# 分布式在线条件梯度优化算法

李德权 董 翘 周跃进

(安徽理工大学数学与大数据学院 安徽 淮南 232001)

**摘 要** 针对现有分布式在线优化算法所面临的高维约束难以计算的问题,提出一种分布式在线条件梯度优化算法(Distributed Online Conditional Gradient Optimization Algorithm,DOCG)。首先,通过多个体网络节点间的相互协作进行数据采集,并通过共享采集的信息更新局部估计,同时引入反映环境变化的局部即时损失函数。然后,该算法利用历史梯度信息进行加权平均,提出一种新的梯度估计方案,其用线性优化步骤替代投影步骤,避免了投影运算在高维约束时难以计算的问题。最后,通过分析表征在线估计性能的Regret界,证明了所提DOCG算法的收敛性。利用低秩矩阵填充问题进行仿真验证,结果表明,相比于现有分布式在线梯度下降法(DOGD),所提DOCG算法具有更快的收敛速度。

**关键词** 条件梯度,无投影,分布式网络,在线学习,Regret界

**中图法分类号** TP301.6 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.03.049

## Distributed Online Conditional Gradient Optimization Algorithm

LI De-quan DONG Qiao ZHOU Yue-jin

(School of Mathematics and Big Data,Anhui University of Science and Technology,Huainan,Anhui 232001,China)

**Abstract** In order to overcome the problem that the high-dimensional constraints in existing distributed online optimization algorithms are hard to be calculated,a distributed online conditional gradient optimization algorithm (DOCG) was proposed in this paper. Firstly,data collection is carried out through mutual cooperation among nodes of the muti-agent distributed network,and then each node updates its local iterate based on new local data,together with an instantaneous local cost functions that reflects the environmental changes. Secondly,by virtue of the historical gradient information for weighted averaging,a new gradient estimation scheme is proposed,in which the sophisticated projection step is replaced by the linear optimization step and thus avoids the disadvantages of the projection operator that is hard to be calculated. Finally,by defining the corresponding Regret bound to characterize the performance of online estimation,the convergence of the DOCG algorithm is proved. Simulation results are conducted on low-rank matrix completion problems,which clearly show that the proposed algorithm has faster convergence rate than the existing distributed online gradient method(DOGD).

**Keywords** Conditional gradient,Projection-free,Distributed network,Online learning,Regret bound

实际应用中,由众多价格低廉、广泛分布的小型设备通过局部信息传递耦合而成的分布式网络,能够方便地对海量数据进行分布式存储和计算。事实上,许多高维优化算法已被相继提出并用来处理具有海量数据集的优化问题,这些分布式优化算法允许节点或个体在网络中合作和共享计算资源,其优点是不仅可以有效降低网络中节点的数据传输速率,而且当出现局部故障时仍能确保系统的鲁棒性[1-2]。实际应用中,多个体网络系统通常处于动态变化和不确定的环境下,如编队控制、可再生能源系统的调度以及配电网络中的能量调度等,而已有的多个体网络分布式优化通常假定节点数据是静态的,且要等到网络中的所有节点数据都被收集后再进行数据处理,这种离线的学习方式会带来高昂的通信代价。在线学习是解决以上问题的一种常用方法,其本质是随机优化方法[3]的一种扩展。在线学习方法的优点是可以利用任意变化的代价函数来表示多个体网络系统的不确定性,同时可以方便地对网络节点的动态数据流进行实时处理。

文献[4]提出的条件梯度法(Frank-Wolfe,FW)已经成为有效求解高维度约束优化问题的热点算法。与经典的投影梯度(Projection Gradient,PG)算法[5-6]相比,FW算法由于具备无投影性质而更具吸引力。具体来说,每当PG算法进行一次优化运算时,都会导致当前的迭代点离开优化问题的可行域,从而得到一个不可行点,此时需要通过投影运算来恢复其可行性。而投影步骤意味着须在可行域内找到一点,且要求该点与当前不可行点的距离最短,这本质上等价于求解一个

到稿日期:2018-01-17 返修日期:2018-03-07 本文受国家自然科学基金资助项目(61472003),高校学科(专业)拔尖人才学术资助重点项目(gxbjZD2016049),安徽省学术和技术带头人及后备人选资助项目(2016H076),安徽省教育厅自然科学基金重点项目(KJ2017A087)资助。

**李德权**(1973-),男,博士,教授,硕士生导师,主要研究方向为分布式优化理论、统计机器学习;**董 翘**(1993-),女,硕士生,主要研究方向为分布式优化理论与应用,E-mail:1047869260@qq.com;**周跃进**(1977-),男,博士,副教授,主要研究方向为高维数据处理和统计机器学习。

凸二次规划问题。然而，在高维约束优化问题中，求解凸二次规划问题非常困难，这促使人们考虑对其进行有效的线性优化。FW 算法在分布式计算以及在线学习方面已有不少研究成果，但基于 FW 算法如何将两者进行有效结合的研究并不多见。此外，一些传统的优化算法，如投影梯度下降法、基于交替方向乘子法和对偶平均法等，近年来在分布式在线优化方面得到了较为广泛的应用。例如，文献[7-9]分别介绍了分布式在线梯度下降法(DOGD)、分布式在线交替方向乘子法(OD-ADMM)和分布式在线对偶平均(ODDA) 3 种方法。事实上，上述分布式在线学习算法已在大规模流数据处理问题方面得到成功应用。但这些算法中的投影运算所需的成本高昂，从而限制了它们在许多实际问题中的进一步应用。因此，需要将分布式条件梯度法拓展到在线学习情形，以解决投影运算所带来的计算成本高昂的问题。

分布式在线学习问题可分为优化计算阶段和一致性阶段两部分。在一致性阶段，网络中的节点或个体与其邻居节点通过局部通信交互进行信息共享。这样，经过 $T$ 次迭代后，所有节点逐步逼近整个网络的全局最优解。通常，节点主要采用以下两种形式达成一致性：1)多个体网络中的节点在优化阶段给出一个局部估计，然后利用分布式加权平均法更新局部估计，以寻找全局网络最优解；2)通过对网络中已有的边添加一致性约束，以等价的可分解形式重新构造优化问题，即如果节点 $i$ 和节点 $j$ 之间存在一条边，则 $x_i=x_j$。

基于上述考虑，本文采用一致性框架 1)，将文献[10]提出的集总式无投影在线学习算法拓展到分布式情形，提出一种分布式在线条件梯度优化算法。受分布式 Frank-Wolfe 优化算法[11]的启发，所提算法不同于分布式在线无投影算法[12]基于对偶平均的设计，其关键步骤是给出一种新的梯度估计方案，该方案主要利用历史信息来实现快速和准确的平均梯度估计，从而解决分布式的在线优化问题。

## 1　问题描述

如无特殊说明，本文所提到的向量均为列向量。$R^m$ 表示 $m$ 维列向量空间；$y^{\mathrm{T}}$ 表示向量 $y$ 的转置；$\langle y,x\rangle$ 表示向量 $y$ 和向量 $x$ 的内积；$\|y\|$ 表示向量 $y$ 的欧氏范数；$[x]_i$ 表示向量 $x$ 的第 $i$ 个分量；一个矩阵 $A\in\mathbb{R}^{n\times m}$，其中 $[A]_{ij}$ 表示第 $i$ 行第 $j$ 列的元素。

### 1.1　图论的相关知识

节点间的信息交互可被建模成图，并且可通过图论中的邻接矩阵或 Laplacian 矩阵[13]来刻画网络拓扑。本文考虑了由 $n$ 个节点构成的多个体网络，$V=\{1,2,\cdots,n\}$ 为节点集合。网络节点之间的通信可被建模成静态加权无向图 $G=(V,E,W)$。$E=\{(j,i)\mid i,j\in V\}$ 表示网络中所有无向边构成的集合。无向边 $(j,i)\in E$ 表示个体 $j$ 与个体 $i$ 互发信息，此时无向边 $(j,i)$ 的边权 $w_{ji}>0$，$w_{ji}\in W$，且称个体 $j$ 是个体 $i$ 的入度邻居，否则 $w_{ji}=0$。由此定义个体 $i$ 的入度邻居集合为 $N(i)=\{j\in V\mid(j,i)\in E\}$。类似地，可以定义个体 $i$ 的出度邻居集合。一个图 $G$ 被认为是无向的，如果任意时刻都有 $(i,j)\in E$，并且 $(j,i)\in E$。对于无向图来说，其入度邻居集合与出度邻居集合是相同的，这意味着无向图 $G$ 也为平衡图。将图 $G$ 的邻接矩阵记为 $A(G)$，当 $(j,i)\in E$ 时，$[A(G)]_{ji}=w_{ji}$，否则 $[A(G)]_{ji}=0$。图的拉普拉斯矩阵 $L(G)=\Delta(G)-A(G)$，其中 $\Delta(G)$ 为对角矩阵，对角元素为对应个体的入度，记作 $d_i$，且 $d_i=\sum_{\{j\mid(j,i)\in E\}}w_{ij}$。本文研究的多个体网络拓扑有以下性质。

**性质 1**[14]　如果 $G$ 是平衡图，定义矩阵 $P=I-\frac{1}{\varepsilon}L(G)$，其中 $\varepsilon=d_{\max}+1$，$d_{\max}=\max_{i\in V}d_i$，那么 $P$ 为双随机矩阵。

### 1.2　Regret 界

在线优化中，在时刻 $t$，每个节点 $i(i\in V)$ 从决策集 $\kappa_i$ 中选择一个状态 $x_{i,t}$ 作为局部估计，决策集 $\kappa_i$ 是 $\mathbb{R}^n$ 的一个封闭有界的子集。提交决策后，节点 $i$ 方可观测事先不知道的局部即时损失函数 $f_{i,t}(x_{i,t})$，其中约束集 $\kappa$ 为凸集，$f_i:\kappa_i\to\mathbb{R}$ 是凸函数。

本文考虑如下的分布式在线优化问题：

$$f_t(x)=\frac{1}{n}\sum_{i}^{N}f_{i,t}(x)\tag{1}$$

其中，向量 $x(x\in\kappa)$ 是所有节点局部估计的集合，即 $x=(x_1,\cdots,x_n)\in\kappa$，$x_i\in\mathbb{R}^m$，$\kappa=\kappa_1\times\cdots\times\kappa_n\subseteq\mathbb{R}^m$。节点 $i$ 仅知道局部损失函数 $\{f_{i,t}\}_{t\geqslant1}$，$f_{i,t}(x_i):\mathbb{R}^m\to\mathbb{R}$。

在线优化的目的是缩小 Regret 界——随着时间推移的累积成本与最佳固定决策所产生的成本之间的差额。目前已有研究定义了两类 Regret 界。第一类是一般情况下的通用 Regret 界：

$$R_T(x_i,x)=\sum_{j=1}^{n}\sum_{t=1}^{T}f_{j,t}(x_{i,t})-\sum_{j=1}^{n}\sum_{t=1}^{T}f_{j,t}(x)\tag{2}$$

文献[8]定义了第二类 Regret 界——加权平均 Regret 界：

$$R_T(\bar{x}_i,x)=\sum_{j=1}^{n}\sum_{t=1}^{T}(f_{j,t}(\bar{x}_{i,t})-f_{j,t}(x))\tag{3}$$

若当迭代次数 $T$ 趋于无穷时，$R(T)/nT$ 趋于 0，这意味着在线学习算法的解在渐近意义上收敛到全局网络最优解，即 $R(T)=o(T)$。

## 2　分布式在线条件梯度算法

本节将文献[10]中的无投影在线优化算法拓展到分布式情形，提出分布式在线条件梯度算法(DOCG)来解决式(1)中的优化问题。

算法 1 给出了分布式在线条件梯度法的具体步骤。与文献[12]不同，所提算法不需要引入对偶变量，同时也不再利用原始梯度 $g_{i,t}$ 信息。算法 1 中第 10－12 行表示更新各节点局部估计时，只需直接采用新的梯度估计方案[11]，具体实现步骤见算法 1 第 8－9 行。

**算法 1**　DOCG 算法

1. 输入：最大迭代次数 T，聚合参数$\{\eta_t\}$和步长参数$\{\gamma_{i,t}\}$；
2. 初始化本地变量：$x_{i,1}\in\kappa$，$\forall i\in V$；
3. for t=1,…,T do
4. 　观测局部即时损失函数 $f_t(t)=\{f_{i,t}(t);\text{for}\,\forall i\in V\}$；
5. 　计算局部估计 $x_{i,t}$ 的加权平均值 $\bar{x}_{i,t}$，$\bar{x}_{i,t}=\sum_{j=1}^{n}P_{ij}x_{j,t}$；
6. 　计算次梯度 $g_{i,t}\in\partial f_{i,t}(\bar{x}_{i,t})$，$\forall i\in V$；
7. 　for 每一个节点 $i\in V$ do
8. 　　$\hat{g}_{i,t}=\bar{g}_{i,t-1}+g_{i,t}-g_{i,t-1}$(当 t=1 时，$\bar{g}_{i,0}=g_{i,0}=0$)；
9. 　　计算加权平均次梯度 $\bar{g}_{i,t}=\sum_{j=1}^{n}P_{ij}\hat{g}_{j,t}$；

10. $F_{i,t}(x)=\eta_i\langle \bar{g}_{i,t},x\rangle+\|x\|^2$;
11. $v_{i,t}=\arg\min_{x\in\kappa}\{\langle \nabla F_{i,t}(\bar{x}_{i,t}),x\rangle\}$;
12. 输出:更新局部估计 $x_{i,t+1}=\bar{x}_{i,t}+\gamma_{i,t}(v_{i,t}-\bar{x}_{i,t})$;
13. end for
14. end for

## 3 收敛性分析

本节分析 DOCG 算法在本地损失函数 $f_{i,t}$ 为凸函数时的第二类加权平均 Regret 界。分析思路如下:首先,对 Regret 界的 $\sum_{t=1}^{T}(f_{j,t}(\bar{x}_{i,t})-f_{j,t}(x))$ 部分添加辅助变量以进行分解,可将其分解为 $\sum_{t=1}^{T}|f_{j,t}(x_{i,t}^*)-f_{j,t}(x)|$ 和 $\sum_{t=1}^{T}|f_{j,t}(\bar{x}_{i,t})-f_{j,t}(x_{i,t}^*)|$ 两项;其次,证明这两项都有上界 $O(T^{\frac{3}{4}})$(见 4.1 节)。

分析过程中,根据文献[10]的定义,假定算法 1 第 10 行的最优点为 $x_{i,t}^*=\arg\min_{x\in\kappa}F_{i,t}(x)$,将算法 1 第 4 行的局部即时损失函数转换为以下形式:$\tilde{f}_{i,t}(x)=f_{i,t}(x+\bar{x}_{i,t}-x_{i,t}^*)$。文献[15]分析了任意网络拓扑的平均一致性问题,发现算法的平均收敛时间依赖于表征算法的双随机矩阵的第二大奇异值,即性质 1 所提的双随机矩阵 $P$ 的第二大奇异值 $\sigma_2(P)$,由此引出谱隙定义 $\gamma(P)=1-\sigma_2(P)$。本文所提算法的收敛速度也依赖于上述谱隙,网络的连通性越好,则谱隙值越大。为进行进一步分析,给出以下假设。

假设 1 1)每个局部即时损失函数 $f_{i,t}$ 对于 $L_2$ 范数是 Lipschitz 连续的,即 $\forall x,y\in\kappa$,$|f_{i,t}(x)-f_{i,t}(y)|\leqslant L\|x-y\|$。

2)约束集 $\kappa$ 的欧氏空间的直径上界为 $D$,即 $\forall x,y\in\kappa$,$\|x-y\|\leqslant D$。

3)每个局部即时损失函数 $f_{i,t}$ 是 $\beta$-smooth 的,且满足 $\sigma$-强凸性,即

$$f_{i,t}(y)\leqslant f_{i,t}(x)+\langle\nabla f_{i,t}(x),y-x\rangle+\frac{\beta}{2}\|y-x\|^2$$

$$f_{i,t}(y)\geqslant f_{i,t}(x)+\langle\nabla f_{i,t}(x),y-x\rangle+\frac{\sigma}{2}\|y-x\|^2$$

### 3.1 辅助引理

本节将证明一系列辅助引理。引理 1 主要将 $\sum_{t=1}^{T}|f_{j,t}(x_{i,t}^*)-f_{j,t}(x)|$ 分解成两项,引理 6 和引理 4 将进一步给出其对应的上界。

**引理 1**[12] 对于 $\forall i,j\in V$ 及 $\forall x\in\kappa$,下式成立:

$$\sum_{t=1}^{T}(f_{j,t}(x_{i,t}^*)-f_{j,t}(x))\leqslant\sum_{t=1}^{T}(\tilde{f}_{j,t}(x_{i,t}^*)-\tilde{f}_{j,t}(x))+2L\sum_{t=1}^{T}\|\bar{x}_{j,t}-x_{j,t}^*\| \tag{4}$$

上述引理 1 中迭代式(4)不等号右端由以下两项构成:$\sum_{t=1}^{T}(\tilde{f}_{j,t}(x_{i,t}^*)-\tilde{f}_{j,t}(x))$ 和 $2L\sum_{t=1}^{T}\|\bar{x}_{j,t}-x_{j,t}^*\|$。下面将分别对这两项进行分析,并建立其上界。为此,令 $h_{i,t}(x)=F_{i,t}(x)-F_{i,t}(x_{i,t}^*)$,$h_{i,t}=h_{i,t}(\bar{x}_{i,t})$。引理 2 将给出 $h_{i,t+1}$ 和 $h_{i,t}$ 的递推关系。

**引理 2** 对于 $\forall i\in V$,$\forall t=0,1,\cdots,T$,得到 $h_{i,t+1}$ 和 $h_{i,t}$ 的递推:

$$h_{i,t+1}\leqslant(1-\gamma_{i,t})h_{i,t}+\gamma_{i,t}^2D^2+\eta_i\|\bar{g}_{i,t+1}-\bar{g}_{i,t}\|\sqrt{h_{i,t+1}} \tag{5}$$

证明:由 $h_{i,t}(x)$ 与 $\bar{x}_{i,t+1}$ 的定义和 $F_{i,t}(x)$ 是 2-smooth 可得:

$$\begin{aligned}h_{i,t}(\bar{x}_{i,t+1})&=F_{i,t}(\bar{x}_{i,t}+\gamma_{t,i}(v_{i,t}-\bar{x}_{i,t}))-F_{i,t}(x_{i,t}^*)\\&\leqslant F_{i,t}(\bar{x}_{i,t})-F_{i,t}(x_{i,t}^*)+\gamma_{i,t}\langle\nabla F_{i,t}(\bar{x}_{i,t}),v_{i,t}-\bar{x}_{i,t}\rangle+\gamma_{i,t}^2\|v_{i,t}-\bar{x}_{i,t}\|^2\\&\leqslant F_{i,t}(\bar{x}_{i,t})-F_{i,t}(x_{i,t}^*)+\gamma_{i,t}\langle\nabla F_{i,t}(\bar{x}_{i,t}),v_{i,t}-\bar{x}_{i,t}\rangle+\gamma_{i,t}^2D^2\end{aligned} \tag{6}$$

由算法 1 第 11 行可知,$v_{i,t}$ 是可行域内与 $\nabla F_{i,t}(\bar{x}_{i,t})$ 内积取值最小的向量,根据其最优性可得:

$$\langle\nabla F_{i,t}(\bar{x}_{i,t}),v_{i,t}\rangle\leqslant\langle\nabla F_{i,t}(\bar{x}_{i,t}),x_{i,t}^*\rangle \tag{7}$$

另外,由 $F_{i,t}(x)$ 的凸性可得:

$$\langle\nabla F_{i,t}(\bar{x}_{i,t}),x_{i,t}^*-\bar{x}_{i,t}\rangle\leqslant F_{i,t}(x_{i,t}^*)-F_{i,t}(\bar{x}_{i,t}) \tag{8}$$

将式(7)、式(8)代入式(9),得:

$$\begin{aligned}h_{i,t}(\bar{x}_{i,t+1})&\leqslant F_{i,t}(\bar{x}_{i,t})-F_{i,t}(x_{i,t}^*)+\gamma_{i,t}(F_{i,t}(x_{i,t}^*)-F_{i,t}(\bar{x}_{i,t}))+\gamma_{i,t}^2D^2\\&=(1-\gamma_{i,t})(F_{i,t}(\bar{x}_{i,t})-F_{i,t}(x_{i,t}^*))+\gamma_{i,t}^2D^2\\&=(1-\gamma_{i,t})h_{i,t}+\gamma_{i,t}^2D^2\end{aligned} \tag{9}$$

由 $h_{i,t}(x)$ 的定义和 $x_{i,t+1}^*$ 的最优性,进一步得到:

$$\begin{aligned}h_{i,t+1}&=F_{i,t+1}(\bar{x}_{i,t+1})-F_{i,t+1}(x_{i,t+1}^*)\\&=F_{i,t}(\bar{x}_{i,t+1})-F_{i,t}(x_{i,t+1}^*)+(F_{i,t+1}(\bar{x}_{i,t+1})-F_{i,t}(\bar{x}_{i,t+1}))-(F_{i,t}(x_{i,t+1}^*)-F_{i,t}(x_{i,t+1}^*))\\&\leqslant F_{i,t}(\bar{x}_{i,t+1})-F_{i,t}(x_{i,t}^*)+(F_{i,t+1}(\bar{x}_{i,t+1})-F_{i,t}(\bar{x}_{i,t+1}))-(F_{i,t+1}(x_{i,t+1}^*)-F_{i,t}(x_{i,t+1}^*))\end{aligned} \tag{10}$$

在此基础上,结合算法 1 第 10 行对函数 $F_{i,t}(x)$ 的定义可得:

$$F_{i,t+1}(x)-F_{i,t}(x)=\eta_i\langle\bar{g}_{i,t+1}-\bar{g}_{i,t},x\rangle \tag{11}$$

根据 $h_{i,t}$ 的定义,将式(11)代入式(10),得:

$$\begin{aligned}h_{i,t+1}&\leqslant h_{i,t}(\bar{x}_{i,t+1})+\eta_i\langle\bar{g}_{i,t+1}-\bar{g}_{i,t},\bar{x}_{i,t+1}\rangle-\eta_i\langle\bar{g}_{i,t+1}-\bar{g}_{i,t},x_{i,t+1}^*\rangle\\&=h_{i,t}(\bar{x}_{i,t+1})+\eta_i\langle\bar{g}_{i,t+1}-\bar{g}_{i,t},\bar{x}_{i,t+1}-x_{i,t+1}^*\rangle\\&\leqslant h_{i,t}(\bar{x}_{i,t+1})+\eta_i\|\bar{g}_{i,t+1}-\bar{g}_{i,t}\|\|\bar{x}_{i,t+1}-x_{i,t+1}^*\|\end{aligned} \tag{12}$$

因此,当 $F_{i,t}(x)$ 为 2-强凸性时(见假设 1):

$$\|x-x_{i,t}^*\|^2\leqslant F_{i,t}(x)-F_{i,t}(x_{i,t+1}^*) \tag{13}$$

同理可得:

$$\begin{aligned}\|\bar{x}_{i,t+1}-x_{i,t+1}^*\|^2&\leqslant F_{i,t+1}(\bar{x}_{i,t+1})-F_{i,t+1}(x_{i,t+1}^*)\\&=h_{i,t+1}\end{aligned} \tag{14}$$

将式(14)两边开平方,得:

$$\|\bar{x}_{i,t+1}-x_{i,t+1}^*\|\leqslant\sqrt{h_{i,t+1}} \tag{15}$$

由迭代式(6)—式(15)易得式(5)。

为了使上述递归更加具体,需要给出迭代式(14)中偏差项 $\|\bar{g}_{i,t+1}-\bar{g}_{i,t}\|$ 的上界。引理 3 给出了该偏差项上界的详细推理过程。

**引理 3** 考虑由算法 1 第 8—9 行定义的新梯度 $\bar{g}_{i,t}$。若 $\forall i\in V$,对于任意时刻 $t$,均有:

$$\|\bar{g}_{i,t}-\bar{g}_{i,t-1}\|\leqslant\left(\frac{1+\sigma_2(P)}{1-\sigma_2(P)}\sqrt{n}+1\right)\nabla p_{t-1} \tag{16}$$

进一步地，由文献[12]可知，$\{\nabla p_{t-1}\}_{t\geqslant 1}$ 是非负递减序列，并且 $\| g_{i,t}-g_{i,t-1}\| \leqslant \nabla p_{t-1}=\mathrm{O}(\frac{1}{(t-1)^{\alpha}})$ 成立。

证明：根据算法 1 第 8－9 行对历史梯度及当前梯度的定义可知：

$$\bar{g}_{i,t}=\sum_{j=1}^{N}P_{ij}(\bar{g}_{j,t-1}+g_{j,t}-g_{j,t-1}) \tag{17}$$

其中，定义 $P^r$ 表示 $P$ 的第 $r$ 次幂，$P_{ij}^r$ 表示矩阵 $P^r$ 第 $i$ 行第 $j$ 列的元素。经代数运算，可得以下递推关系式：

$$\bar{g}_{i,t}=\sum_{j=1}^{n}P_{ij}^{t-s+1}\bar{g}_{j,s-1}+\sum_{r=s}^{t}\sum_{i=1}^{n}P_{ij}^{t-r+1}(g_{i,r}-g_{i,r-1}) \tag{18}$$

令 $\bar{g}_{j,0}=0$，当 $s=1$ 时，式(18)变为：

$$\bar{g}_{i,t}=\sum_{r=1}^{t}\sum_{i=1}^{n}P_{ij}^{t-r+1}(g_{i,r}-g_{i,r-1}) \tag{19}$$

假设迭代过程中 $P^0$ 为单位矩阵 $I_n$，由式(19)可推出 $\bar{g}_{i,t}-\bar{g}_{i,t-1}$ 的等式关系：

$$\begin{aligned}\bar{g}_{i,t}-\bar{g}_{i,t-1}&=\sum_{r=1}^{t}\sum_{i=1}^{n}P_{ij}^{t-r+1}(g_{i,r}-g_{i,r-1})-\sum_{r=1}^{t-1}\sum_{i=1}^{n}P_{ij}^{t-r}(g_{i,r}-g_{i,r-1})\\&=\sum_{r=1}^{t}\sum_{i=1}^{n}P_{ij}^{t-r+1}(g_{i,r}-g_{i,r-1})-\sum_{r=1}^{t}\sum_{i=1}^{n}P_{ij}^{t-r}(g_{i,r}-g_{i,r-1})+\sum_{i=1}^{n}P_{ij}^{0}(g_{i,t}-g_{i,t-1})\\&=\sum_{r=1}^{t}\sum_{i=1}^{n}(P_{ij}^{t-r+1}-P_{ij}^{t-r})(g_{i,r}-g_{i,r-1})+\sum_{i=1}^{n}P_{ij}^{0}(g_{i,t}-g_{i,t-1})\end{aligned} \tag{20}$$

又因为 $\| g_{i,t}-g_{i,t-1}\| \leqslant \nabla p_{t-1}$，由范数的性质以及双随机权重矩阵 $P$ 的对称性，可以得到偏差项 $\| \bar{g}_{i,t}-\bar{g}_{i,t-1}\|$ 的上界：

$$\begin{aligned}\| \bar{g}_{i,t}-\bar{g}_{i,t-1}\| &=\| \sum_{r=1}^{t}\sum_{i=1}^{n}(P_{ij}^{t-r+1}-P_{ij}^{t-r})(g_{i,r}-g_{i,r-1})+\sum_{i=1}^{n}(g_{i,t}-g_{i,t-1})\|\\&\leqslant\sum_{r=1}^{t}\sum_{i=1}^{n}\| (P_{ij}^{t-r+1}-P_{ij}^{t-r})\| \| (g_{i,r}-g_{i,r-1})\| +\sum_{i=1}^{n}\| (g_{i,t}-g_{i,t-1})\|\\&\leqslant\sum_{r=1}^{t}\| P_i^{t-r+1}-P_i^{t-r}\|_1\nabla p_{t-1}+\nabla p_{t-1}\end{aligned} \tag{21}$$

其中，$P_i^r$ 表示矩阵 $P^r$ 的第 $i$ 列元素。

为了给出式(21)中 $L_1$ 范数的上界，进一步引入一个列向量 1，该列向量的元素全为 1，且其维数与 $P_i^r$ 相同。注意到：

$$\begin{aligned}&\sum_{r=1}^{t}\| P_i^{t-r+1}-P_i^{t-r}\|_1\\&=\sum_{r=1}^{t}\| (P_i^{t-r+1}-\frac{1}{n})-(P_i^{t-r}-\frac{1}{n})\|_1\\&\leqslant\sum_{r=1}^{t}(\| (P_i^{t-r+1}-\frac{1}{n})\|_1-\| P_i^{t-r}-\frac{1}{n}\|_1)\\&\leqslant\sum_{r=1}^{t}(\sigma^{t-r+1}(P)-\sigma^{t-r}(P))\sqrt{n}\\&=\frac{(1-\delta_2^t(P))(1+\delta_2(P))}{1-\delta_2(P)}\sqrt{n}\\&\leqslant\frac{1+\delta_2(P)}{1-\delta_2(P)}\sqrt{n}\end{aligned} \tag{22}$$

将式(22)代入式(21)，即可得式(16)。

引理 4 给出引理 1 中迭代式(4)中不等式右端第二项 $\sum_{t=1}^{T}\| \bar{x}_{j,t}-x_{j,t}^*\|$ 的上界为 $\mathrm{O}(\sqrt{T^{\frac{3}{4}}})$。

**引理 4**[12]　对于 $\forall i\in V$，$\bar{x}_{i,t}$ 和 $x_{i,t}^*$ 的关系满足：

$$\sum_{t=1}^{T}\| \bar{x}_{i,t}-x_{i,t}^*\| \leqslant\frac{8}{3}DT^{\frac{3}{4}} \tag{23}$$

**引理 5**　对于 $\forall i\in V$，$\forall t=0,1,\cdots,T$，$\bar{g}_{i,t}$ 和 $g_{\mathrm{avg}}$ 的关系满足：

$$\| \bar{g}_{i,t}-g_{\mathrm{avg}}\| \leqslant\frac{\sigma_2(P)}{1-\sigma_2(P)}\sqrt{n}\nabla p_{t-1} \tag{24}$$

其中，$g_{\mathrm{avg}}$ 表示网络中所有节点的平均梯度，并且 $g_{\mathrm{avg}}=\frac{1}{n}\sum_{i=1}^{n}g_{i,t}$。

证明：考虑迭代式(17)和式(18)得到 $\bar{g}_{i,t}$，结合 $g_{\mathrm{avg}}$ 的定义可以得出：

$$\bar{g}_{i,t}-g_{\mathrm{avg}}=\sum_{s=1}^{t}\sum_{i=1}^{n}(\frac{1}{n}-P_{ij}^{t-s+1})(g_{i,s}-g_{i,s-1}) \tag{25}$$

采用与引理 3 相同的推导方法，可以类似地得到偏差项 $\| \bar{g}_{i,t}-g_{\mathrm{avg}}\|$ 的一个上界：

$$\begin{aligned}\| \bar{g}_{i,t}-g_{\mathrm{avg}}\| &=\| \sum_{s=1}^{t}\sum_{i=1}^{n}P_{ij}^{t-s+1}(g_{i,s}-g_{i,s-1})\|\\&\leqslant\sum_{s=1}^{t}\sum_{i=1}^{n}\| P_i^{t-s+1}\| \| (g_{i,s}-g_{i,s-1})\|\\&\leqslant\sum_{s=1}^{t}\| P_i^{t-s+1}-\frac{1}{n}\|_1\nabla p_{t-1}\\&\leqslant\sum_{s=1}^{t}\sigma_2^{t-s+1}(P)\sqrt{n}\nabla p_{t-1}\\&\leqslant\frac{\sigma_2(P)}{1-\sigma_2(P)}\sqrt{n}\nabla p_{t-1}\end{aligned} \tag{26}$$

引理 6 给出引理 1 中迭代式(4)中不等式右端第一项 $\sum_{t=1}^{T}(\tilde{f}_{j,t}(x_{i,t}^*)-\tilde{f}_{j,t}(x))$ 的上界。

**引理 6**　记算法 1 中第 10 行中的正则项为 $\psi(x)=\frac{1}{2}\| x\| ^2$，并且设置聚合参数 $\eta_1=\cdots=\eta_N=\eta$，$\forall i\in V$，令 $\alpha_{i,t}=\eta$，转换后带有欧氏范数的局部即时损失函数 $\tilde{f}_{t,j}(x)$ 满足以下 Regret 界：

$$R_T^a(x_i,x)\leqslant\frac{L^2}{2}T\eta+\frac{3\sigma_2(P)}{1-\sigma_2(P)}\sqrt{n}\nabla p_{t-1}+\frac{1}{\eta}D^2 \tag{27}$$

证明：参考文献[12]中引理 7 的证明，利用本文引理 5 给出的式(24)，可推导出 $\tilde{f}_{t,j}(x)$ 的 Regret 界为：

$$\begin{aligned}R_T^a(x,x_i)&\leqslant\frac{L^2}{2}\sum_{t=1}^{T-1}\alpha_{i,t}+\frac{1}{\eta}\psi(x)+L\sum_{t=1}^{T}\alpha_{i,t}(\| \bar{g}_{i,t}-g_{\mathrm{avg}}\| +\frac{2}{n}\sum_{i=1}^{n}\| \bar{g}_{j,t}-g_{\mathrm{avg}}\| )\\&\leqslant(\frac{L}{2}+\frac{\sqrt{n}\sigma_2(P)}{1-\sigma_2(P)}\nabla p_{t-1}+\frac{2\sqrt{n}\sigma_2(P)}{1-\sigma_2(P)}\nabla p_{t-1})LT\eta+\frac{1}{\eta}D^2\\&\leqslant(\frac{L}{2}+\frac{3\sqrt{n}\sigma_2(P)}{1-\sigma_2(P)}\nabla p_{t-1})LT\eta+\frac{1}{\eta}D^2\end{aligned} \tag{28}$$

### 3.2　分布式在线 Regret 界

基于上节的准备知识，本节将给出主要结论，即解析地给出分布式在线条件梯度算法 $R_T(\bar{x}_i,x)$ 的上界。进一步地，当局部即时损失函数是凸函数时，可证明 DOCG 算法是收敛的，并给出其收敛速度。

**定理 1**　在算法 1 即 DOCG 算法中，若取 $\eta_i=$

$\frac{1-\sigma_2(P)D}{2(\sqrt{n}+1+(\sqrt{n}-1)\sigma_2(P))\nabla p_{t-1}T^{\frac{3}{4}}}$，$\gamma_{i,t}=\frac{1}{\sqrt{t}}$，则对于 $\forall i\in V$，$\forall t=0,1,\cdots,T$，加权平均 Regret 界满足：

$$R_T(\bar{x}_i,x)\leqslant 8nLDT^{\frac{3}{4}}+\frac{L^2}{2}T\eta+\frac{3\sqrt{n}\sigma_2(P)}{2(\sqrt{n}+1+(\sqrt{n}-1)\sigma_2(P))}LDT^{\frac{1}{4}}+\frac{2(\sqrt{n}+1+(\sqrt{n}-1)\sigma_2(P))}{1-\sigma_2(P)}LDT^{\frac{3}{4}} \quad (29)$$

证明：由迭代式(3)可知，加权平均 Regret 界可改写为 $R_T(\bar{x}_i,x)=\sum_{j=1}^{n}\sum_{t=1}^{T}f_{j,t}(\bar{x}_{i,t})-\sum_{j=1}^{n}\sum_{t=1}^{T}f_{j,t}(x)$。而由引理 1 可知加权平均 Regret 界中 $\sum_{t=1}^{T}(f_{j,t}(\bar{x}_{i,t})-f_{j,t}(x))$ 部分可分解成如下形式：

$$\begin{aligned}&\sum_{t=1}^{T}(f_{j,t}(\bar{x}_{i,t})-f_{j,t}(x))\\&=\sum_{t=1}^{T}(f_{j,t}(\bar{x}_{i,t})-f_{j,t}(x_{i,t}^*)+f_{j,t}(x_{i,t}^*)-f_{j,t}(x))\\&\leqslant\sum_{t=1}^{T}|f_{j,t}(x_{i,t}^*)-f_{j,t}(x)|+\sum_{t=1}^{T}|f_{j,t}(\bar{x}_{i,t})-f_{j,t}(x_{i,t}^*)|\\&\leqslant\sum_{t=1}^{T}|\tilde{f}_{j,t}(x_{i,t}^*)-\tilde{f}_{j,t}(x)|+2L\sum_{t=1}^{T}\|\bar{x}_{j,t}-x_{j,t}^*\|+L\sum_{t=1}^{T}\|\bar{x}_{i,t}-x_{i,t}^*\| \quad (30)\end{aligned}$$

根据式(30)，结合引理 2－引理 6，给出 $R_T(\bar{x}_i,x)$ 的上界：

$$\begin{aligned}R_T(\bar{x}_i,x)&=\sum_{j=1}^{n}\sum_{t=1}^{T}(f_{j,t}(\bar{x}_{i,t})-f_{j,t}(x))\\&\leqslant\sum_{j=1}^{n}\sum_{t=1}^{T}|\tilde{f}_{j,t}(x_{i,t}^*)-\tilde{f}_{j,t}(x)|+2L\sum_{j=1}^{n}\sum_{t=1}^{T}\|\bar{x}_{j,t}-x_{j,t}^*\|+nL\sum_{t=1}^{T}\|\bar{x}_{i,t}-x_{i,t}^*\|\\&\leqslant\frac{1}{\eta}D^2+\frac{3\sqrt{n}\sigma_2(P)}{1-\sigma_2(P)}\nabla p_{t-1}LT\eta+8nLDT^{\frac{3}{4}} \quad (31)\end{aligned}$$

将式(23)和式(27)代入式(31)，即可得式(29)成立。

$$\begin{aligned}R_T(\bar{x}_i,x)&\leqslant 8nLDT^{\frac{3}{4}}+R_T^a(x_i,x)\\&\leqslant 8nLDT^{\frac{3}{4}}+\frac{L^2}{2}T\eta+\frac{3\sqrt{n}\sigma_2(P)}{2(\sqrt{n}+1+(\sqrt{n}-1)\sigma_2(P))}LDT^{\frac{1}{4}}+\frac{2(\sqrt{n}+1+(\sqrt{n}-1)\sigma_2(P))}{1-\sigma_2(P)}LDT^{\frac{3}{4}}\end{aligned}$$

迭代式(29)表明，在时间平均意义下，DOCG 算法具有 $O(T^{\frac{3}{4}})$ 的 Regret 界。

## 4 数值实验

本节将本文所提 DOCG 算法与最早提出的经典分布式在线梯度下降法(DOGD)[7] 进行数值实验比较，以验证 DOCG 算法的性能。为了更好地进行对比，两种分布式在线优化算法均采用相同的参数，如步长参数、聚合参数等。

近年来，矩阵填充问题(Matrix Completion)已成为机器学习的一个研究热门。文献[16-17]给出了矩阵填充问题的数学原理，本文主要基于文献[17]提出的低秩矩阵填充问题来进行数值仿真验证。

网络中的节点观测到一个不完整的矩阵 $X_{\text{true}}$，其维数为 $h\times k$。在 $t$ 时刻，第 $i$ 个节点从训练集 $\Omega_{i,t}\subset[ht]\times[kt]$ 中获得带有噪声的观测值 $Y_{ht,kt}=[X_{\text{true}}]_{ht,kt}+Z_{ht,kt}$，$\forall(ht,kt)\in\Omega_{i,t}$，恢复完整矩阵 $X_{\text{true}}$ 即为低秩矩阵填充问题。

实验中使用一个包含 300 个训练样本和 300 个测试样本的数据集。训练样本在一个由 15 个节点组成的网络中通过在线训练的方式完成训练，训练时长 $T=150$，300 个训练样本等分给 15 个节点。本文多个体网络采用随机图，且网络中每条边的连接概率为 0.3。

在 $t$ 时刻，节点 $i$ 的即时损失函数是 $f_{i,t}([X]_{ht,kt})=(1/\sigma_{i,t}^2)\cdot([X]_{ht,kt}-Y_{ht,kt})^2$，$Y_{ht,kt}$ 为 $t$ 时刻节点 $i$ 对未知矩阵 $X_{\text{true}}$ 的带有噪声的观测值，噪声 $Z_{ht,kt}=p_{ht,kt}\cdot\tilde{Z}_{ht,kt}$，其中 $t$ 时刻 $p_{ht,kt}$ 服从伯努利分布 $P(p_{ht,kt}=1)=0.1$，且 $\tilde{Z}_{ht,kt}$ 服从 N(0,1)的正态分布，$\sigma_{i,t}^2$ 为噪声的方差，$\kappa=\|X\|_{tr}\leqslant R$。在本文的数值实验中，预先选定一个秩 $K=3$ 的矩阵 $X_{\text{true}}$，其维数 $h=20$，$k=30$，$X_{\text{true}}=\sum_{i=1}^{K}u_iv_i^{\mathrm{T}}/K$，$u_i$ 和 $v_i$ 是独立同分布的，且均服从N(0,1)的正态分布，取 $R=2\|X_{\text{true}}\|_{tr}$。训练任务是将 300 个测试样本的均方误差(MSE)收敛到 0 的邻域内，$MSE\triangleq|\Omega_{\text{test}}|^{-1}\sum_{(ht,kt)\in\Omega_{\text{test}}}|[X_{\text{true}}]_{ht,kt}-[\hat{X}]_{ht,kt}|^2$，其中 $\hat{X}$ 表示算法产生的估计值。为更好地比较两种分布式在线学习算法，定义两种性能指标：1)平均损失 $\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}(f_{i,t}(\bar{x}_{i,t})-f_{i,t}(x^*))$，它将 $\bar{x}_{i,t}$ 代入即时损失函数，然后在时间段 1 到 $T$ 上进行平均；2)测试样本的均方误差 $MSE$。

两种分布式在线学习算法的平均损失和测试集的均方误差 $MSE$ 如图 1 和图 2 所示。图 1 的纵坐标表示的平均损失以 10 为底取对数。图 1 清楚地表明了 DOCG 算法在开始迭代时收敛速度不及 DOGD 算法，但随着迭代次数的增加，收敛速度明显优于 DOGD 算法。这一现象表明，虽然 DOCG 的 Regret 界 $O(T^{\frac{3}{4}})$ 比 DOGD 的 Regret 界 $O(\sqrt{T})$ 高，且 DOCG 算法的迭代次数也在不断增加，但与其每次迭代的低计算成本相比，该算法总体上仍不失为一种快速算法。图 2 则进一步表明，测试样本在 DOCG 和 DOGD 两种算法的作用下，均方误差 $MSE$ 均收敛到 0 的邻域内，但 DOCG 具有更小的均方误差 $MSE$。

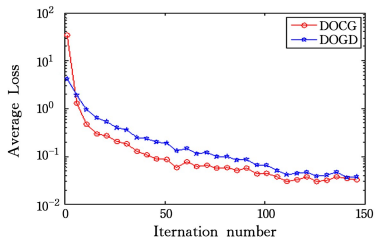


图 1 DOCG 与 DOGD 算法平均损失的比较

Fig. 1 Comparison of average loss of DOCG and DOGD algorithms

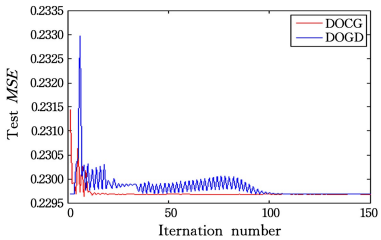


图 2 两种算法测试集的均方误差比较

Fig. 2 Comparison of mean square error on testing set of DOCG and DOGD algorithms

文献[12]的数值实验将分布式在线优化算法与集总式在线优化算法进行对比，实验结果表明，分布式环境相对集总式环境不会损失太多性能，却可以有效地降低网络中节点的数据传输速率，同时确保系统的鲁棒性。因此，本节将分布式在线条件梯度算法(DOCG)与集总式在线条件梯度算法(OCG)进行数值实验比较。图 3 中的结果表明，DOCG 算法与 OCG 算法的收敛性能几乎相同，能够得到与集总式算法类似的结果；但与集总式环境相比，分布式系统的经济成本更低且运行更高效，可以用来处理集总式难以解决的海量数据。这说明了运用 DOCG 算法的必要性和有用性。

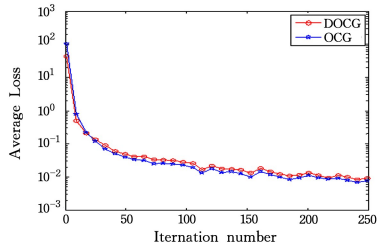


图 3　DOCG 与 OCG 种算法平均损失的比较

Fig. 3　Comparison of average loss of DOCG and OCG algorithms

图 4 表明了网络拓扑对算法性能的影响。数值实验的网络拓扑选取了循环图及小世界网络。循环图网络中每个节点只有两个邻居，连通度较低；小世界网络设置的两个参数分别为：网络图的平均度 $k=4$，每条边的连接概率 $p=0.3$，连通度较好。由图 4 可知，DOCG 算法在连通度较好的网络拓扑下，收敛速度稍快。仿真结果与第 3 节中的相关结论吻合。

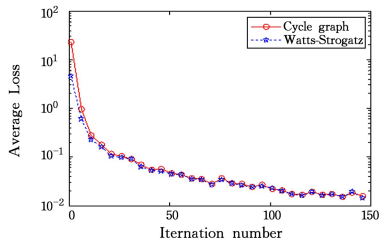


图 4　两种网络拓扑下平均损失的比较

Fig. 4　Comparison of average loss of cycle graph and watts-strogatz network topologies

**结束语**　针对分布式网络中流式数据实时分析这一重要实际问题，本文提出了分布式在线条件梯度算法(DOCG)。理论分析证明 DOCG 算法具有 $O(T^{\frac{3}{4}})$ 的 Regret 界，同时相关数值仿真实验进一步验证了 DOCG 算法具有更好的收敛性能。而实际分布式网络中数据的异步处理更具普遍性，因此研究相关的在线异步算法将是我们下一步的工作。

## 参 考 文 献

[1] AKBARIM, GHARESIFARD B, LINDER T. Distributed Online Convex Optimization on Time-Varying Directed Graphs[J]. IEEE Transactions on Control of Network Systems, 2017, 4(3): 417-428.

[2] LEES, NEDICH A, RAGINSKY M. Stochastic Dual Averaging for Decentralized Online Optimization on Time-Varying Communication Graphs[J]. IEEE Transactions on Automatic Control, 2017, PP(99): 1-1.

[3] RAMS S, NEDIC, VEERAVALLI V V. Incremental Stochastic Subgradient Algorithms for Convex Optimization[J]. Siam Journal on Optimization, 2009, 20(2): 691-717.

[4] FRANKM, WOLFE P. An algorithm for quadratic programming [J]. Naval Research Logistics, 1956, 3(1/2): 95-110.

[5] BECKA, TEBOULLE M. A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems[J]. Siam Journal on Imaging Sciences, 2009, 2(1): 183-202.

[6] RAM S S, NEDIC A, VEERAVALLI V V. Distributed Stochastic Subgradient Projection Algorithms for Convex Optimization [J]. Journal of Optimization Theory & Applications, 2010, 147(3): 516-545.

[7] YANF, SUNDARAM S, VISHWANATHAN S V N, et al. Distributed Autonomous Online Learning: Regrets and Intrinsic Privacy-Preserving Properties[J]. IEEE Transactions on Knowledge & Data Engineering, 2013, 25(11): 2483-2493.

[8] XU H F, LING Q. Decentralized online alternating direction method of multipliers[J]. Journal of Computer Applications, 2015, 35(6): 1595-1599. (in Chinese)
许浩锋，凌青. 分布式在线交替方向乘子法[J]. 计算机应用，2015, 35(6): 1595-1599.

[9] HOSSEINIS, CHAPMAN A, MESBAHI M. Online Distributed optimization via dual averaging[C] // Decision and Control. IEEE, 2014: 1484-1489.

[10] HAZANE. Introduction to online convex optimization[J]. Foundations and Trends © in Optimization, 2016, 2(3-4): 157-325.

[11] WAIH T, LAFOND J, SCAGLIONE A, et al. Decentralized Frank-Wolfe Algorithm for Convex and Non-convex Problems [J]. IEEE Transactions on Automatic Control, 2017, PP(99): 1-1.

[12] ZHANG W P, ZHAO P L, ZHU W W, et al. Projection-Free Distributed Online Learning in Networks[C] // International Conference on Machine Learning. 2017: 4054-4062.

[13] GODSILC, ROYLE G. Algebraic graph theory[M]. New York: Springer, 2001.

[14] HOSSEINIS, CHAPMAN A, MESBAHI M, et al. Online Distributed ADMM on Networks[J/OL]. https://www.researchgate.net/publication/269933107_Online_Distributed_ADMM_on_Networks.

[15] DUCHIJ C, AGARWAL A, WAINWRIGHT M J. Dual Averaging for Distributed Optimization: Convergence Analysis and Network Scaling[J]. IEEE Transactions on Automatic Control, 2011, 57(3): 592-606.

[16] NEDICA, OLSHEVSKY A, OZDAGLAR A, et al. On Distributed Averaging Algorithms and Quantization Effects[J]. IEEE Transactions on Automatic Control, 2012, 54(11): 2506-2517.

[17] LING Q, XU Y, YIN W, et al. Decentralized low-rank matrix completion[C] // IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2012: 2925-2928.