总结报告28

论文Trace Norm Regularization: Reformulations, Algorithms, and Multi-task Learning

及论文A distributed Frank–Wolfe framework for learning low-rank matrices with the trace norm

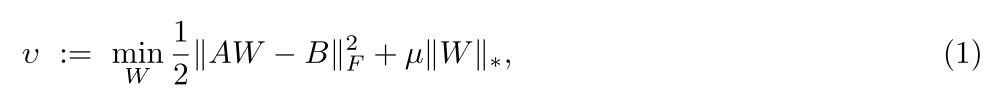
（2020.4.11）

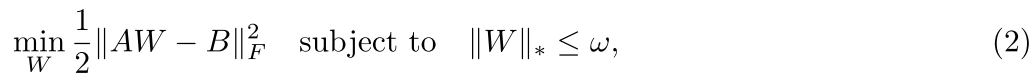
一、论文1

Pong, T. K., Tseng, P., Ji, S., & Ye, J. (2010). **Trace norm regularization: Reformulations, algorithms, and multi-task learning**. SIAM Journal on Optimization, 20(6), 3465–3489.

1. 目标函数

Primal：



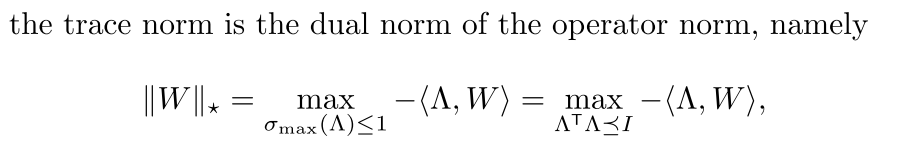




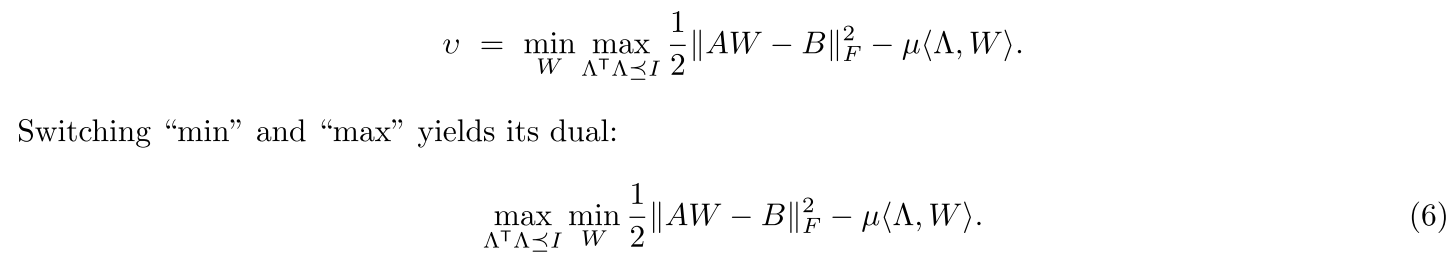
1. 当A∈Rm\*n，rank（A）=r<n时，可以归成r=n的情形，所以，下面针对r=n的情形推出dual：

**推导dual的过程：**

**首先，把核范数写成对偶的形式，带入primal**

where Λ ∈ Rr×m.

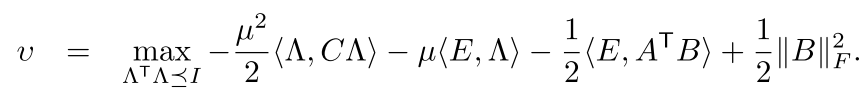
（1）可以写成：



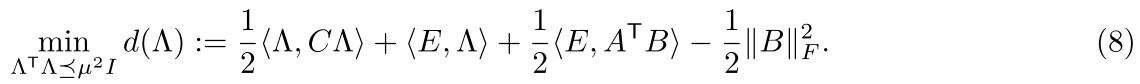
**然后，对W求导，解得min在W=μCΛ+E取得**



带入（6），



改变负号成min，再把Λ由μ scale，得到**对偶形式（8）**：



另外，原始解和对偶解的最优互相取得，根据



**一个限制条件是：Since the dual feasible set {Λ | ΛTΛ =I} is convex and compact, there is no duality gap, i.e., the optimal value of (6) equals v**

when r ≥ m, minimizing a linear function over the constraint set of (8) and projecting onto it requires **only an SVD of an r × m matrix**. This will be **key** to the efficient implementation of solution methods for (8).

1. 解决方法

|  |  |  |
| --- | --- | --- |
| **Class** | **Approaches** | **Details** |
| Dual gradient method | Dual conditional gradient (DCG) |  |
| Dual gradient-projection (DGP)  上一个方法slow，所以利用Gradient-projection method to be faster |  |
| Dual accelerated gradient-projection (DAGP)  Gradient-projection method can be accelerated using Nesterov’s extrapolation technique |  |
| Primal gradient method | Primal accelerated proximal gradient (PAPG) |  |
| Primal accelerated proximal gradient-average (PAPGavg) method  the method uses a **weighted average** of all past gradients instead of the most recent gradient  **more** **robustly** than PAPG |  |

二、论文2

Zheng, W., Bellet, A., & Gallinari, P. (2018). **A distributed Frank–Wolfe framework for learning low-rank matrices with the trace norm**. Machine Learning, 107(8–10), 1457–1475.

1. 目标函数

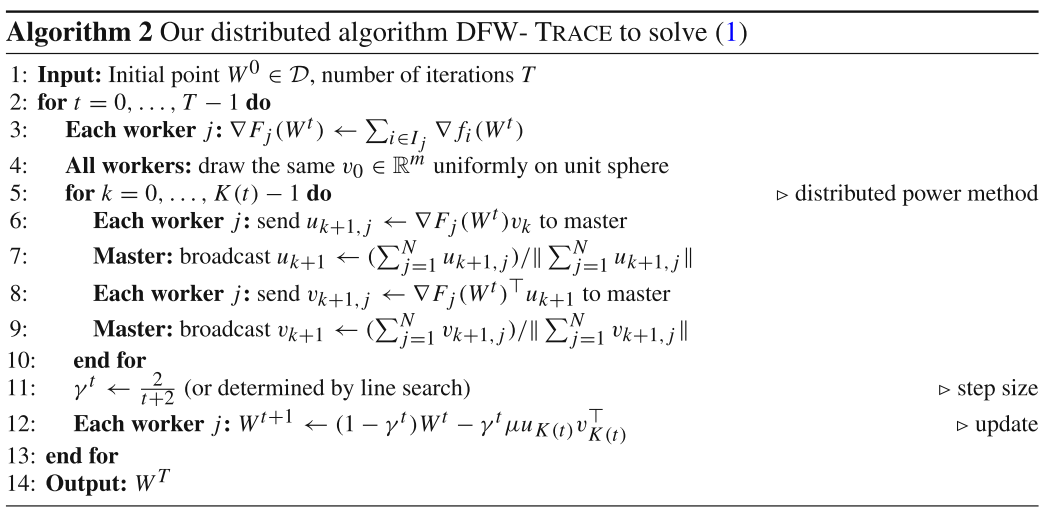


1. Approach——DFW-TRACE（Distributed Frank-Wolfe algorithm）

提出的方法结合了下面两个strategies：

1. Naive DFW：running the centralized Frank-Wolfe Algorithm in the distributed setting.每一个worker计算梯度后分送给master，master聚合总的梯度更新W。
2. Singular Vector Averaging：the master averages the singular vectors (weighted proportionally to nj ), resolving the sign ambiguity by setting the largest entry of each singular vector to be positive and using appropriate normalization.

DFW-TRACE结合了上面两个strategies，就会有Singular Vector Averaging的communication cost，和Naive DFW的收敛速率。



单位球

v右奇异向量

u左奇异向量

三、summary

1．这两篇论文针对trace norm，都用到了**SVD**的方法，都在试图减少SVD所带来的计算成本。第一篇是只计算原矩阵中的一个子矩阵的SVD分解，而第二篇用的是左右最大的奇异向量，因为trace norm是一些rank=1的凸包，所以一定存在rank=1的子问题的解。

2. 第一篇论文中的方法是不断地加速收敛速度，从一开始的条件梯度，再到应用projection（其实就是一种proximal？），再到使用Nesterov’s accelerate。

3. 第一篇论文的dual推导要求primal和dual是强对偶的。这样算法的终止条件就可以利用duality gap is small，也就是计算出Λ，可以得到W，将两者分别带到d（Λ）和p（W），如果这两个函数相差不大，说明算法可以停止了。可以每10次迭代，验证一下终止条件。

4. 第二篇论文的算法是基于FW算法应用到trace norm上改进的。FW算法就是：将问题线性化，得到这次迭代的可行解S，再和上次迭代的解Wt相组合，通过stepsize加权得到新的Wt+1。

四、问题

1. 第一篇论文gradient-projection的方法是proximal的一种吗？或者projection怎么实现的加速呢？