总结报告26

Primal-dual Problem

（红字为修改部分）

（2020.4.6-2020.4.7）

一、理论基础

|  |  |  |  |
| --- | --- | --- | --- |
| Source | Details | Remark | |
| 张贤达：矩阵分析与应用P216-P222 |  | 对于正交约束问题，ΘTΘ-I=0，由于它不是仿射函数（大概是说形如Ax+b，线性函数）的形式，所以，它**不能够**借助Lagrangian松弛方法换成一个凹函数的对偶无约束问题。  对于Loss（f0（x））它可以不是凸函数。对偶之后仍然是凹函数优化问题，但是这个时候就会有duality gap，不是原问题的最优解，而是次优解。 | |
| Yuxin Chen（2019），Princeton University：Dual and primal-dual methods PPT  （接上） | Dual的推导：  Primal-dual的推导：    Can we update both primal and dual variables simultaneously and take advantage of both proxf and proxh?也就是这种情况利用primal-dual problem | 从推导可以看出，是否把f（x）+<λ，Ax>结合起来写成共轭的形式，就决定了最后到底产生的是dual形式还是primal dual形式。  Dual形式，只有λ一个变量，很多文献也给出了λ转化成x的公式；  而primal-dual形式，有x和λ两个变量，也就是一个鞍点问题。  这个PPT当中也给出了dual和primal-dual的prox算法，原始-对偶在每一次迭代中，原始变量和对偶变量都要更新，而dual可以只更新对偶变量，最后根据等式转换成原始变量解，（~~当然也可以x和λ都更新：delete说的不准确~~）。  这个PPT中dual和primal-dual methods~~本质（delete，说的不准确）~~区别在于是否原始变量和对偶变量的迫近算子都用到了。  目标函数是： | |
| TinTin博客：Primal-dual problem  （Posted by Tintin on April 20, 2019）  <https://tintin.space/2019/04/20/Primal/> | Primal： | 其中，是regularizer，f(x)是分布式的loss | |
| 百度文库：对偶和鞍点问题PPT |  | | 鞍点定理其实和第一行张贤达说的是一个道理，要转换成鞍点问题，约束要满足凸函数、线性函数的条件；  鞍点与KKT条件之间的关系表明，满足KKT条件的点就是鞍点。 |

二、论文

|  |  |  |
| --- | --- | --- |
| Title | Objective functions | Note  (The property of regularizer) |
| Shalev-Shwartz, S., & Zhang, T. (2013). **Stochastic Dual Coordinate Ascent methods for regularized loss minimization**. Journal of Machine Learning Research, 14(1), 567–599. | Primal:  Dual: | 正则项：L2-norm |
| Yang, T. (2013). **Trading computation for communication: Distributed stochastic dual coordinate ascent**. Advances in Neural Information Processing Systems, 1–9. | Primal:    Dual: | 正则项：g(w)要求是凸函数，但不要求光滑。比如，g（w）中含有L1-norm |
| Zheng, S., Wang, J., Xia, F., Xu, W., & Zhang, T. (2017). **A general distributed dual coordinate optimization framework for regularized loss minimization.** Journal of Machine Learning Research, 18, 1–52. | Primal：  分布式原始：  分布式对偶： | g(w) is a strongly convex regularizer,h(w) is another convex regularizer  比如： |
| Zhang, Y., & Xiao, L. (2017). **Stochastic primal-dual coordinate method for regularized empirical risk minimization**. Journal of Machine Learning Research, 18, 1–42. | **Stochastic Primal-Dual Coordinate (SPDC)method：**basic idea: to approach the saddle point of f(x, y) defined in (4), we alternatively maximize f with respect to y, and minimize f with respect to x. Since the dual vector y has n coordinates and each coordinate is associated with a feature vector ai ∈ Rd, maximizing f with respect to y takes O(nd) computation, which can be very expensive if n is large. We reduce the computational cost by **randomly picking a single coordinate of y at a time, and maximizing f only with respect to this coordinate**. Consequently, the computation of each iteration is O(d). | g（x）凸 |
| Chambolle, A., & Pock, T. (2011). **A first-order primal-dual algorithm for convex problems with applications to imaging. Journal of Mathematical Imaging and Vision**, 40(1), 120–145. <https://doi.org/10.1007/s10851-010-0251-1> |  | 正则项要求凸。 |
| 鞍点问题和约束优化的几个一阶算法 尤燕飞（南京大学博士论文2015） |  | 凸，非光滑。 |
| Tan, C. (2018). **Stochastic Primal-Dual Method for Empirical Risk Minimization with O（1）Per-Iteration Complexity**. Advances in Neural Information Processing Systems 31 (NIPS 2018).1, 1–10. |  | 有一类算法的目标函数为  f（w）+g（Aw），这里的A可以data matrix. |

三、summary

总的来说，对于应用primal-dual，因为Loss应该就是ERM，那么对于regularizer来说，都是要求是凸的，但是可以不光滑，比如L2-norm，L1-norm,至于非凸的目前还没有看到。根据张贤达的矩阵分析与应用，应该这种正交约束难以用对偶的方法解决，但是，一些论文给出了正交约束的算法的迭代步骤。也就是当单独优化Θ：，这种算法是有一些的。

对于regularizer来说，有的文章用核范数，谱范数来做正则，那么这两个范数如何做分布式？如何做原始对偶？他们的共轭函数的形式是什么样的？即：是否可以做出类似如下问题的分布式算法？原始对偶优化算法？



四、response

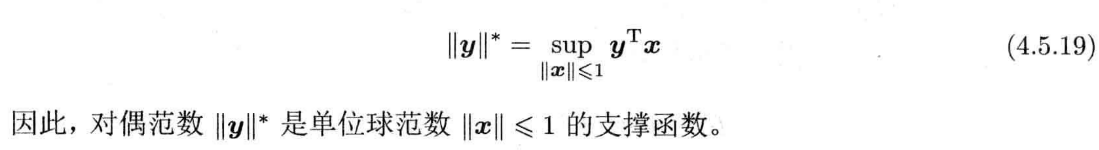
4.1 略

4.2略

4.3 首先要区别**对偶范数**和**范数的共轭函数**两个的概念

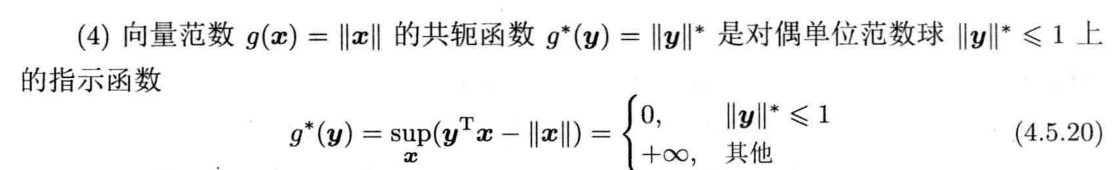
**·对偶范数:**

**定义** 向量范数||x||的对偶范数为||y||\*

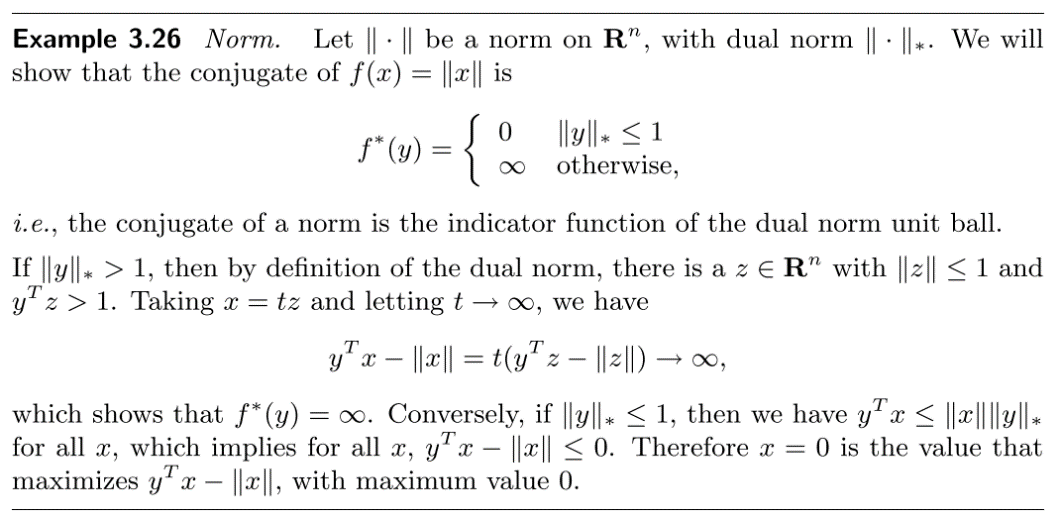


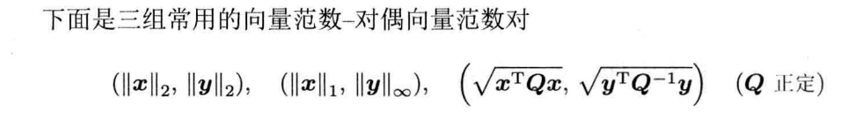
**·范数的共轭函数：**

**定义**

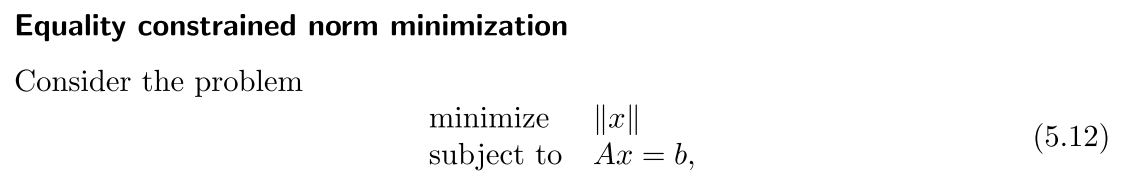
****

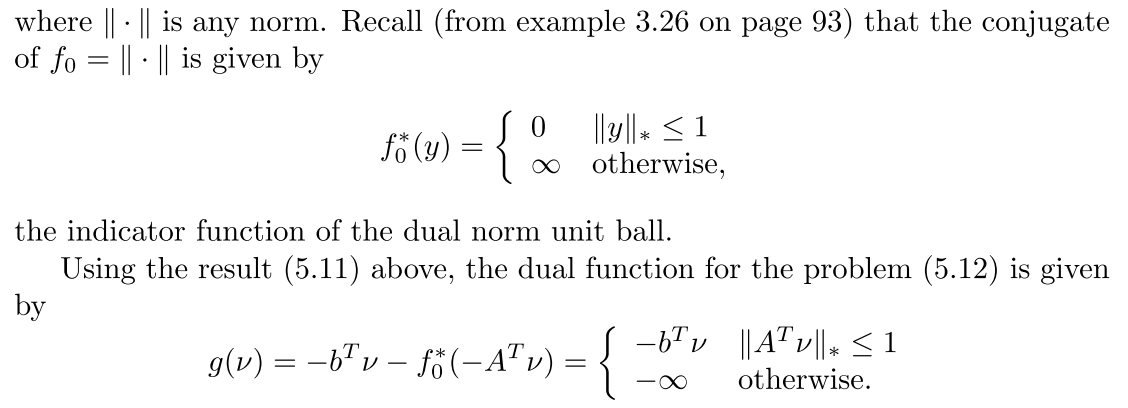
**举例：**



****

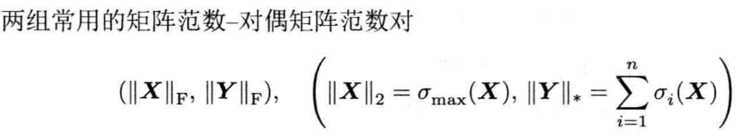
举例：带向量范数的对偶问题





**----------------------------------------------------------------------------------------------------------------------------------------------**

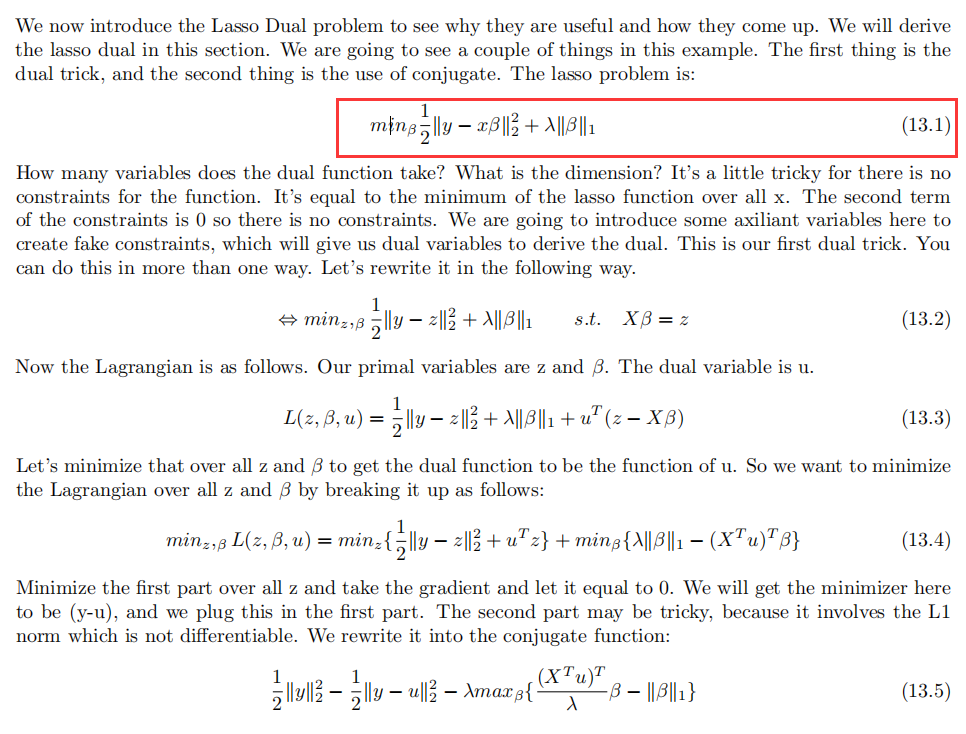
**以上说的都是针对向量的范数，对于矩阵范数：**

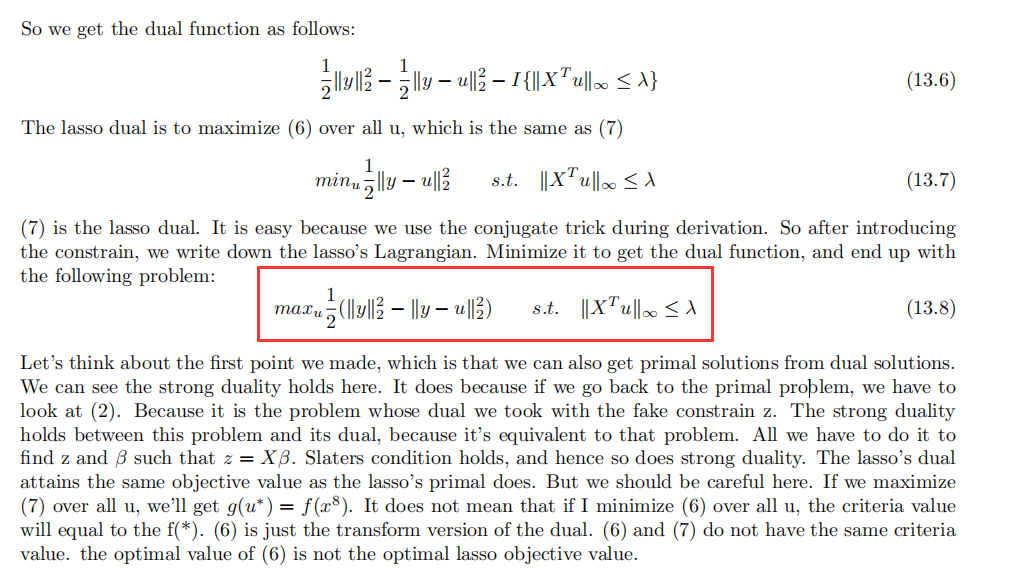
****

**矩阵核范数的对偶范数是谱范数。**

**弹性网的对偶函数：**

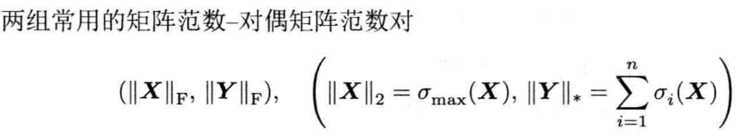
这个例子说明弹性网的对偶函数是可解的。





注：

1. 基本上查找了关于范数的对偶和共轭问题的资料，现在绝大部分都围绕着**向量范数**进行讨论的。对于矩阵范数的讨论，基本上只有这两个

****

而且，这仅仅是矩阵范数的对偶范数，如果想要去求其共轭函数，得出的结果也比较繁琐。并且目前还没有关于**矩阵范数**的共轭函数讨论的资料和实例，包括对于上面弹性网的实例，它也是定义在w（共识问题），而不是W上的。

1. 上面的论文，除了最后三篇，实际上都是DCA（Dual Coordinate Asent）算法的延伸和改进（***方法上的改进***），比如让它Stochastic，收敛的更快等等。所以，它们的论文当中，仅仅是对于方法上比较抽象的，比如他说regularizer g（x）是光滑的还是非光滑，会导致什么样的效果，它还没有给出一个非常具体的应用，比如还没有明确的给出一个目标函数长什么样，然后应该怎么推导。实验部分，也仅仅是非常简单的，比如针对只有L1-norm或者L2-norm的SVM问题，几种方法（比如和以前的SGD）比较一下效果怎么样之类的。（也就是现在的论文是对于DCA方法的改进和讨论，还没有实际应用）
2. 基于2），上面论文中给出的目标函数，可能用原始方法直接优化就十分简单，但是它可能为了讨论这种方法，就用对偶去做。至于给出一个具体的目标函数的时候是否满足它给的如下的比较抽象的条件（比如λ小、没有终止条件）从而用它的方法，就还得具体分析。

原问题直接用Stochastic gradient descent(SGD)可以解，但是SGD approach has several disadvantages. It does not have a clear stopping criterion; it tends to be too aggressive at the beginning of the optimization process, especially when λ is very small; while SGD reaches a moderate accuracy quite fast, its convergence becomes rather slow when we are interested in more accurate solutions.

So an alternative approach is dual coordinate ascent (DCA)

DCA是更早期的文章提出的，这篇论文提出的叫SDCA（Stochastic dual coordinate ascent），是对DCA的改进。

The purpose of this paper is to develop theoretical understanding of the convergence of the duality gap for SDCA.

1. 有一个视频是ICML 2017 Tutorial：**Zeyuan Allen-Zhu朱泽园：Recent Advances in Stochastic Convex and Non-convex**，这里面讲的primal和dual和primal-dual的若干论文方法一个总结，我明后两天可能要好好看看，以及他列出的论文的list。

<https://www.bilibili.com/video/BV1M441177kK>