总结报告16

论文4篇（修改版）

（2020.2.18—2020.2.19）

红字标识为修改部分

一、contents

Wang, W., Wang, J., Kolar, M., & Srebro, N. (2018). **Distributed Stochastic Multi-Task Learning with Graph Regularization**.

Xu, J., Tan, P. N., Luo, L., & Zhou, J. (2016). **GSpartan: A geospatio-temporal multi-task learning framework for multi-location prediction**. 16th SIAM International Conference on Data Mining 2016, SDM 2016, 657–665.

Li, C., Huang, S., Liu, Y., & Zhang, Z. (2018). **Distributed jointly sparse multitask learning over networks**. IEEE Transactions on Cybernetics, 48(1), 151–164.

Verma, V. K., & B, P. R. (2017). **Distributed Multi-task Learning for Sensor Network**. 1, 792–808.

二、总结

Wang, W., Wang, J., Kolar, M., & Srebro, N. (2018). **Distributed Stochastic Multi-Task Learning with Graph Regularization**.

主旨说明：

1. 对W的限制条件之一：两个近的点间w相差不大，利用图矩阵L，变成正则项进入目标函数中
2. 使用随机梯度SGD，减少计算次数
3. 我认为这篇论文实际上创新点一是加入了Laplacian 图矩阵这样一个先验知识，而且作者在后面的mini-batch SGD中sampling也强调要尊重graph，也即利用graph所包含的先验知识。二是对于算法的讨论，作者认为把F（W）和R（W）其中一个线性化可能就足够了，也能得到收敛后的W.对于这两种方案，作者认为他们都是在不同的通信体制下使用，作者的目的在于减少交流和计算成本。利用stochastic、prox加快计算。

论文脉络：

1. Introduction
2. Graph-based multi-task learning
3. Distributed algorithm for ERM

3.1 Directly solving the regularizer

3.2 Directly optimizing the loss

4. Stochastic algorithms

4.1 Directly solving the regularizer

4.2 Directly optimizing the loss

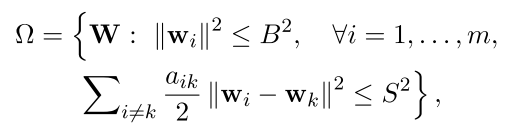
5. Connection to consensus learning

6. Experiments

Details:

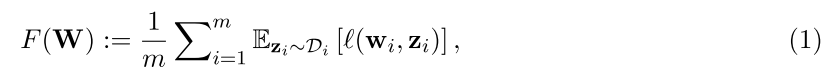
1. Introduction
2. Graph-based multi-task learning：提出了基于图矩阵的目标函数

权重矩阵要满足下面的要求：



这个Ω的意义是：范数有界，且related的predictors不相似度小

全局的优化目标为：



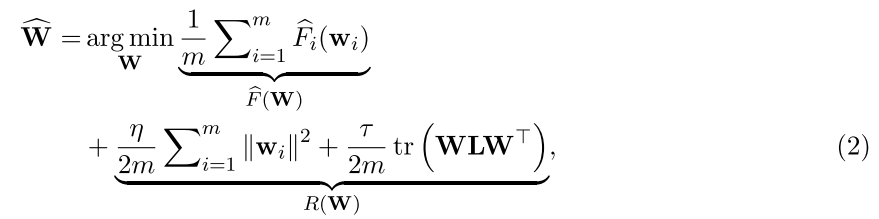
其中，为每个node的优化目标。

注意，对集Ω的应用：

对集Ω其中一个条件利用Laplacian 图矩阵，即L，则该条件可以转化为



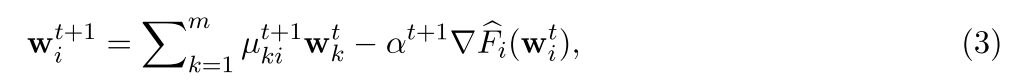
Regularized ERM (Empirical Risk Minimization) Objective：



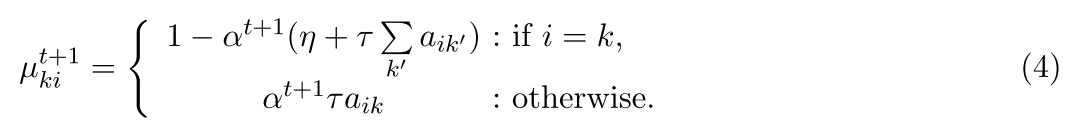
1. Distributed algorithm for ERM

解决目标函数（2）：

首先对使用gradient descent



其中，为weights for combining neighboring predictors



问题：作者想要解决问题（2），原文说

The simplest approach is perhaps to perform gradient descent on .

With an appropriate step-size schedule (or even a fixed stepsize if the loss is smooth), this method converges to? 这里为什么只对使用了梯度下降，而没有管R(W)?

原文：Taking steps based on the gradients amounts to considering, in each iteration, a linearization of the objective, that is of both the empirical lossand the regularizer R(W).

根据梯度采取的步骤就相当于在每次迭代中，对目标函数同时线性化。

即正常要对和都线性化，但是作者接下来仅对其中一个线性化，并且对另一项给出明确的处理。作者认为下面两种方法是强大的alternatives。

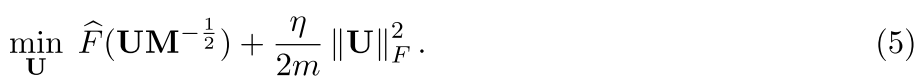
3.1 Directly solving the regularizer

首先，利用变换



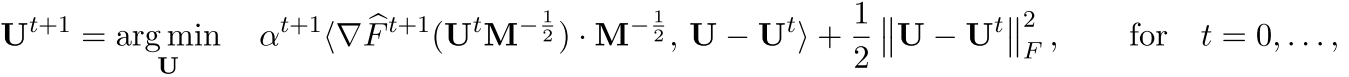
我认为在这里使用变化U并没有什么特别的地方，只是为了将R(W)中的两项结合成一项||U||2F, 3.1节的重点是将F（W）线性化，最后得到一个计算w的算法

将目标函数改为，



通过Appendix D的叙述，可见这里对进行了线性化，二阶近似形式

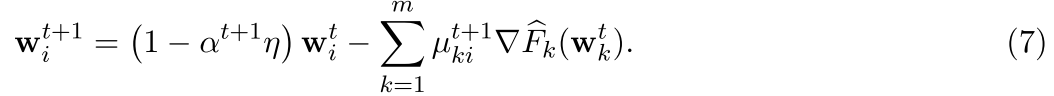
推导的理论基础<https://blog.csdn.net/qq_38290475/article/details/81052206>

（10）

关于U使用GD，在W-space，上述公式简化为:



分布到每一个node：

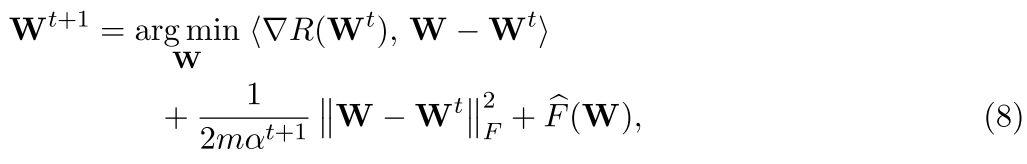


Note：通过通信一圈得到每个node的梯度，M-1提前计算

3.2 Directly optimizing the loss

由于上面的算法需要通信一圈，不太符合分布式的设置，于是得到下面的算法：

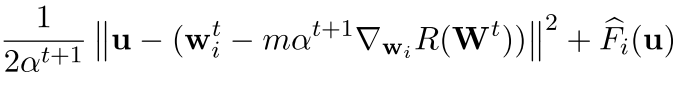
linearize the graph regularizer but fully optimize over the loss:



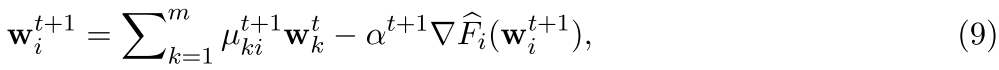
公式（8）也使用了近端梯度下降方法，||W-Wt||2F是在二次近似后自然而然出来的，线性化后可能有不好的性质，强凸保证

分配到node：

下面的公式利用了近端梯度算法，



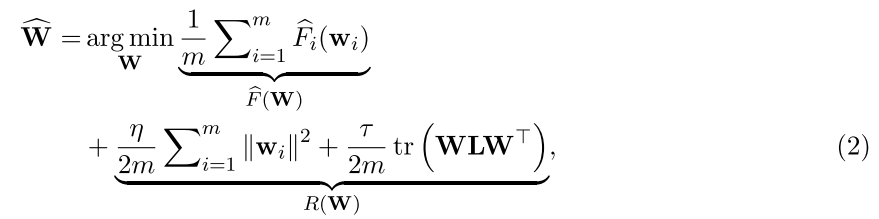
根据优化的最优条件，



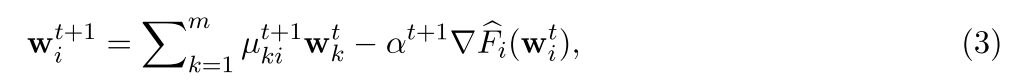
Summary：

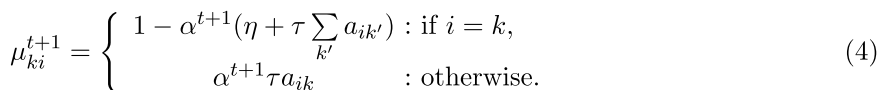
上述三种方法：

问题：

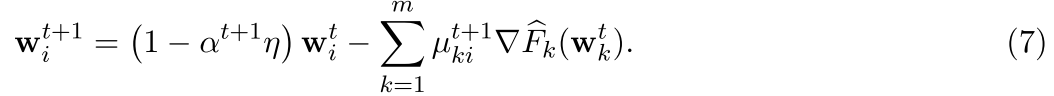


1. F和R全部线性化：得到的是公式（3）



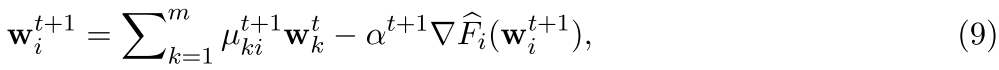
Where，

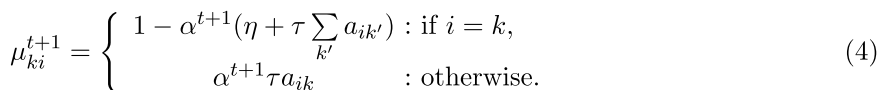
1. 线性化处理F：得到公式（7）



Where，

1. 线性化处理R：得到公式（9）



Where, 

作者认为2）3）两种方案，每次只处理F或R其中一项，到底选择哪种要看图关系、结构和交流的形式。

作者认为公式（7）需要using one round of global, all-to-all communication

认为公式（9）需要using a local，peer-to-peer communication

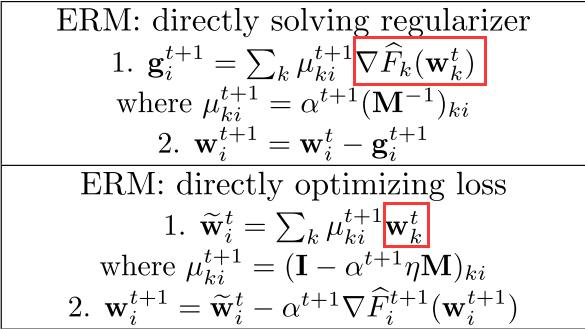
这个地方的global应该指全局通信，通信一圈，而local应该指只和邻居通信。但是从公式没看出来？

作者说，Comparing (9) with the similar update (3) where we linearized both the regularizer and the loss, we observe that (9) is also a form of gradient method, with the gradient of loss evaluated at the “future” point.

也即（3）和（9）公式前面都是相同的，唯一的区别在于（9）的最后一项是对使用梯度，而（3）是对使用梯度。

作者说，The advantage of (9) is that the gradient ∇R(W) is data-independent and is obtained using only one round of local communication from each machine to its neighbors.

确实，R(W)是跟data或example无关，而是data-dependent。但是后面的只需要和邻居一圈的当地通信即可得到，没太懂?

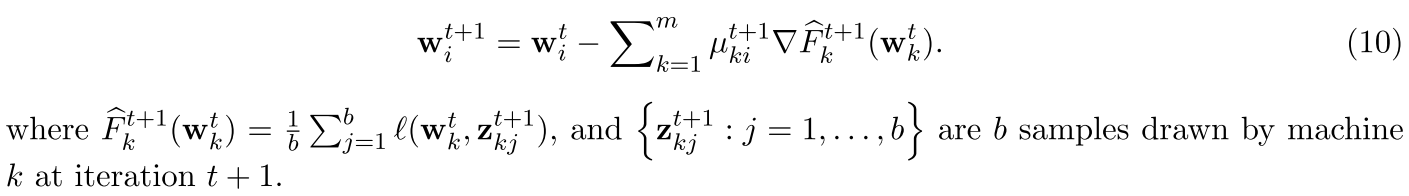


作者强调交流的信息message有两种，一种是gradient表格第一行，一种是iterates，即表格下一行。但是都要通信，通信gradient和parameter有什么区别呢？

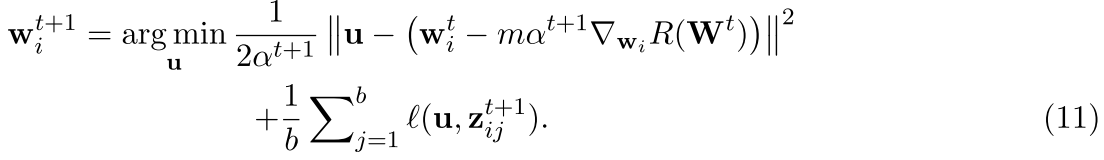
4. Stochastic algorithms

4.1 Directly solving the regularizer

minibatch SGD with **b samples** per machine



4.2 Directly optimizing the loss



5. Connection to consensus learning

在这一小节作者强调的是，当参数怎样变化时，可以把multi-solution变成共识consensus问题，共识问题指的就是所有的nodes最后都得到了一个模型，一个参数W。

作者想说他们的模型对于这种consensus问题是兼容的。

Xu, J., Tan, P. N., Luo, L., & Zhou, J. (2016). **GSpartan: A geospatio-temporal multi-task learning framework for multi-location prediction**. 16th SIAM International Conference on Data Mining 2016, SDM 2016, 657–665.

主旨说明：提出GSpartan模型

1）Local models共享一些公共的、低秩representations

2）每个局部模型（local models）作为基模型（base models）的线性组合

3）有两种变量：response variable，即我们想要预测的变量；

predictor variables，是由全局或地区模型产生的其他输出

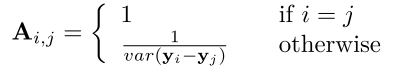
比如，预测每月降雨量是response variable ;mean temperature at 2 meters, mean sea level pressure, 500 hPa geopotential height, and near surface relative humidity是predictor variables。

也就是说response variable 是因变量y；predictor variables是自变量X

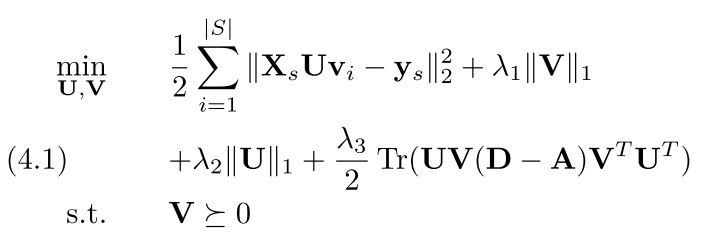
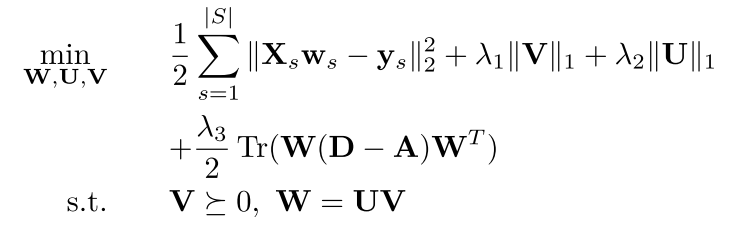
4）实验部分通过去掉拉普拉斯图矩阵的正则化进行比较，还和一些其他方法进行了比较

5）拉普拉斯图矩阵：通过修正变差函数的逆（inverse of a modified variogram）给出

Variogram：方差图是一种空间统计中发展起来的一种度量方法，用来确定一对位置之间的空间依赖性。见引用文献[10]



目标函数：

Note：W=UV

U：a feature representation of the base models

V：the weighted combination of the base models that form the local model at each location.

矩阵U是基本模型的特征表示，V表示在一点形成局部模型的基本模型的加权组合。

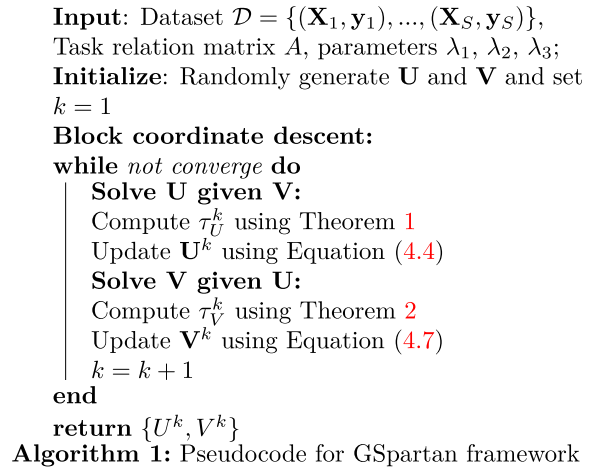
A：拉普拉斯图矩阵，Task relation matrix，表示空间自相关，引入任务关系的先验知识

为了模型的可解释性，引入稀疏和非负。作者在这里希望两个稀疏，一个是每个base model里面的变量尽量少，即如果有1000个变量，最好只需要5个用来解释。另一个稀疏是local用的base model少一些。所以，分别在U,V上加1-norm稀疏，另外保证非负，即不引入负向的解释。

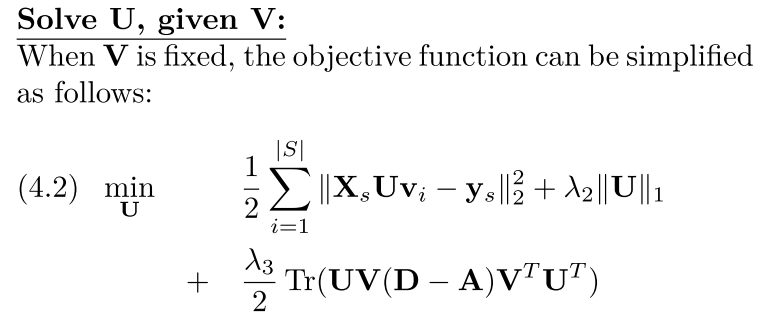
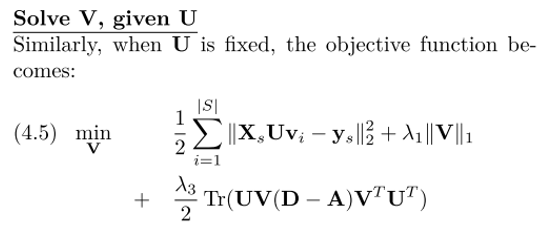
Tr（）：类似理解成F范数的形式，用来表征任务之间尽可能相似的penalty term。即A表示空间自相关，两个点地理上相近，那么他们的参数也相近。

优化算法：

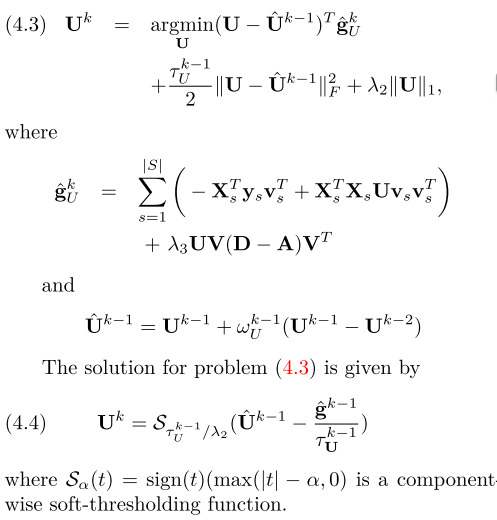
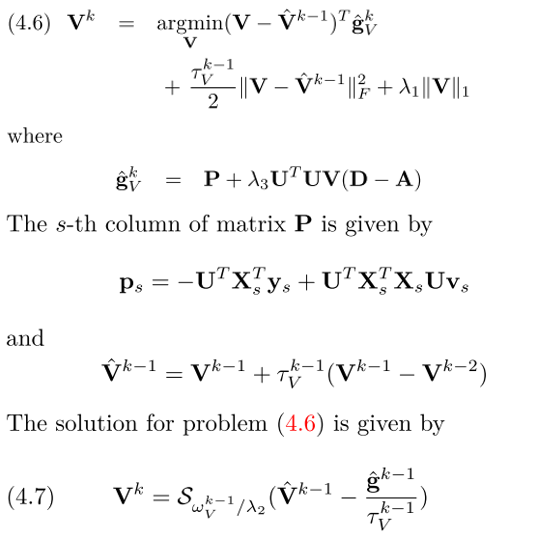
通过BCD，交替优化U和V

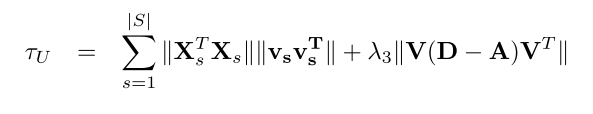


这里可以看到，虽然V≥0，是有约束的优化，但是可以看到通过分步优化，就相当于没有约束条件。



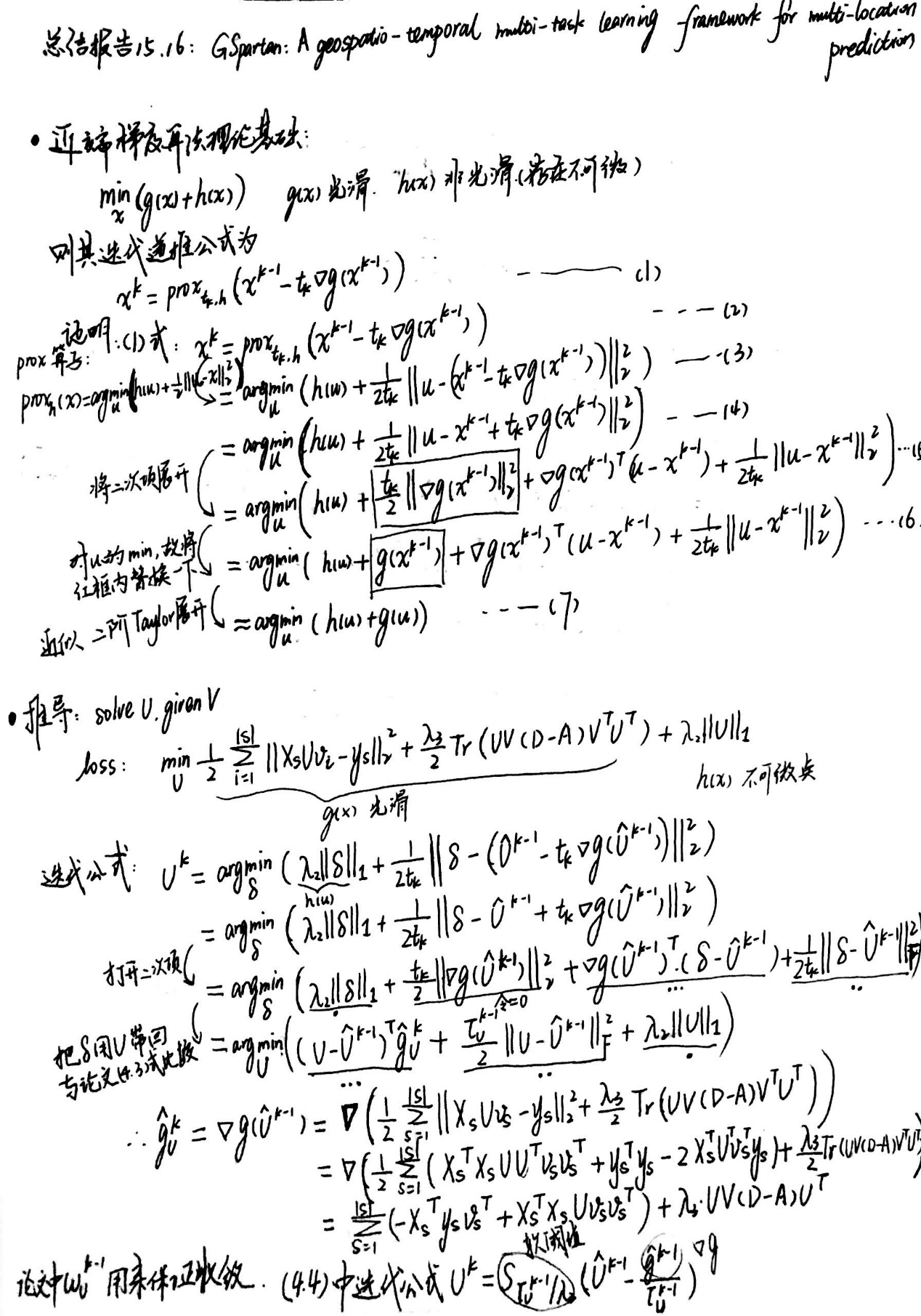
通过proximal gradient descent进一步计算， 同样通过proximal gradient descent，





推导过程：

利用了近端梯度算法。推导U，同理可得V



Li, C., Huang, S., Liu, Y., & Zhang, Z. (2018). **Distributed jointly sparse multitask learning over networks**. IEEE Transactions on Cybernetics, 48(1), 151–164.

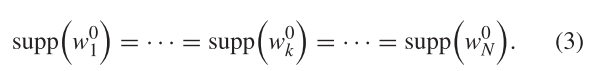
主旨说明：AC-dJSMT algorithm

1. 文章考虑了多任务的两点：

的一项为惩罚intertask similarities；的一项用来惩罚joint sparsity。

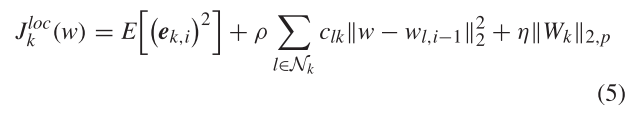
如果只考虑其中一种情况，只须让对应的项的系数为0.

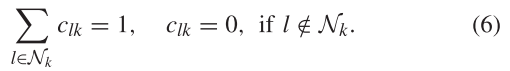
1. 为了分布式，将W分配给每个节点Wk：a local parameter matrix（consisting of the parameter vectors of all its neighbors），作者在这里说明了使用Wk最终也会收敛到min
2. 作者认为joint sparsity意味着每一个node的参数向量哪个位置是0，哪个位置不是0，对于所有的节点都是一样的，比如第一个分量all nodes全是0，第二个分量全不是0…这样的结构，即



4) intertask similarities的部分intertask combiner，见下优化算法所述。

目标函数：



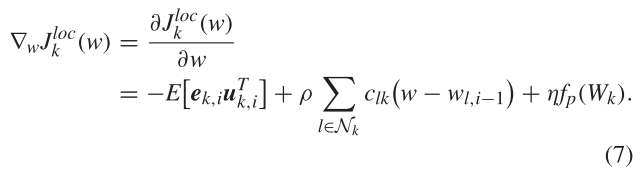
Note：.另外有：intertask combiners

：similarity-promoting term，想让intertask即related tasks邻居之间的参数相似一些

：用来惩罚稀疏，sparse-promoting term，分别讨论了L2,1-norm, reweighted L2,1-norm, L2，0-norm

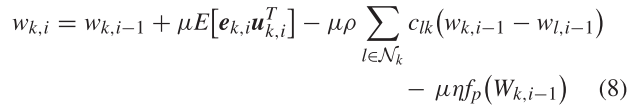
解决方法：

steepest-descent method：这里只是简单的用gradient descent 方法求解，重点在于对L2，p-norm范数求次梯度.

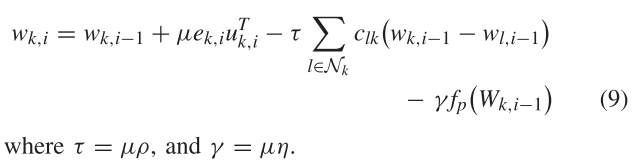


Note:是关于w的L2，p范数的次梯度

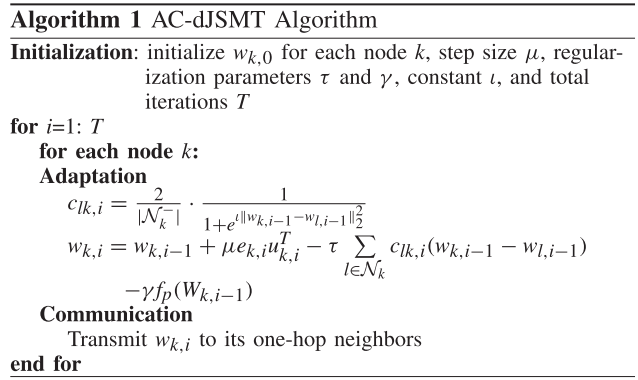
迭代更新公式：



用来代替得到，



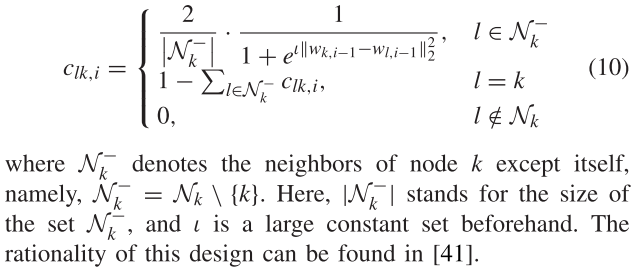
优化算法：



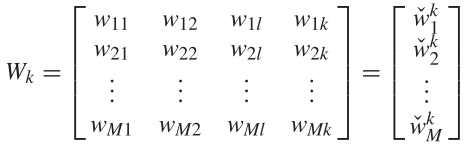
·当is fixed，得到了FC-dJSMT（fixed  distributed jointly sparse multitask）

·当is adaptive，得到了AC-dJSMT（adaptive  distributed jointly sparse multitask）

由下面的公式给出的取值，

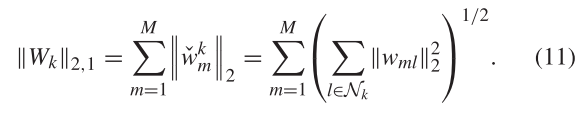


·对于稀疏惩罚项||W||2，p的范数的讨论

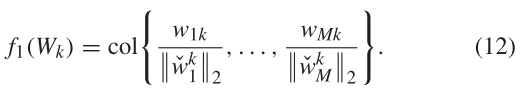


其中，

1. L2，1-norm

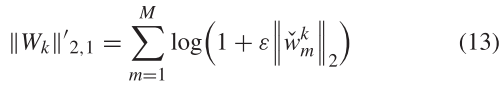


它的次梯度为：

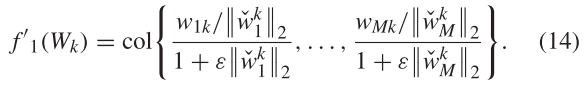


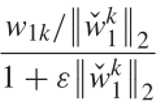
其中，就是Wk矩阵的第一行

1. Reweighted L2，1-norm ：RW范数

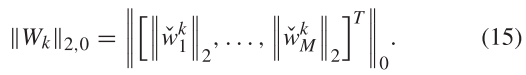


它的次梯度为：

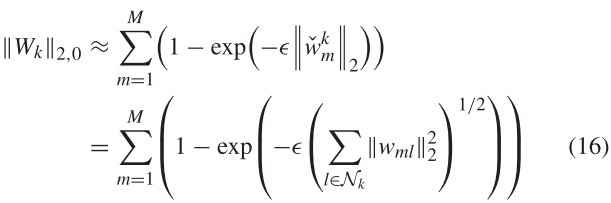


把分子的除下来，可以发现，越小，越趋近于0，促进稀疏（zero attractor）

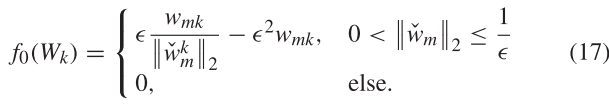
1. L2，0-norm

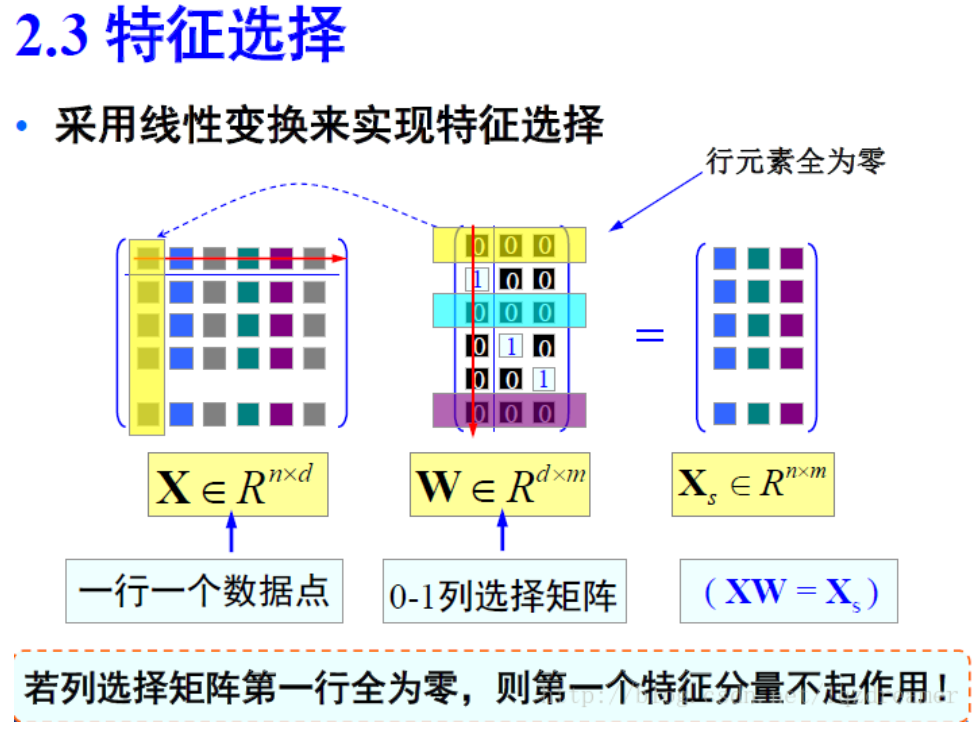
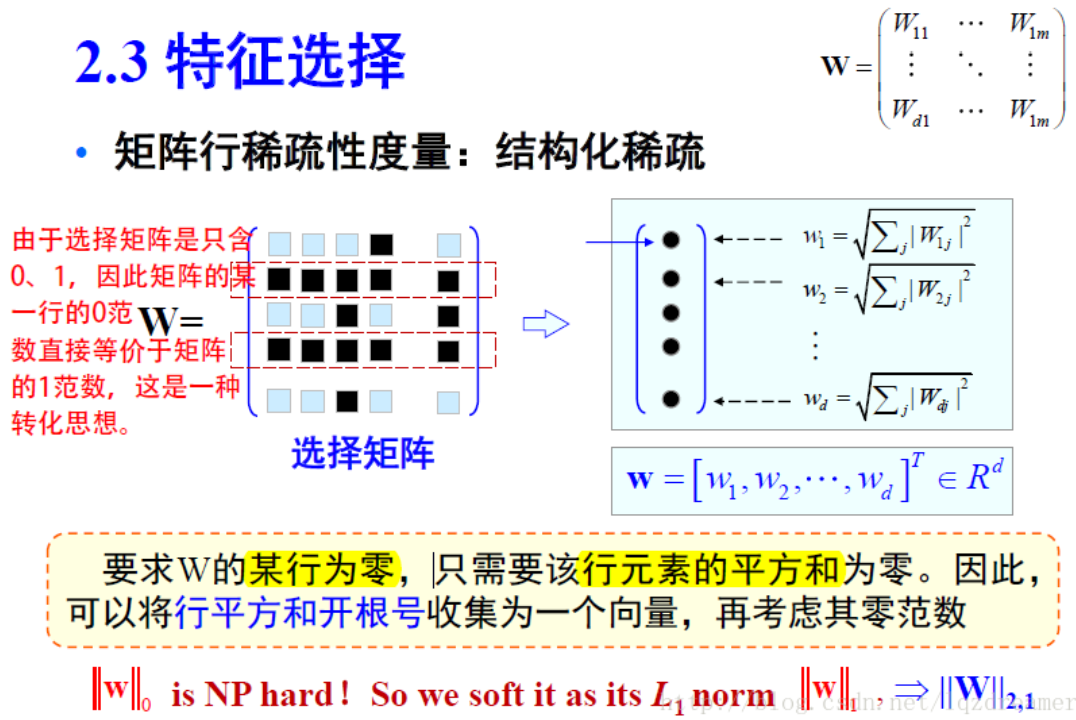


L2，0-norm非凸：



利用上式（16）L2,0-norm的近似再将指数函数用一阶泰勒展开式展开，得到下面，它的次梯度，



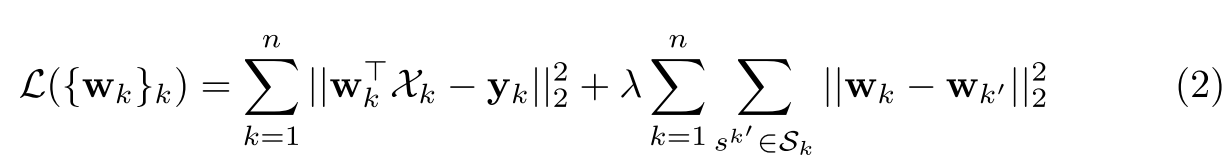
Li, Jiyi, et al. "**Distributed Multi-task Learning for Sensor Network**." Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, Cham, 2017.

主旨说明：

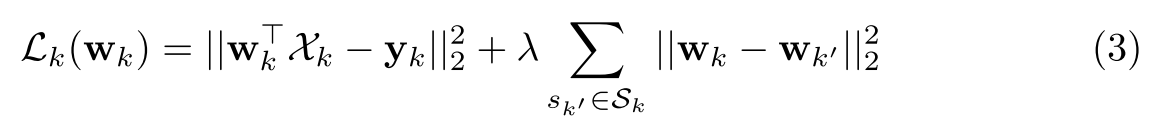
比较中规中矩的一篇分布式+多任务学习，在neighbors间传递**参数parameter**

目标函数：

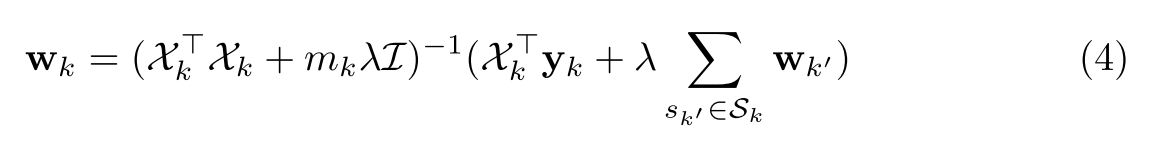
全局的：



分配到每一个节点的：

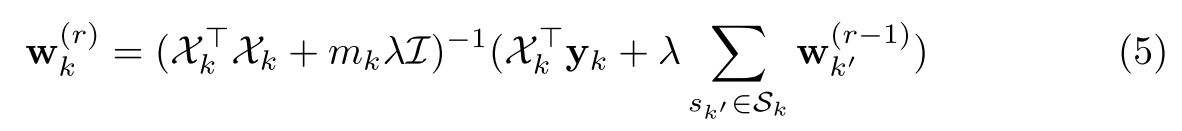


得到：

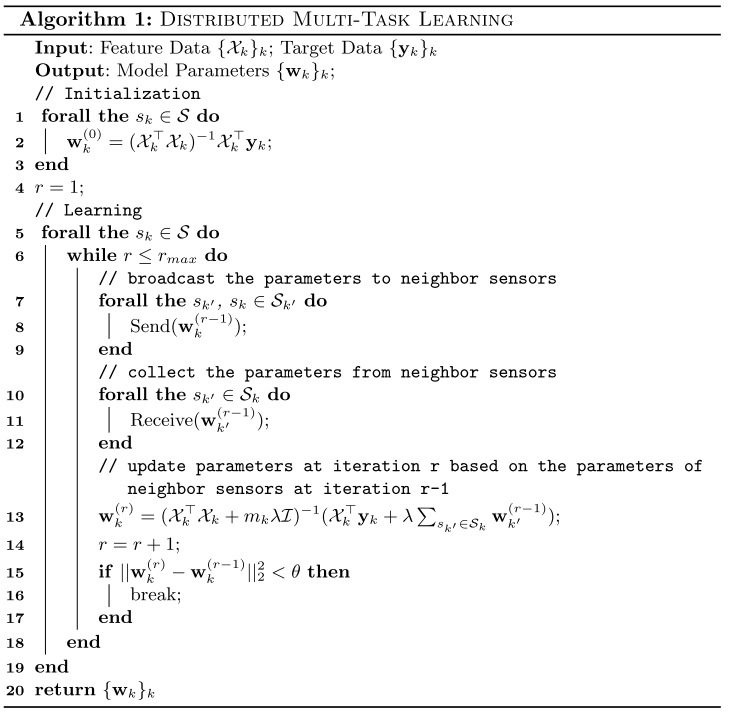


优化算法：

第r次迭代：



DMTL算法:



推导过程：

