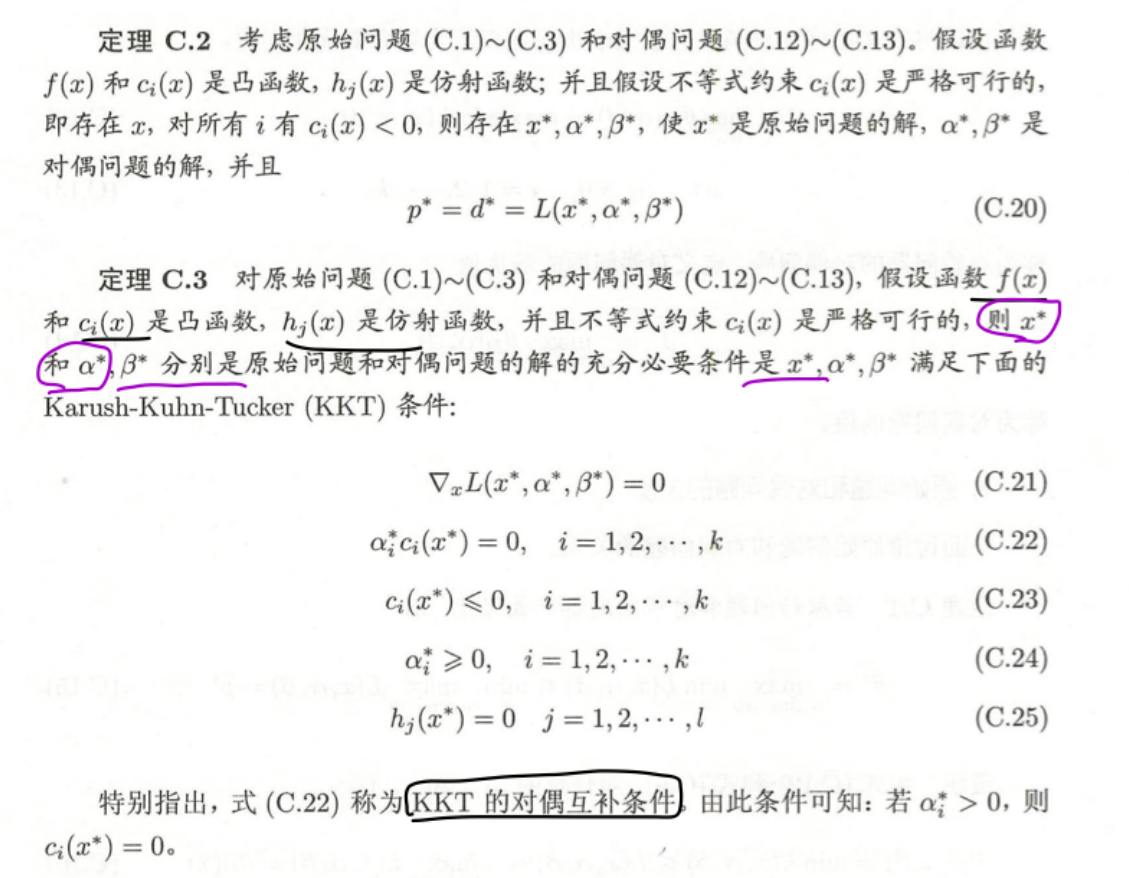
1. （必填）自己提出的问题的理解（罗列全部）：
2. 提出的问题1：和梯度下降法相比，牛顿法虽然具有收敛速度快的优点，但是求解海森矩阵的过程中计算二阶偏导数过程复杂，如何在选择这两中算法？

讨论后的理解：

1. 梯度下降法是一阶优化算法，牛顿法是二阶优化算法；
2. 牛顿法的收敛速度相比梯度下降法常常较快；
3. 牛顿法每次需要更新一个二维矩阵，计算代价很大，实际使用中常使用拟牛顿法；
4. 牛顿法对初始值有一定要求，在非凸优化问题中（如神经网络训练），牛顿法很容易陷入鞍点（牛顿法步长会越来越小），而梯度下降法则很容易逃离鞍点（因此在神经网络训练中一般使用梯度下降法，高维空间的神经网络中存在大量鞍点）；
5. 梯度下降法在靠近最优点时会震荡，因此步长调整在梯度下降法中是必要的，具体有adagrad, adadelta, rmsprop, adam等一系列自适应学习率的方法；
6. 提出的问题2：P449 原始问题和对偶问题都有最优值的时候，满足什么条件两个问题的最优值相同？

讨论后的理解：



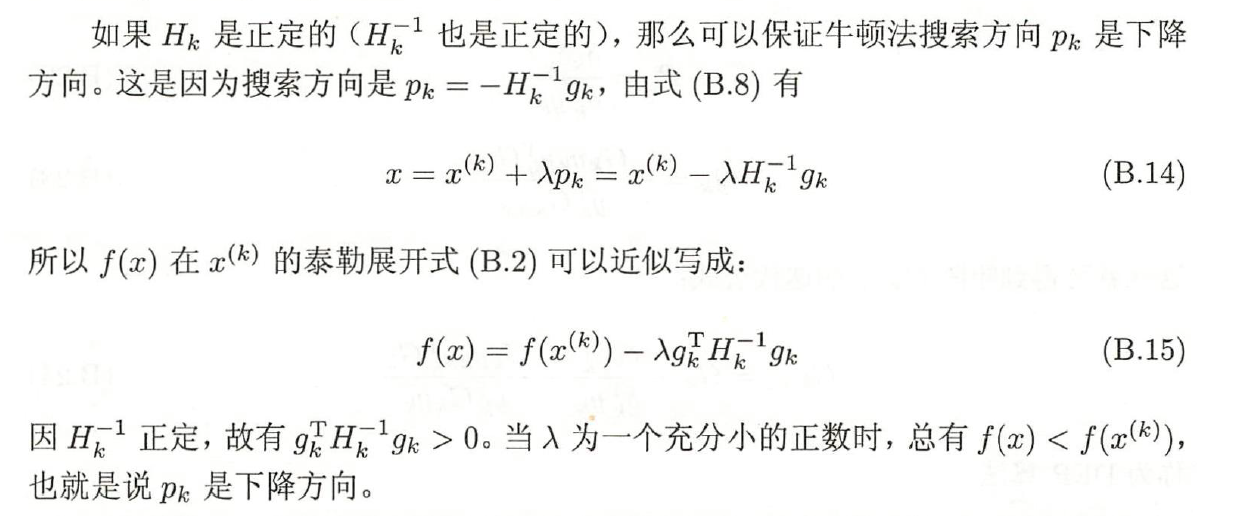
1. （必填）别人提出的问题的理解（选择几个问题罗列，并给出理解）：
2. 问题3：牛顿法和梯队下降法有什么区别和联系？

自己的理解：

* 1. 牛顿法：是通过求解目标函数的一阶导数为0时的参数，进而求出目标函数最小值时的参数，收敛速度很快；海森矩阵的逆在迭代过程中不断减小，可以起到逐步减小步长的效果。缺点：海森矩阵的逆计算复杂，代价比较大，因此有了拟牛顿法。
  2. 梯度下降法：是通过梯度方向和步长，直接求解目标函数的最小值时的参数。越接近最优值时，步长应该不断减小，否则会在最优值附近来回震；

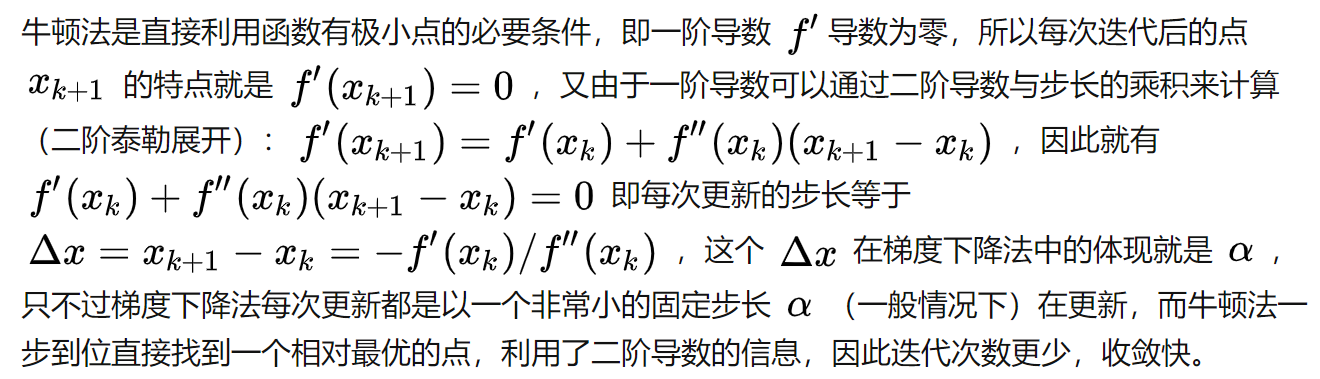
1. 问题4：为什么Hk是正定的可以保证牛顿法搜索方向pk是下降的方向

自己的理解：、



1. 问题5：从数学上来说，为什么牛顿法和拟牛顿法能够更快的找到极小值

自己的理解：



举例子，比如一个简单的强凸函数，无论起始点是哪，牛顿法总能一次迭代就找到最小值点0, 而梯度下降法只能根据步长和起始点得位置慢慢逼近0，但是实际上函数可能要复杂得多，函数可能是整体上非凸，但是局部是凸函数（比如最优点附近），直接用牛顿法迭代多次也不能保证收敛到最优点，因此需要先用梯度下降找到一个相对好的解后再用牛顿法可能效果比较好（根据曾文俊的回答），从这里我们也可以看出牛顿法对函数的性质以及初始点的位置选择比较挑剔，另外一个是二阶导数矩阵的逆矩阵计算量比较大（当为参数矩阵时，此时为一个二阶偏导数矩阵，即Hesse 矩阵），通常使用拟牛顿法。

1. （必填）读书计划

1、本周完成的内容章节：《统计学习方法》附录A-C

2、下周计划：《统计学习方法》第七章

四、（选做）读书摘要及理解或伪代码的具体实现（读书摘要、伪代码的具体实现代码等可以写到这个部分）

读书摘要及理解（选做）

关于对偶问题:

1、为什么要使用对偶问题（SVM）

1. 对偶问题将原始问题中的约束转为了对偶问题中的等式约束
2. 方便核函数的引入
3. 改变了问题的复杂度。由求特征向量w转化为求比例系数a，在原始问题下，求解的复杂度与样本的维度有关，即w的维度。在对偶问题下，只与样本数量有关。

2、对偶问题就是使求解更加高效且目标函数值不变，通过先消去w，b，得到关于α的函数，然后单独计算 α，通过得到的α反求w，b，最后获得超平面的参数，相比于先对α的不等式约束进行计算，对偶的方式使得计算更加便捷。另外KKT条件就是在约束下求得目标函数极值时αi满足的条件，只有满足了kkt条件，才算是满足了目标函数和约束函数，因此之后介绍的计算迭代算法也是基于KKT条件，通过不断修改不满足KKT条件的α，使其满足KKT条件，从而求出目标函数的最优值。下篇文章将主要推导一种计算w和b的高效算法-SMO算法，看看实际中如何通过对偶问题公式推出α，从而由α推出w和b.