总结报告17

Communication-efficient distributed multi-task learning with matrix sparsity regularization

（2020.2.26）

1. 题目

Zhou, Q., Chen, Y., & Pan, S. J. (2019). **Communication-efficient distributed multi-task learning with matrix sparsity regularization**. Machine Learning.

2. 论文内容

2.1 亮点

1）利用dual problem

2）为了reduce communication cost，利用了一个data screening approach（数据筛选方法），在nodes之间只传active features，inactive features为0，不起作用。

3）matrix sparsity regularization

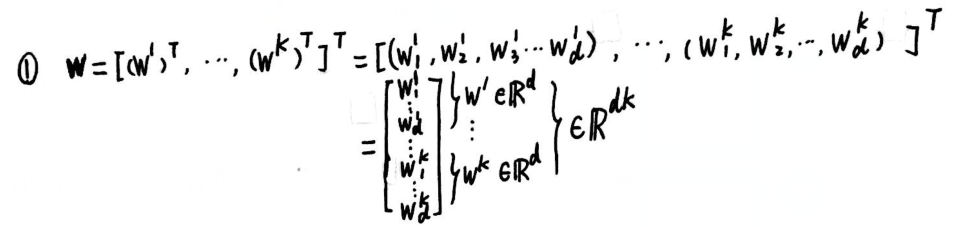
4）最大的优点是：附录中给了非常详尽的primal-dual推导

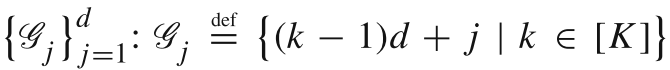
2.2 论文重点提取

1）首先交代一下符号中重要的：

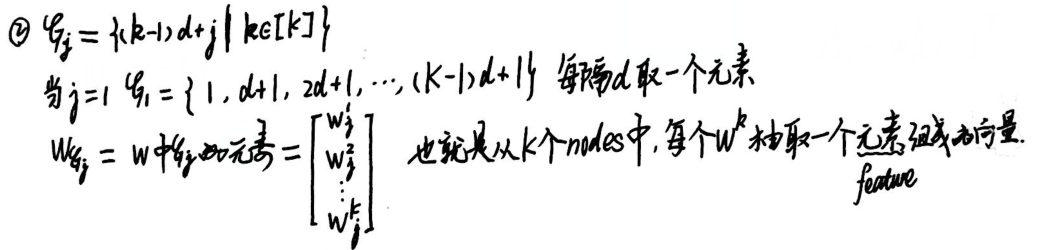
有K个nodes，每一个node有个examples，则所有nodes一共有n（）个examples.



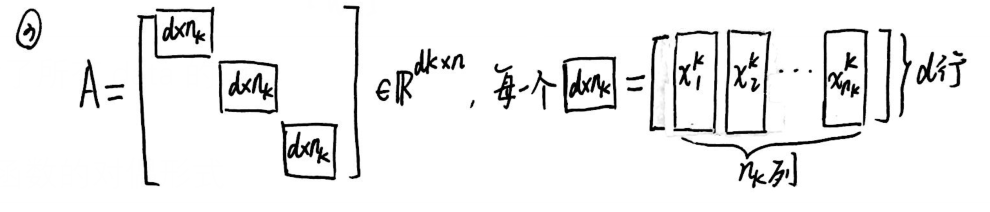


，其中

所以，的解释

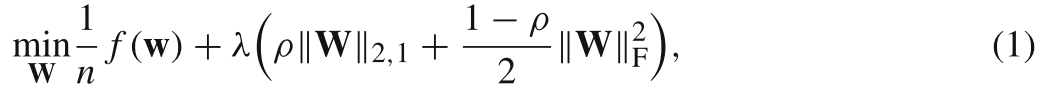


A为包括了所有data的矩阵



2）目标函数的对偶形式

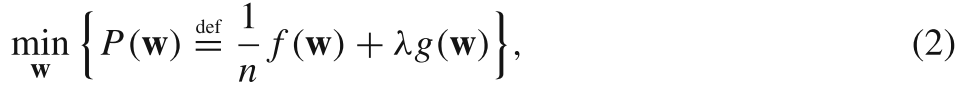
Primal-global loss：



其中，



Primal-distributed loss：



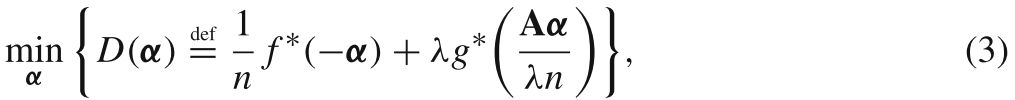
其中，



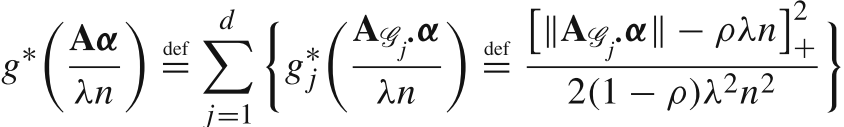
注意：这里分布式是依据feature而不是依据instance的，因为根据上面的定义：是在K个nodes上每个node抽取了wk的一个特征，组成的，然后g（w）是d个特征Σ，而不是把g（w）分布到K个nodes上（K个nodes的Σ）。作者说这是和现实有点不太相符，现实多是依据instance的分布式。

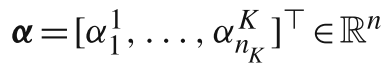
却是根据K个node上的examples分布的，所以作者下面要对loss修改，以便能够在K个nodes上分布

Dual-distributed loss：



其中，



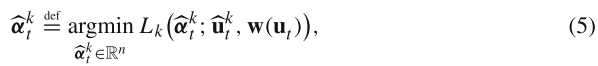


最后附上了推导过程，在推导过程中出现了疑问。

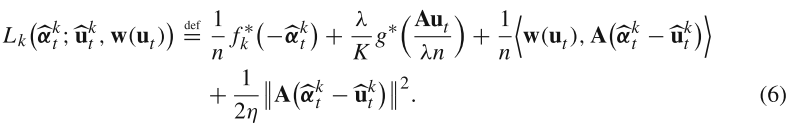
w和α的关系转换式如下：



3）分布到每一个node上，node要处理的子问题为(实际每个node处理的subproblem)：



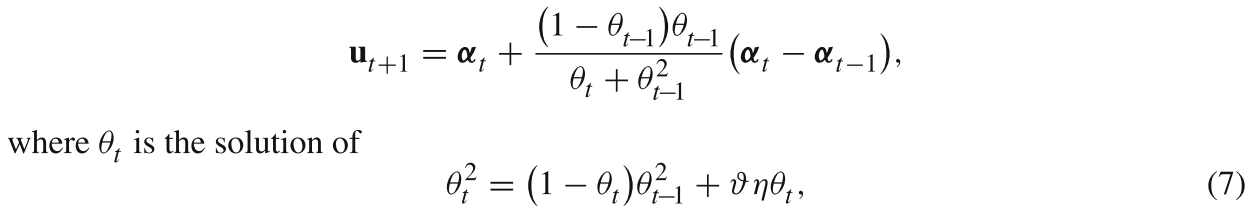
其中，（）



Note：

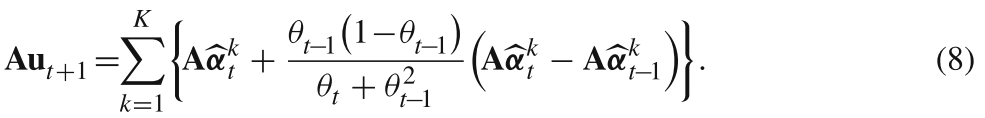
a）这里的改变是因为是根据feature求和的形式，所以没办法分布到K个nodes上，而f()不是这样。故由于这个原因，干脆把整个loss改写成（5），这里的理论证明在于。

b）u是一个参考点（reference point），实际就是α。比如，在COCOA+中，为上一次迭代的解，但是作者在这里没有按照COCOA+中的定义，而是采用了下面的定义，下面这样定义好处在于利用了 和，即两次迭代的结果，这样利用的信息多一些。



其中，

那么有，

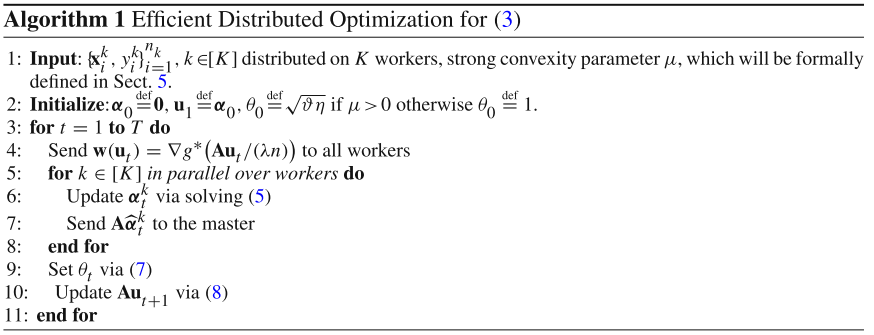


Note：

对于这样一个master-worker框架下，worker把传给master，master把传给worker。对于每个worker实际上就成了计算和传递，所以A要乘u后再传给worker。

4）算法：

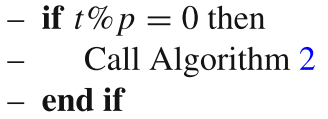
有了上面的说明铺垫，得到下面的算法：

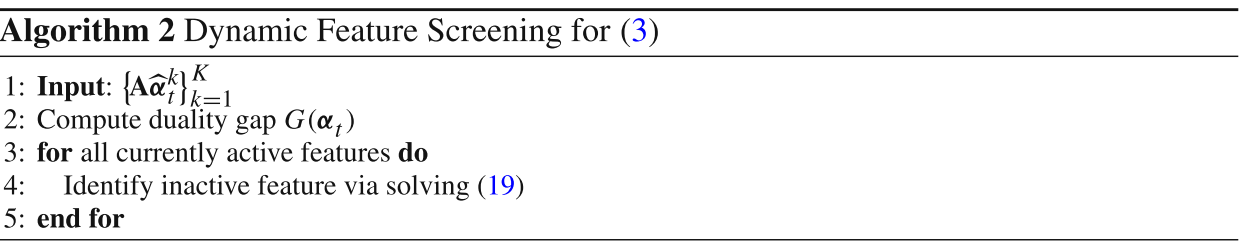


5）最后一点

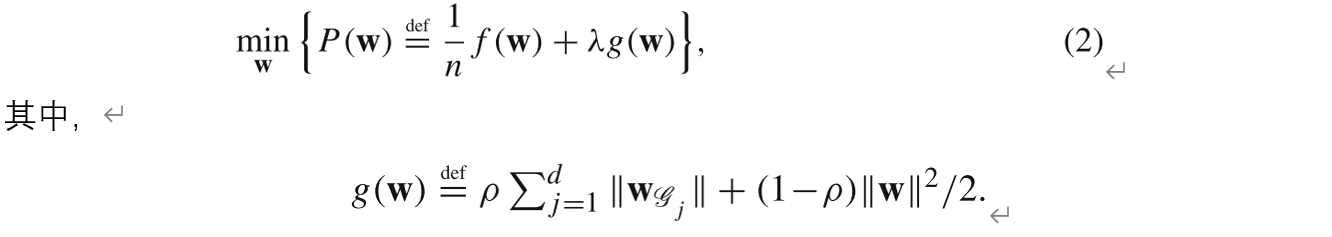
作者为了减少communication cost，提出了dynamic feature screening，意义在于每次识别一些inactive feature，他们对应的w的分量的最优值为0，而active feature才是真正有意义的。每p次迭代，进行一次feature screening，识别筛选一次。

这个idea反应到伪代码上，就是把下面三条语句加到before line 4 in Algorithm 1

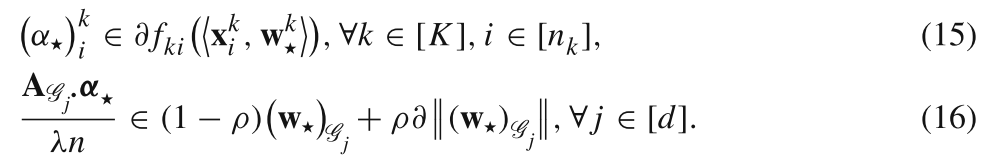




计算理论略，但是推导过程中，利用了（2）式的KKT条件得到式（15）（16）

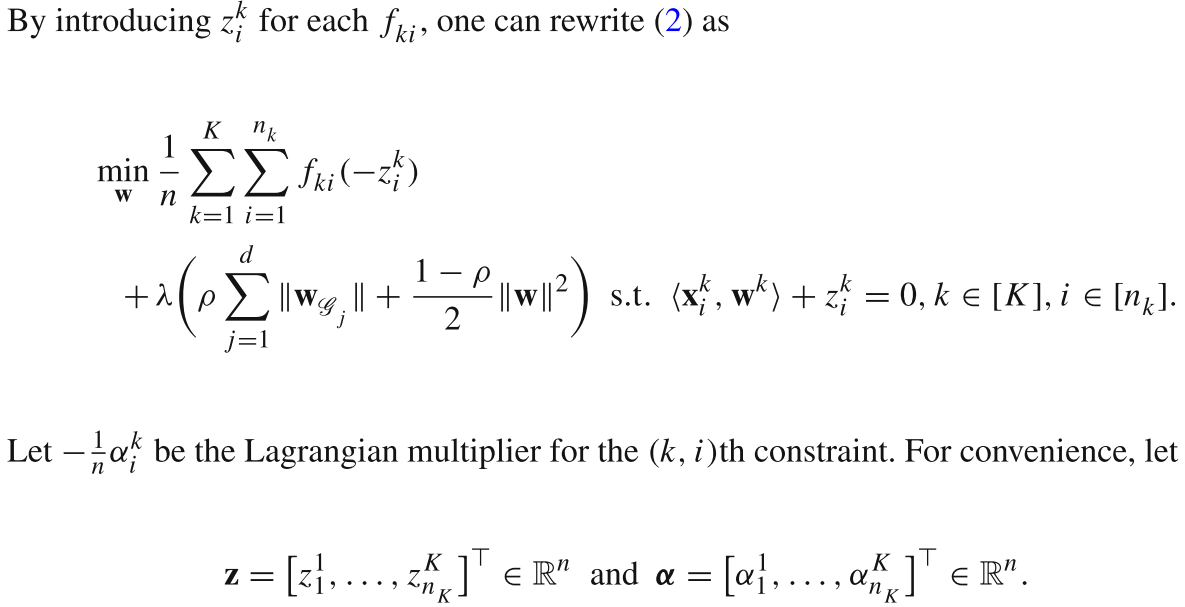


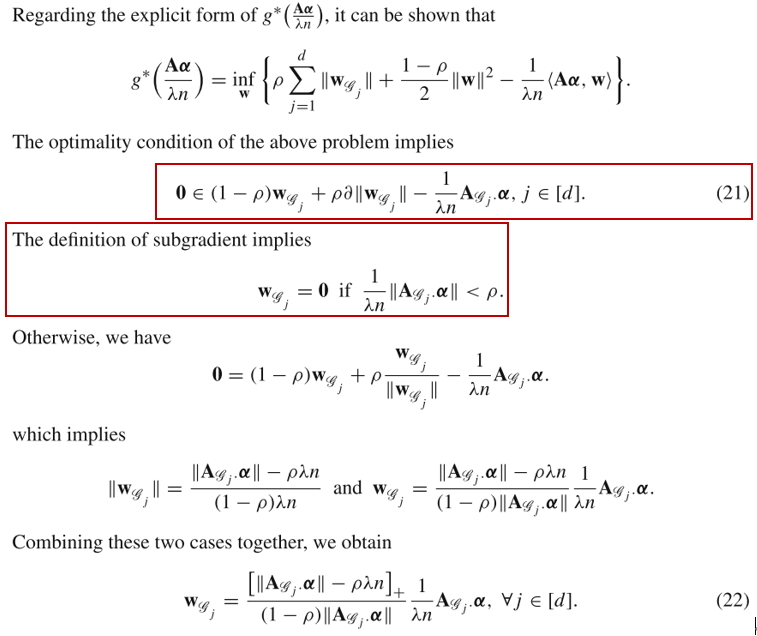
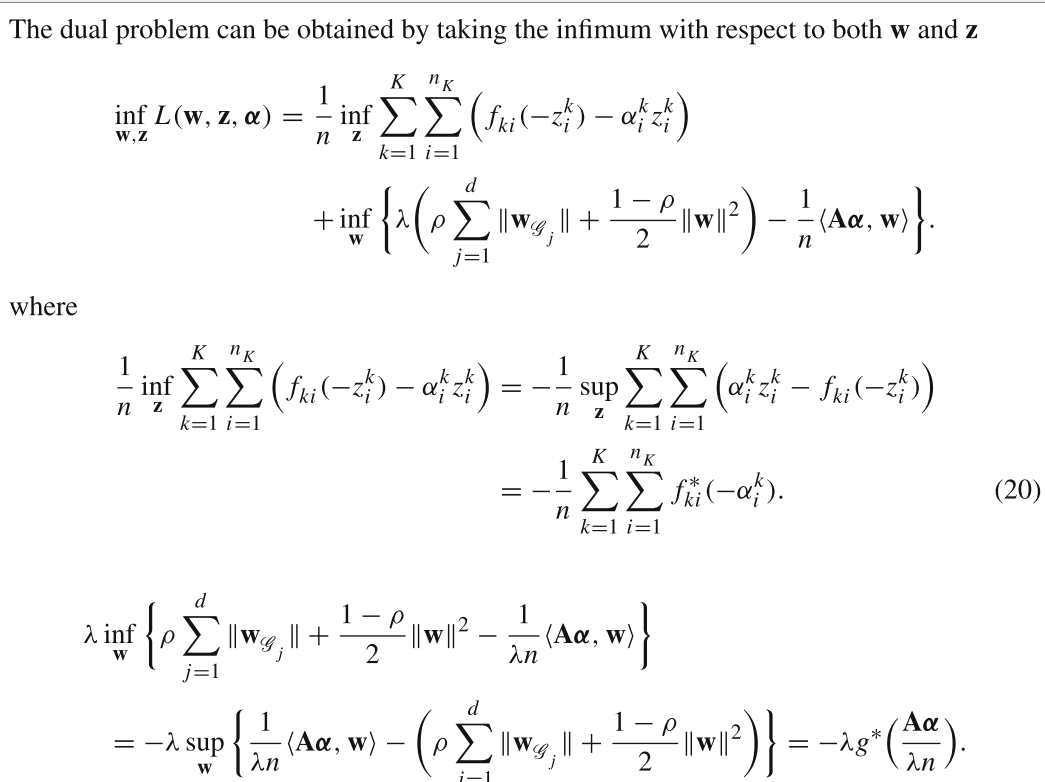
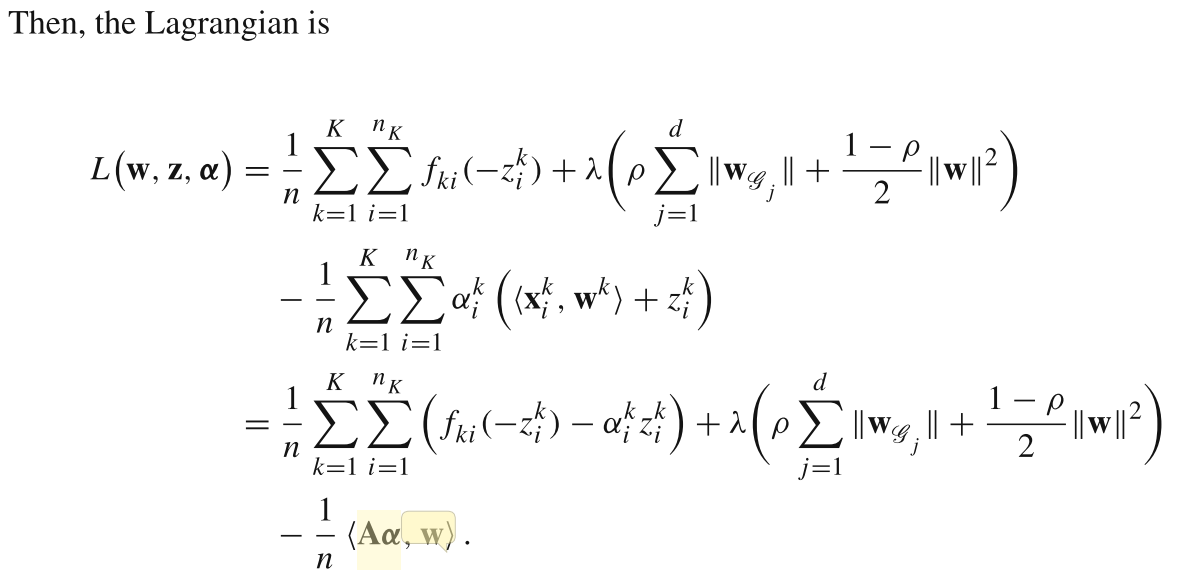
是怎样利用KKT条件得到下面的式子（15）（16）的？

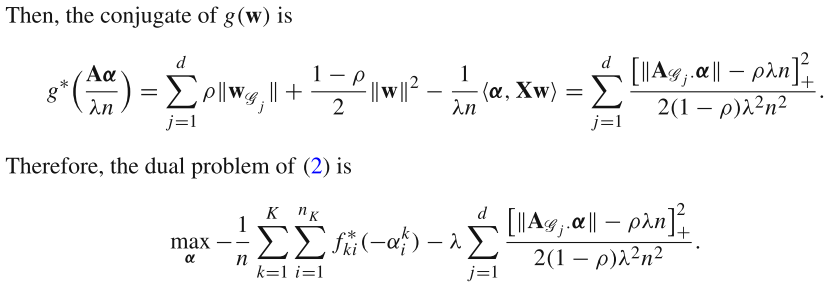


Note:

Dual problem推导过程：







我的问题在推导的红框处

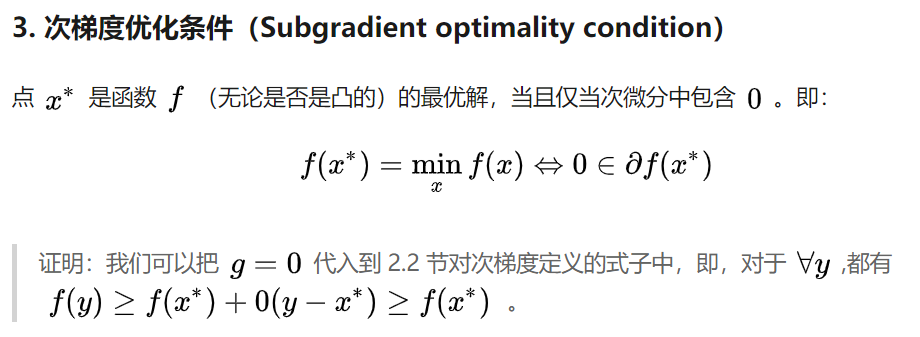
问题1：第一个红框处

次梯度的理论基础网站：

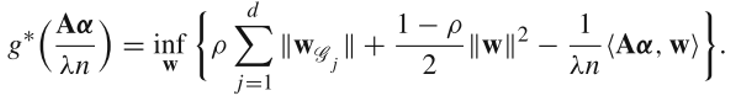
<https://www.zhihu.com/search?type=content&q=%E6%AC%A1%E6%A2%AF%E5%BA%A6>

note：次梯度与次微分存在区别

次梯度的理论基础：

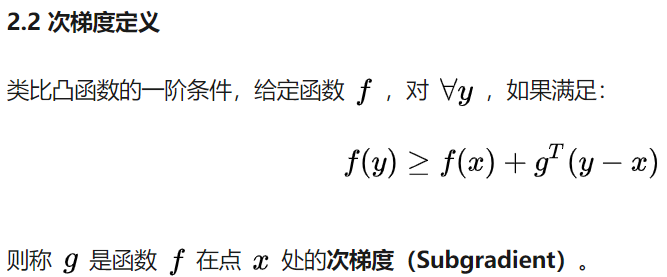


根据



应该是对w求解次微分（Subdifferential），但为什么（21）式对求解了次梯度（Subdifferential）后，然后利用了最优条件？

问题2：第二个红框



为什么根据次梯度（subgradient）的定义可以得到第二个红框？