总结报告22

论文Communication-Censored Linearized ADMM for Decentralized Consensus Optimization及其他论文

（2020.3.14）

一、contents

Li, W., Liu, Y., Tian, Z., & Ling, Q. (2020). **Communication-Censored Linearized ADMM for Decentralized Consensus Optimization**. IEEE Transactions on Signal and Information Processing over Networks, 6(1), 18–34.

Richtárik, P., & Takáč, M. (2016). **Distributed coordinate descent method for learning with big data**. Journal of Machine Learning Research, 17.

Duchi, J. C., Agarwal, A., & Wainwright, M. J. (2012). **Dual Averaging for Distributed Optimization: Convergence Analysis and Network Scaling**. IEEE TRANSACTIONS ON AUTOMATIC CONTROL, 57(3), 1564–1565.

二、论文1

Li, W., Liu, Y., Tian, Z., & Ling, Q. (2020). **Communication-Censored Linearized ADMM for Decentralized Consensus Optimization**. IEEE Transactions on Signal and Information Processing over Networks, 6(1), 18–34.

1. 内容

这篇论文提出的模型叫COLA（communication-censored linearized ADMM）：

论文有两个点：一个是linearize，即线性近似来减少computation cost；另一个是communication-censoring strategy来降低communication cost，即有一个阈值，只有变量的变化大于阈值，才开始进行交流，弥补了线性化步骤造成的通信效率低下的问题

ADMM好处：fast & exact convergence

1. 符号标记

1）双向网络：

n个nodes

m条有向边

2）

3）extended block arc **source** matrix： ，包括m× n方块 

如果边e=（i，j）属于，则为单位阵，否则，等于0.

extended block arc **destination** matrix：，同样包括m× n方块 

但是，它的方块当且仅当，边e=（i，j）属于，且有向边的终点朝向j，才等于单位阵，否则都是0.

extended **oriented** incidence matrix： ；**oriented** Laplacian：

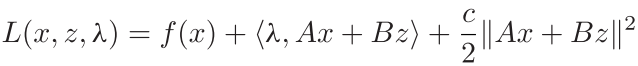
the **unoriented**：；**unoriented** Laplacian：

degree matrix：，其中

1. ADMM



增广拉格朗日量：

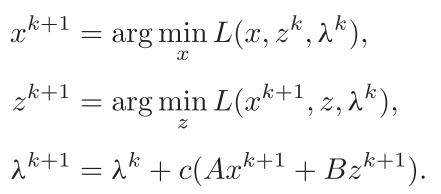


Note:

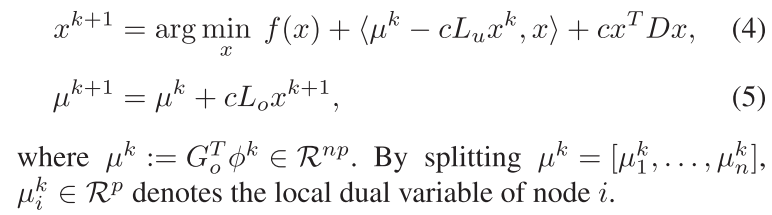
，z为辅助变量

是Lagrangian multipliers，满足限制

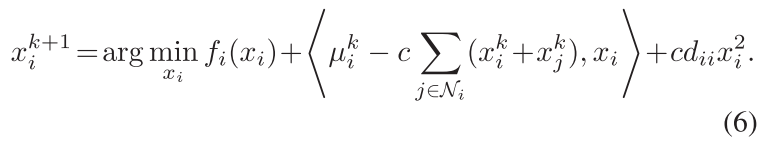
更新步骤：

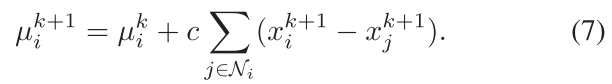


通过方法删除z，将λ用低维变量代替，可以得到简单更新形式：



利用上面的符号定义，这篇论文中更新步骤变成：





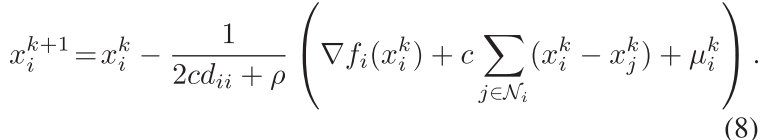
1. COLA

模型特别之处有两点：①线性化；②选择性的交流

1. Linearization：

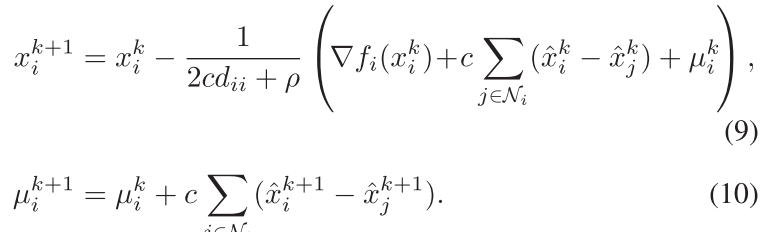
二次逼近： 

于是，（6）变成：

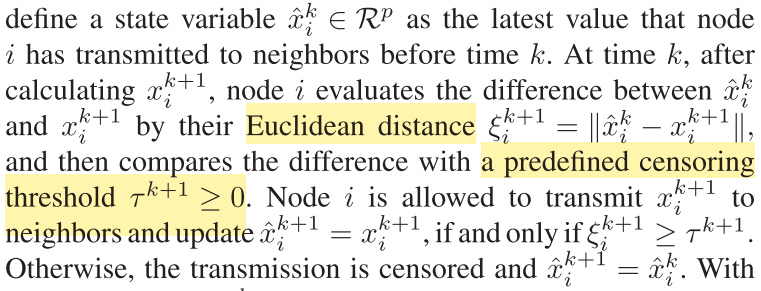


1. Communication censoring：

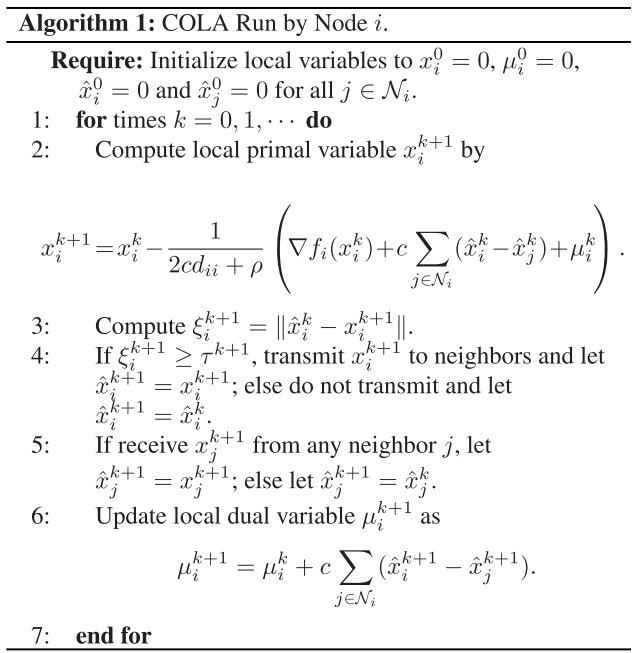
更新步骤变为：（注意下面式子中变量头顶带的）



Note：对于communication-censoring strategy及头顶的解释：



总的算法为：



三、论文2

Richtárik, P., & Takáč, M. (2016). **Distributed coordinate descent method for learning with big data**. Journal of Machine Learning Research, 17.

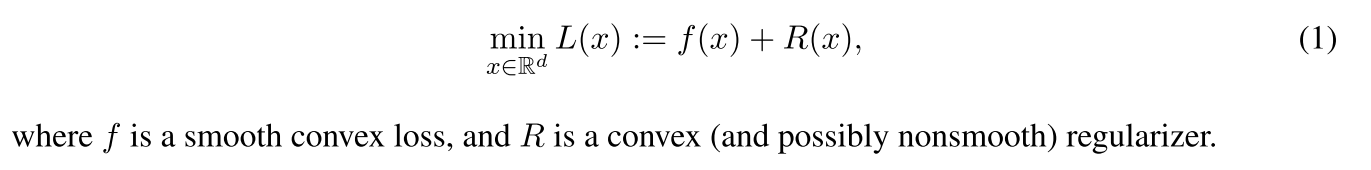
1. 内容

坐标下降**coordinate descent**，实际上就是每一次优化变量x（p维）其中的几个维度（或者叫feature），也就是coordinate之意：坐标轴。而Randomized coordinate descent methods就表示每次选择的那几个维度是随机的。

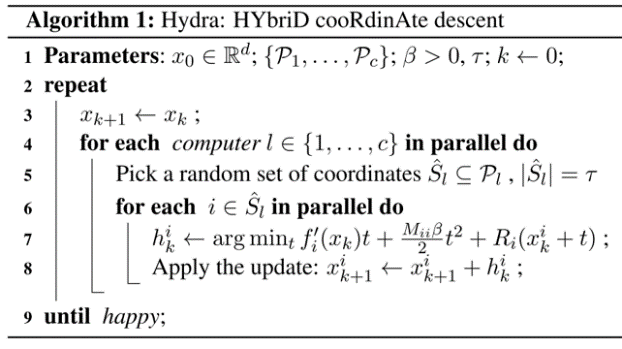
一共有c的nodes和d维的变量，每一次迭代每个node都从分配给它们的维度中选择个coordinate进行优化，所有的nodes都完成了，就得到了cτ个coordinates.（每次并不是所有的都更新了≠d）

The resulting vector, stored as  vectors of size s = d/c each, in a distributed way, is the new iterate.

1. 目标函数



1. 算法



Step 3：modify cτ entries of xk in total the result is called xk+1

step 4-5：注意每一轮迭代，x只有cτ个coordinates更新了，不是全部d个维度都更新了。

Step 7：需要计算f关于x第i维的偏导

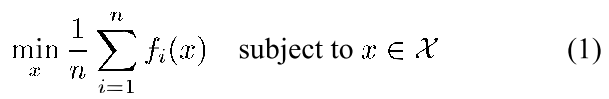
四、论文3

Duchi, J. C., Agarwal, A., & Wainwright, M. J. (2012). **Dual Averaging for Distributed Optimization: Convergence Analysis and Network Scaling**. IEEE TRANSACTIONS ON AUTOMATIC CONTROL, 57(3), 1564–1565.

1. Idea

dual subgradient averaging

1. 目标函数



Note：eachis convex and hence sub-differentiable, but need not be smooth.

1. Standard dual averaging

首先，论文中假设每一个都是L-Lipschitz，也就是说，



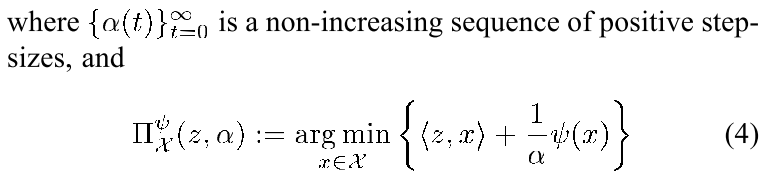
也就意味着，对于任意的次微分，我们都有其中， notes the 

那么，dual averaging algorithm就是产生序列

对于任意的次梯度



Note：



is a type of projection.

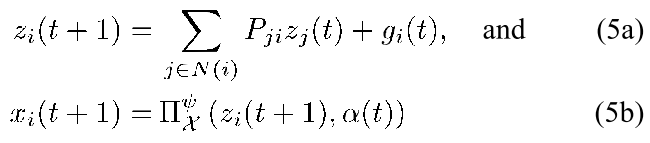
dual averaging algorithm也就是说，每次迭代产生，要使得minimize an averaged first-order approximation to the function，并且proximal function 和stepsize α(t)＞0是使得不至于振荡的太剧烈。

1. Distributed Dual Averaging

每一个node都计算它的次梯度，从它的邻居节点得到z（t）后加权平均，之后再计算它的下一次更新x（t+1）利用一个（由和α(t)定义的）投影。

也就是下面的更新步骤：

Update：



其中，P是一个doubly stochastic matrix：non-negative weights that respects the structure of the graph

