估计一个声源的位置[6] 是生物和军事应用中的一个重要的问题, 在这类问题中, 声源被描述为监测区域的一个未知的位置 θ. 假设传感器均匀地分布在一个边长大于等于 1 的正方体中或分布在一个立方体中, 每个传感器节点知道自身的位置 $r_i(i=1,2,\cdots,N)$, 然后对节点 i 的第 j 个信号强度进行测量, 应用等向性能量传播模型得

$$x_{ij}=\frac{A}{\|\theta-r_i\|^{\beta}}+\omega_{ij},$$

其中, A 是正常数; 对所有的节点 $i, \|\theta-r_i\|>1$; 指数 $\beta\geqslant 1$, 描述了声信号通过传播媒介的衰减特性; ω_{ij} 是独立同分布的零均值、方差为 σ^2 的高斯噪声. 则声源的最大似然估计是求解下面的优化问题:

$$\theta^*=\arg\min_{\theta}\frac{1}{mN}\sum_{i=1}^{N}\sum_{j=1}^{m}\left(x_{ij}-\frac{A}{\|\theta-r_i\|^{\beta}}\right)^2.$$

截至目前, 已经涌现了大量的分布式优化算法来求解优化问题式 (1.1). 下面将结合现有的几类经典的分布式优化算法, 简要分析分布式优化算法的研究现状.

1.1.2 几类经典的分布式优化算法

在分布式优化理论研究方面, 众多著名学者如 Nedić等 [20, 21]、Bertsekas 等 [35-37]、Nesterov [38] 、Boyd 等 [39] 和 Tsitsiklis 等 [40] 都做了大量的开创性工作, 他们的研究为分布式优化理论奠定了坚实的基础. 本小节将简要介绍三类分布式优化算法: 第一类是离散时间分布式次梯度算法; 第二类是连续时间和离散时间辅助变量算法; 第三类是连续时间零梯度和算法.

1. 分布式次梯度算法

按照现有的文献, 可以粗略地把次梯度算法分为两类: 一类是增量次梯度算法[6,16,18-21,32]; 另一类是一致性次梯度算法[15,17,22-31,33]. 在增量算法中, 节点之间需要存在一个环状的路径, 最优值的估计值按照环状的路径通过网络在节点之间传递直到得到最优解. 然而, 在网络中确定一个覆盖所有节点的循环路径是非常困难的, 尤其在一个分散的去中心化网络中. 另外, 当环状路径上的任何一条边出现故障时, 数据的传送就会中断, 算法的迭代过程就会停止. 一致性次梯度算法是一种

结合了经典次梯度算法和多智能体一致性算法[41-45]的网络优化算法, 一经提出就受到广泛的关注. 它有两种不同的形式, 其中一种是由 Nedić 等在文献 [17] 中首次提出的, 算法的描述如下:

$$x_i(k+1)=\sum_{j\in\mathcal{N}_i}a_{ij}x_j(k)-\alpha_i(k)d_i(k), \tag{1.2}$$

其中, k 是非负整数; $x_i(k)$ 是节点 i 在时刻 t_k 储存的最优值 x^* 的一个估计; a_{ij} 是对应于一个无向通信拓扑的行随机矩阵 A 的第 i 行第 j 列的元素; $\mathcal{N}_i$ 表示节点 i 可以接收到信息的节点的集合, 或称为节点 i 的邻居节点的集合; 实标量 $\alpha_i(k)$ 是迭代步长, 且满足 $\sum_{k=0}^{\infty}\alpha_i(k)=\infty$ 和 $\sum_{k=0}^{\infty}\alpha_i^2(k)<\infty$; 向量 $d_i(k)$ 是节点 i 的目标函数 f_i 在 $x_i(k)$ 处的一个次梯度, 即 $d_i(k)\in\partial f_i(x_i(k))$. 显然, 如果 f_i 是可微的, $d_i(k)$ 就是函数 f_i 在 $x_i(k)$ 处的梯度, 即 $d_i(k)=\nabla f_i(x_i(k))$.

注意到当所有的 f_i 都等于 0 时, 式 (1.1) 和式 (1.2) 就简化成一致性问题[41, 44]. 式 (1.2) 的迭代过程可以解释为两步: 第一步叫做一致性项, 节点 i 用自己当前的估计值和从周围邻居节点接收到的局部估计值组成一个凸组合 $\sum_{j\in\mathcal{N}_i}a_{ij}x_j(k)$, 组合的系数就是行随机矩阵 A 相应的元素 a_{ij}. 一致性项的作用在于处理节点对优化问题的信息不完整的不足. 第二步, 将一致性项得到的凸组合 $\sum_{j\in\mathcal{N}_i}a_{ij}x_j(k)$ 沿着负次梯度方向迭代更新. 以上迭代过程不断更新, 在梯度项 $\alpha_i(k)d_i(k)$ 的作用下, 所有节点信息或状态值 $x_i(k)$ 最终一致收敛于全局目标函数的最优值 x^*. 另外一种形式是由 Johansson 等在文献 [15] 中提出的. 这种形式的两个步骤和式 (1.2) 正好相反, 它先把对最优值的估计值沿着负次梯度方向迭代, 然后再与周围相邻节点分享迭代后的结果, 算法具体表示如下:

$$x_i(k+1)=\sum_{j\in\mathcal{N}_i}a_{ij}\big[x_j(k)-\alpha_j(k)d_j(k)\big], \tag{1.3}$$

其中, 向量 $d_j(k)$ 是节点 j 的目标函数 f_j 在 $x_j(k)$ 处的次梯度, 即 $d_j(k)\in\partial f_j(x_j(k))$, $\partial f_j(x_j(k))$ 表示函数 f_j 在点 $x_j(k)$ 处的次微分; 其他参数和符号的解释及满足的条件都和式 (1.2) 相同. 文献 [31] 对这两种不同形式的一致性次梯度算法在收敛误差和收敛速度等方面做了比较, 证明了式 (1.3) 比式 (1.2) 在性能上要好一些.

虽然以上的次梯度算法在比较弱的假设条件下可以求解优化问题式 (1.1), 但算法的收敛性却受到步长选择的牵制. 一般来说, 步长有三种选择方式: 常值步