

homework-sto

黄晃 数院 1701210098

2018 年 6 月 17 日

1 问题

原问题为

$$\min \frac{1}{n} f_i(w) + \lambda \|w\|_1 \quad (1)$$

where $f_i(w) = \log(1 + \exp(-y^i w^T x^i))$ and $\lambda > 0$

为了使用stochastic optimization, 可以将问题改写成

$$\min \frac{1}{n} f_i(w) \quad (2)$$

where $f_i(w) = \log(1 + \exp(-y^i w^T x^i)) + \lambda \|w\|_1$

1.1 数据集

使用要求的两个数据集:MINIST和Coverttype.

MINIST 一共70000个样本,其中每个 x^i 是28*28的灰度矩阵向量化的结果,而 y^i 是该幅图片对应的label

Coverttype 一共581012个样本,其中每个 x^i 是54维向量

数据归一化 对于 $X = [x_1, x_2, \dots, x_N] = [a_1, a_2, \dots, a_p]^T$,我们对其进行行归一化,即取 $\hat{X} = [\hat{a}_1, \hat{a}_2, \dots, \hat{a}_p]^T$,其中 $\hat{a}_i = a_i / \max(a_i)$

1.2 算法

一共实现了要求的两个算法:SAG和SVRG以及附加任务:使用BB步长的SG方法

1.2.1 SAG

Algorithm 1: Learning paramters for BPR

输入:

The observed user-item pair set S ;
 The feature matrix of items F ;
 The content features entities $A := \{A^u, A^v\}$;

输出:

$\Theta := \{Y^u, Y^v\}$;
 $W := \{W^u, W^v\}$;

```

1 initialize the model parameter  $\Theta$  and  $W$  with uniform
    $(-\sqrt{6}/k, \sqrt{6}/k)$ ;
2 standarized  $\Theta$ ;
3 Initialize the popularity of categories  $\rho$  randomly;
4 repeat
5   Draw a triple  $(m, i, j)$  with 算法??;
6   for each latent vector  $\theta \in \Theta$  do
7      $\theta \leftarrow \theta - \eta \frac{\partial L}{\partial \theta}$ 
8   end
9   for each  $W^e \in W$  do
10    Update  $W^e$  with the rule defined in Eq.??;
11  end
12 until convergence;
```

$$w_{k+1} = w_k - \alpha_k \left(\frac{1}{n} (\nabla f_{s_k}(w_k) - g_{k+1}^{s_k}) + \frac{1}{n} \sum_{i=1}^N g_{k-1}^i \right)$$

其中

$$g_k^i = \begin{cases} \nabla f_i(w_k) & \text{if } i = s_k \\ g_{k+1}^{s_k} & \text{else} \end{cases}$$

所以需要存储每个 f_i 随着随机量 s_k 最近一次被计算时得到的值作为 g_k^i .此外,为了启动算法,在第一次迭代开始前,对每个 f_i 计算一次相应的次梯度

1.2.2 SVRG

如课件所示,每求一次完整的函数 f 的梯度,进行 m 次随机梯度的迭代,然后将结果的平均作为外部循环的结果.每次内循环的下降方向选择为

$$v_k = \nabla f_{s_k}(w_k) - \nabla f_{s_k}(y) + \nabla f(w_i)$$

其中 w_i 为外循环第 i 步的值

1.2.3 SG-BB(Extra-credit)

$$w_{k+1} = w_k + \alpha_k \nabla f_{s_k}(w_k)$$

其中

$$\begin{aligned} \alpha_k &= \frac{(s^{k-1})^T y^{k-1}}{(y^{k-1})^T y^{k-1}} \\ s^{k-1} &= w^k - w^{k-1}, \\ y^{k-1} &= g^k - g^{k-1} \end{aligned}$$

其中 $g^k = \nabla f_{s_k}(w_k)$ 为第 k 次计算的随机梯度,而不是完整函数的梯度

1.3 计算细节

在计算函数值以及梯度时会遇到如下两类问题

- 计算 $\frac{e^x}{1+e^x}$ 时 $e^x = \inf$
- 计算 $\log(1+e^x)$ 时 $e^x = \inf$

我们将其处理成 $\frac{e^x}{1+e^x} = 1, \log(1+e^x) = x$

2 计算结果

对 $\lambda = 10, 1, 0.1, 0.001, \frac{1}{n}$ 进行了实验

参数 SVRG中 $m = \frac{n}{10}$,SVRG以及SAG中 $\alpha = 0.01$,SVRG一共进行了20次迭代,SAG进行了n次迭代,sgBB进行了2n次迭代.

误差定义 参考文献定义Testing Error $R(w) = \sum_{i=1}^n \log(1 + \exp(-b^i w^T a^i))$

结果取样 由于 $R(w)$ 的计算量太大,在展示结果时,我们只等距的选取了20个点的值来作图

x轴 仿照参考文献,做了三类图,其x轴分别为迭代次数,求导次数,时间.在这里求导次数定义为单个 f_i 的求导次数.所以SVRG一次迭代求导次数了增加 $n+m$ 次.

图注 在每一组图的图注位置,在相应算法后记录最后时刻的 $R(w)$

2.1 MINIST

其中最后两幅图与其他略有区别:1 是去掉初始点后,SAG与SVRG的比较;2是去掉初始点后SAG的 $R(w)$.以上两幅图的单独列出是为了排除掉初始点,让函数值的变化看的更清晰(初始点值太大产生比例尺的干扰)

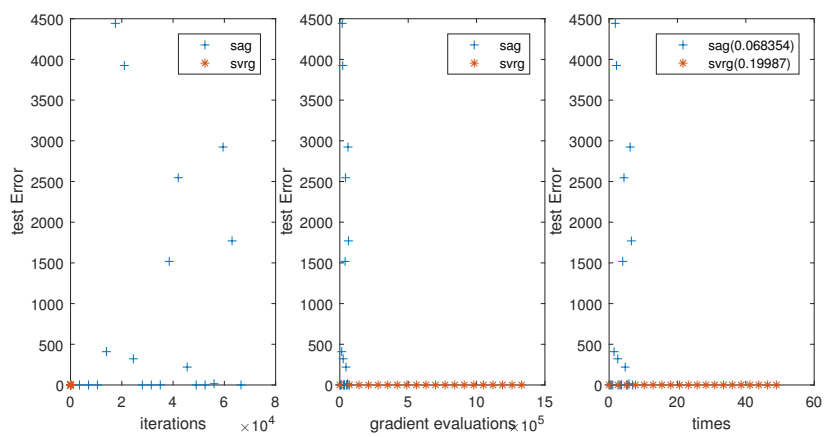
2.2 COVERTYPE

与之前一样,最后两幅图与其他略有区别:3 是去掉初始点后,SAG与SVRG的比较;4是去掉初始点后SAG的 $R(w)$.

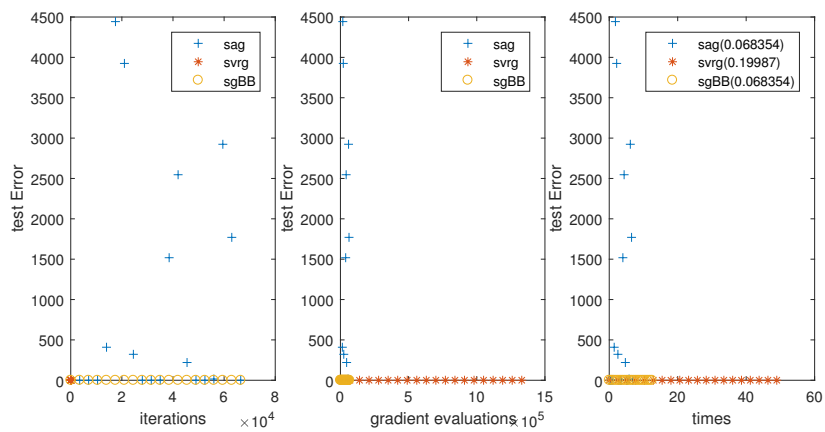
2.3 结果分析

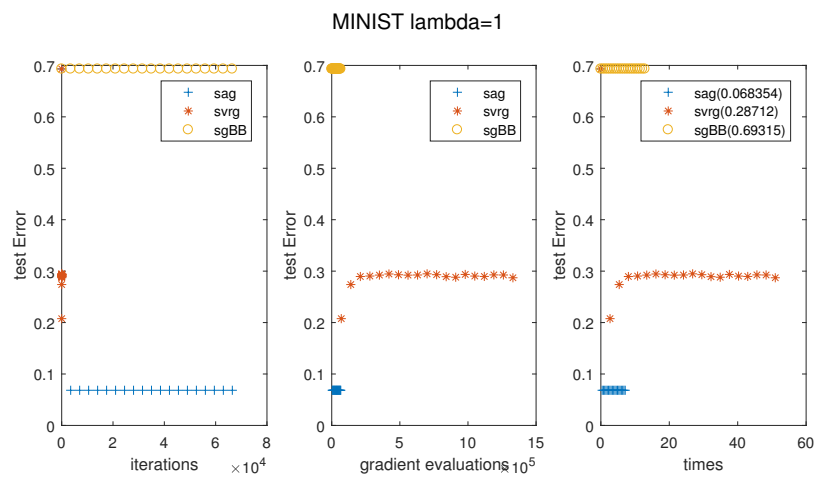
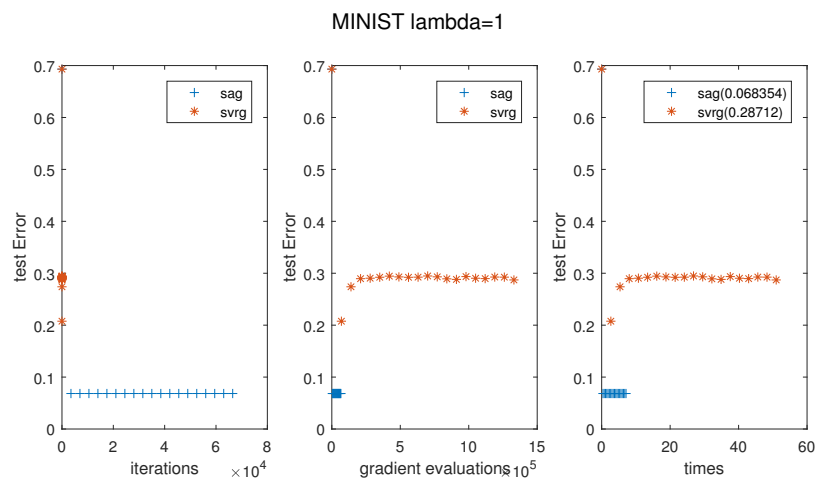
- 对于两个数据集,随着 λ 的减小,SVRG得到的ERROR随之减小,直至 $1/n$ 时,error能接近SAG的结果
- 对于 $\lambda \geq 0.1$ 的情况,SVRG虽然ERROR比SAG大,然而往往较为稳定
- 对于较大的 λ ,SAG的TESTING ERROR有较大的波动

MINIST lambda=10

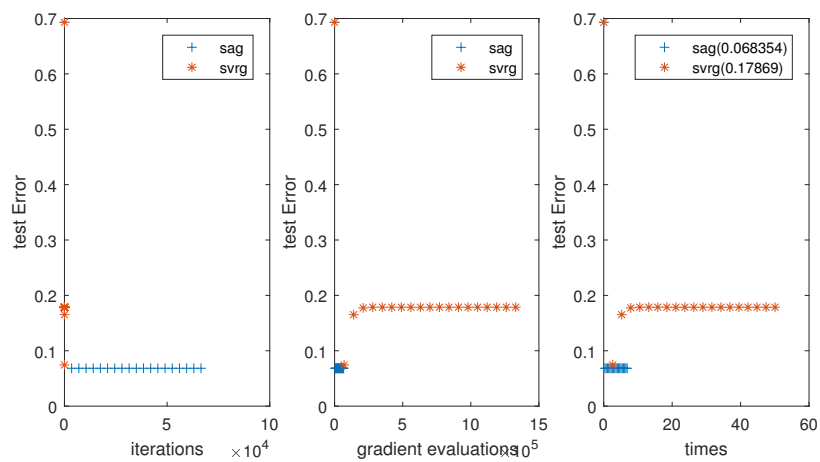


MINIST lambda=10

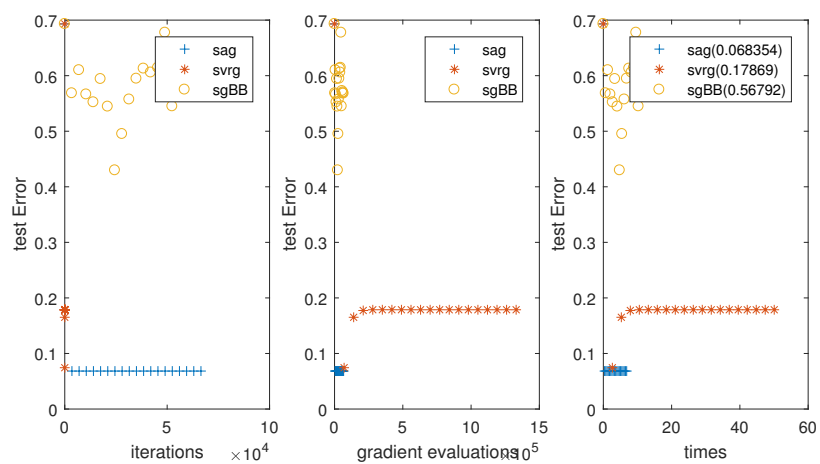


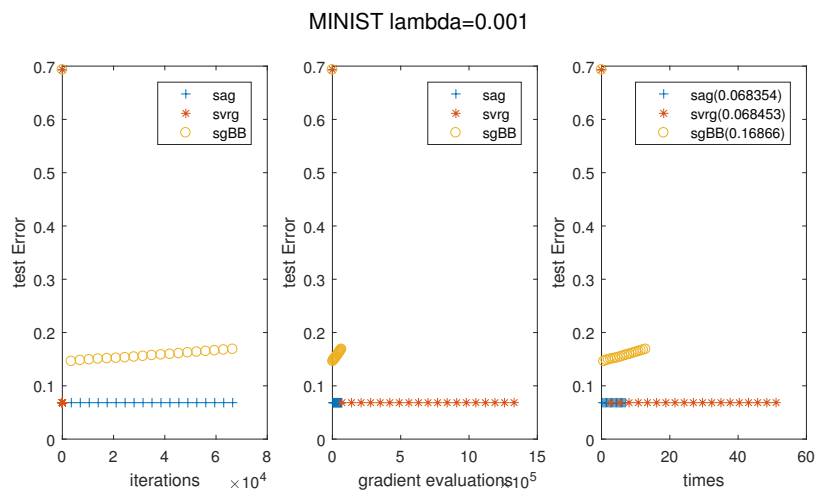
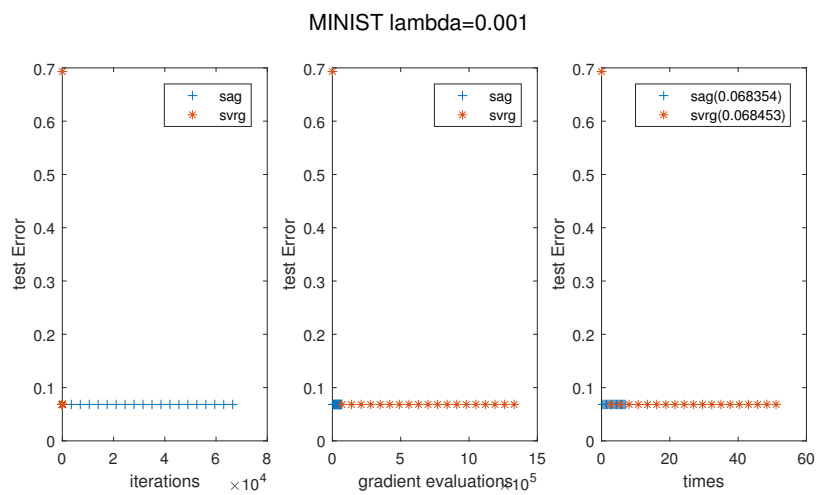


MINIST lambda=0.1

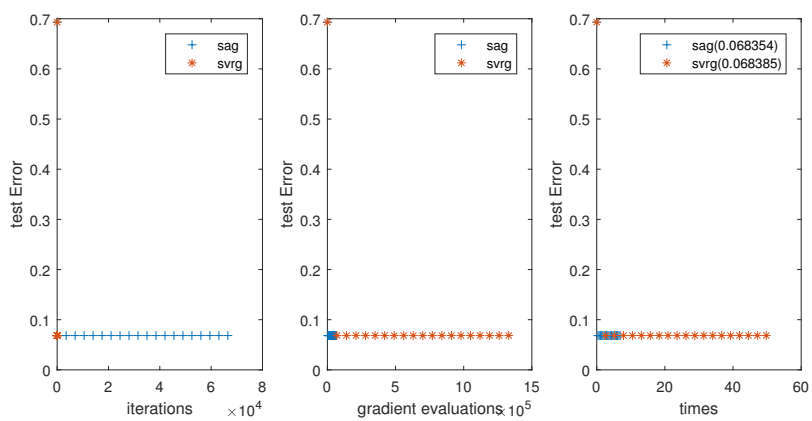


MINIST lambda=0.1

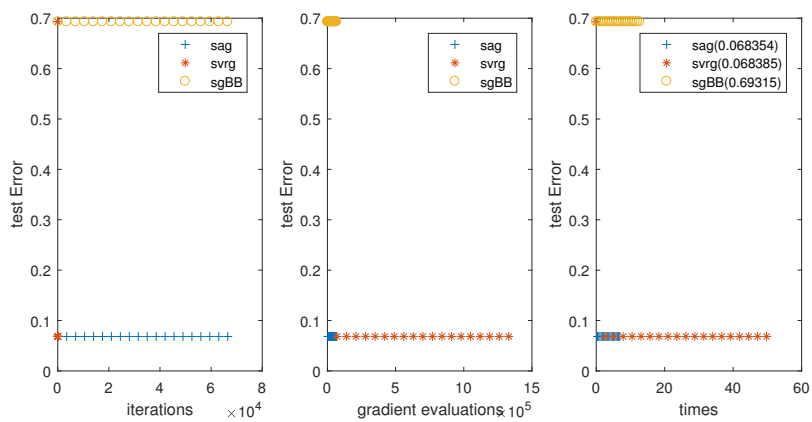




MINIST lambda=1.4286e-05



MINIST lambda=1.4286e-05



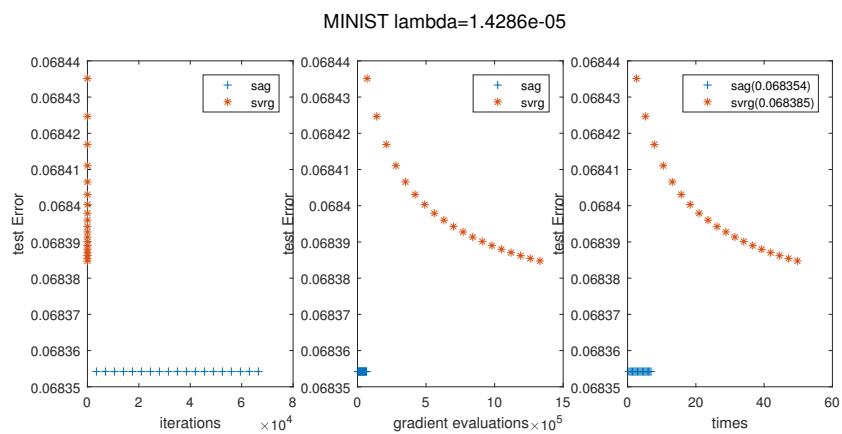


图 1:

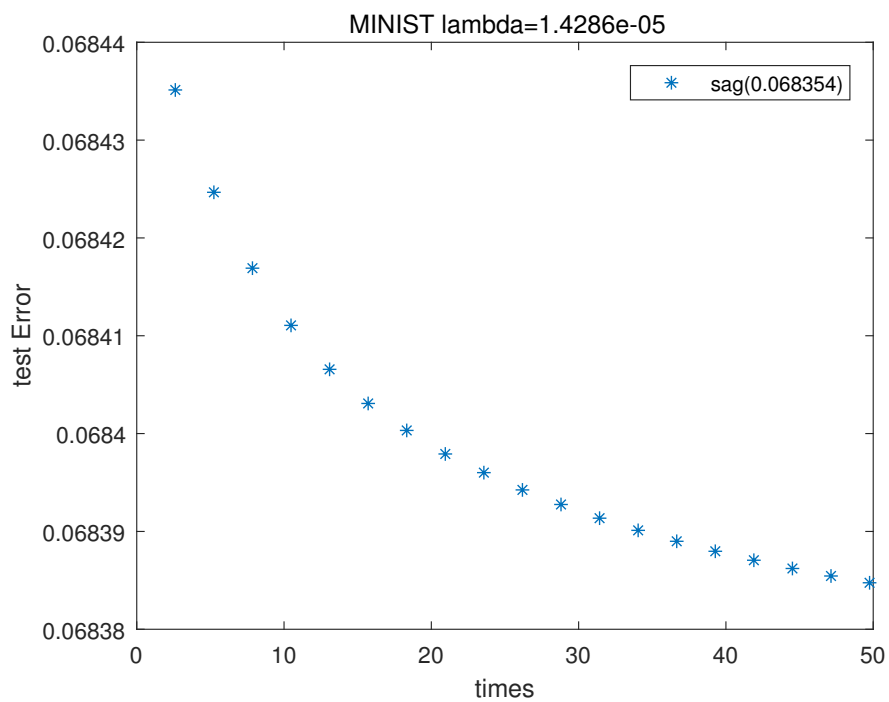
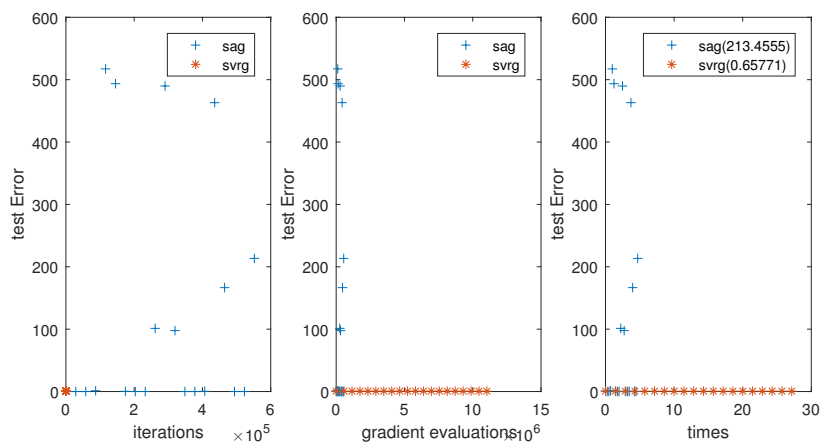
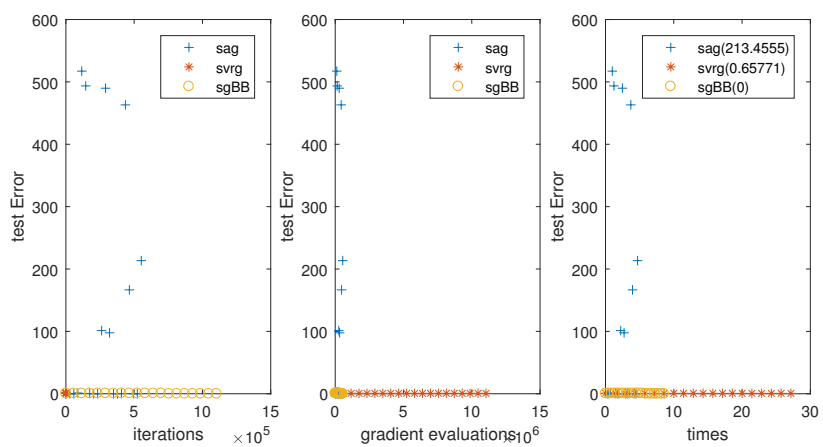


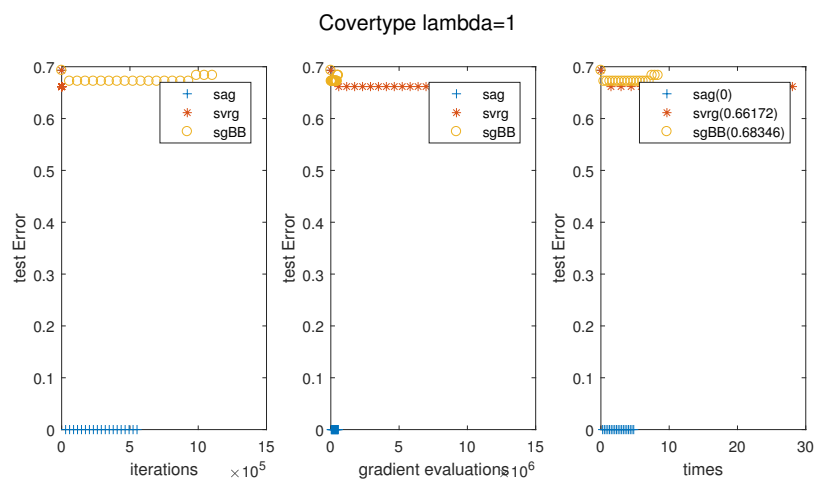
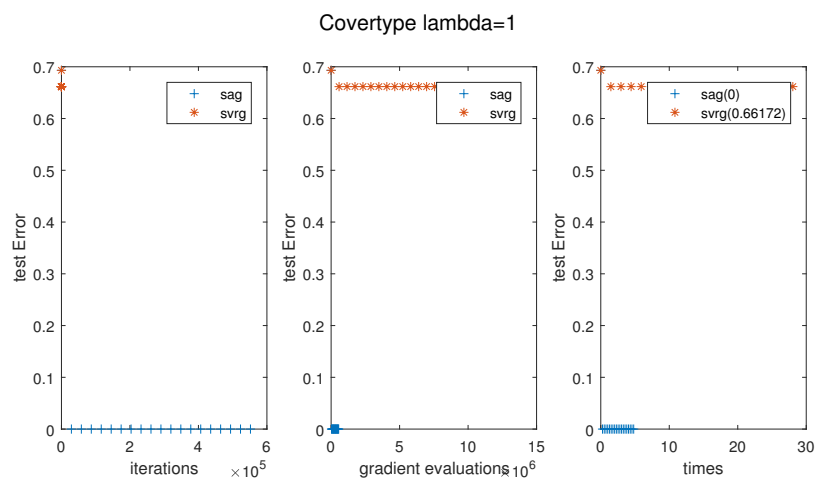
图 2:

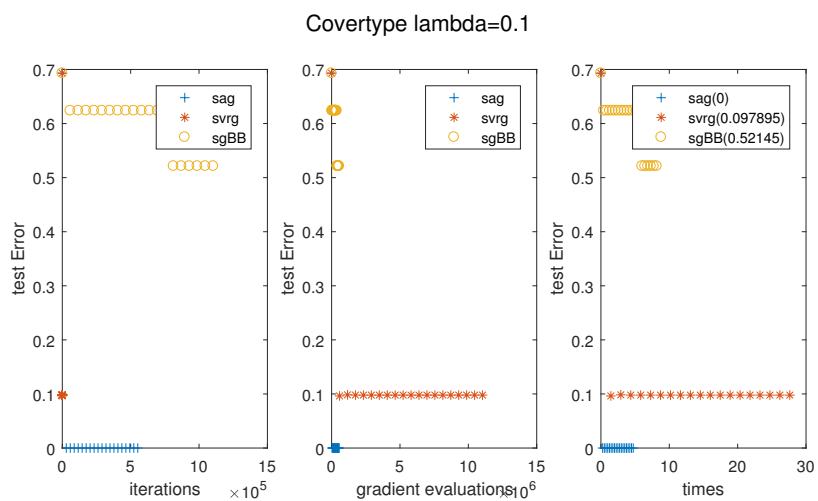
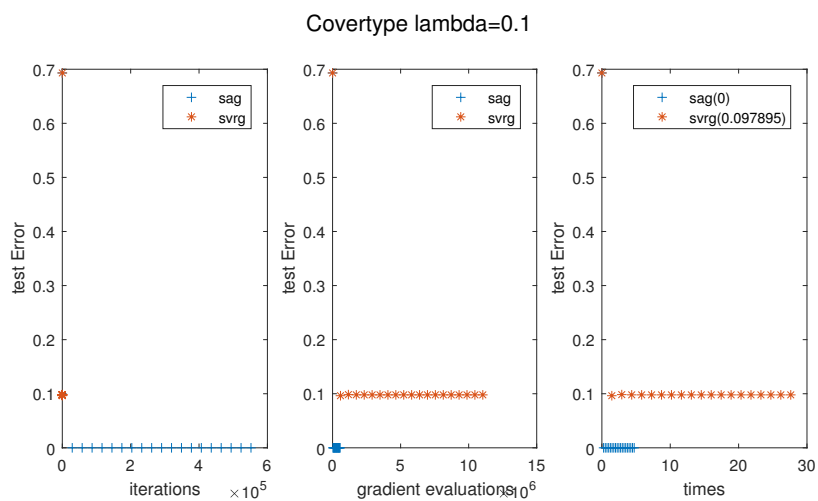
Covertypes lambda=10



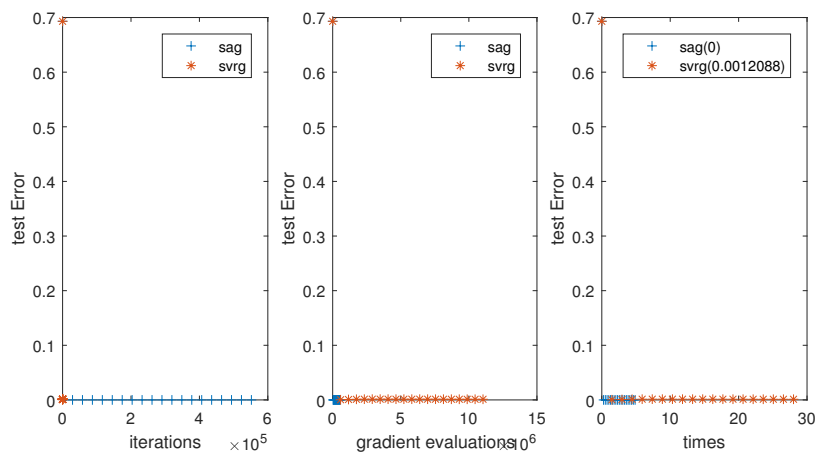
Covertypes lambda=10



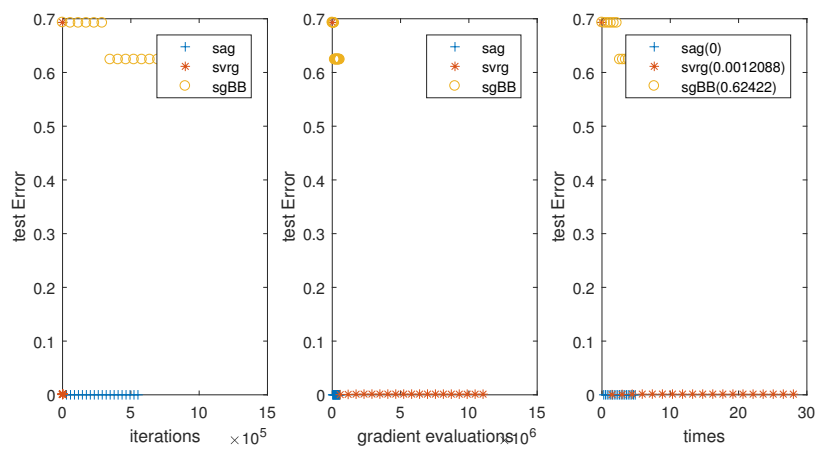


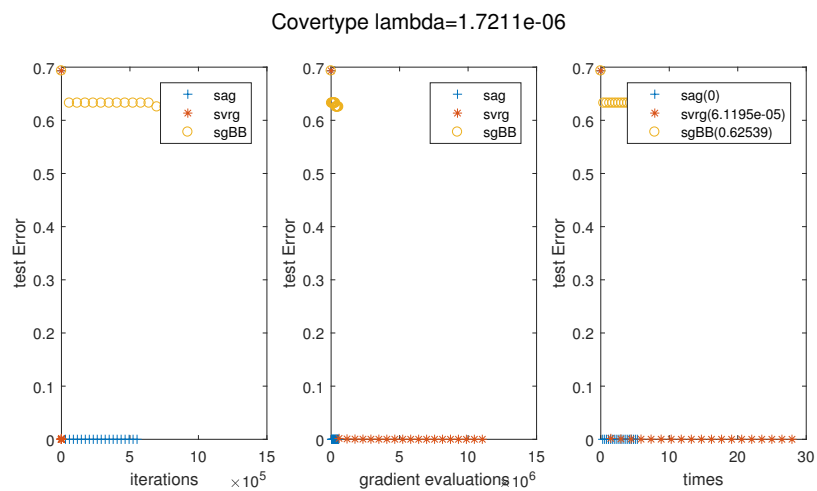
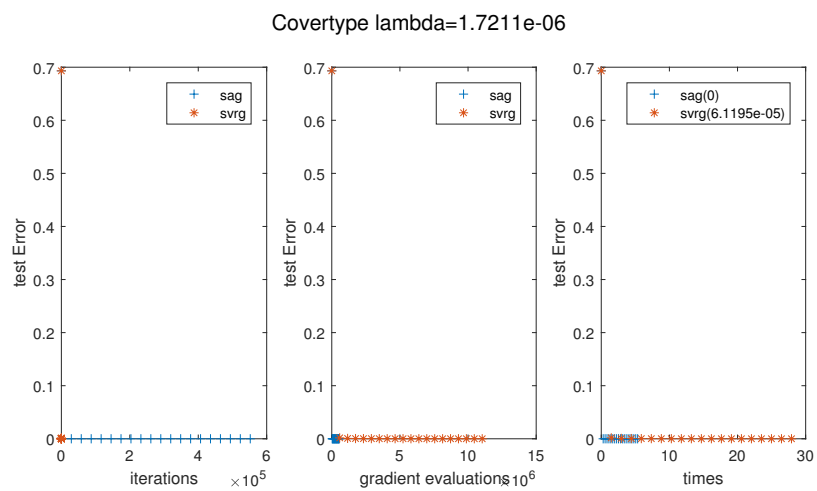


Covertypes lambda=0.001



Covertypes lambda=0.001





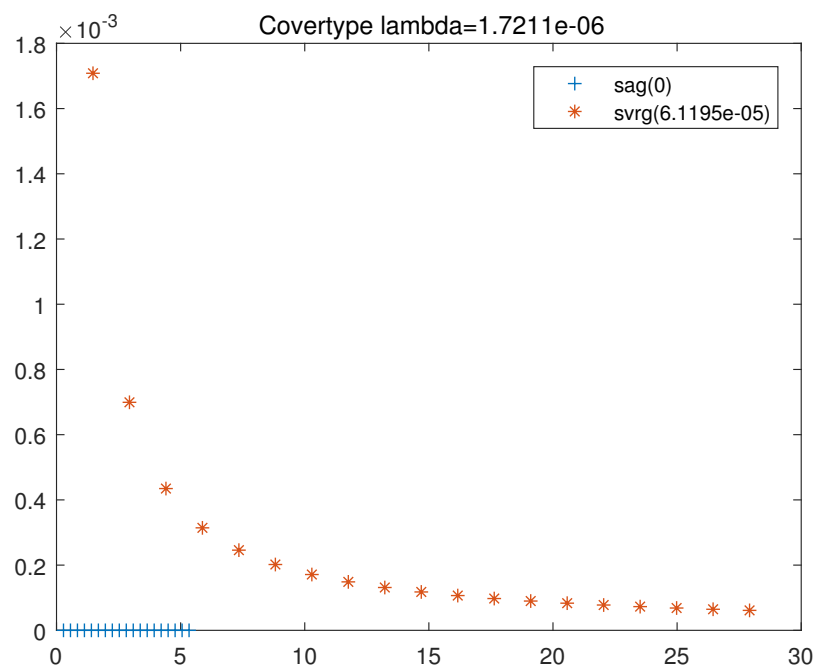


图 3:

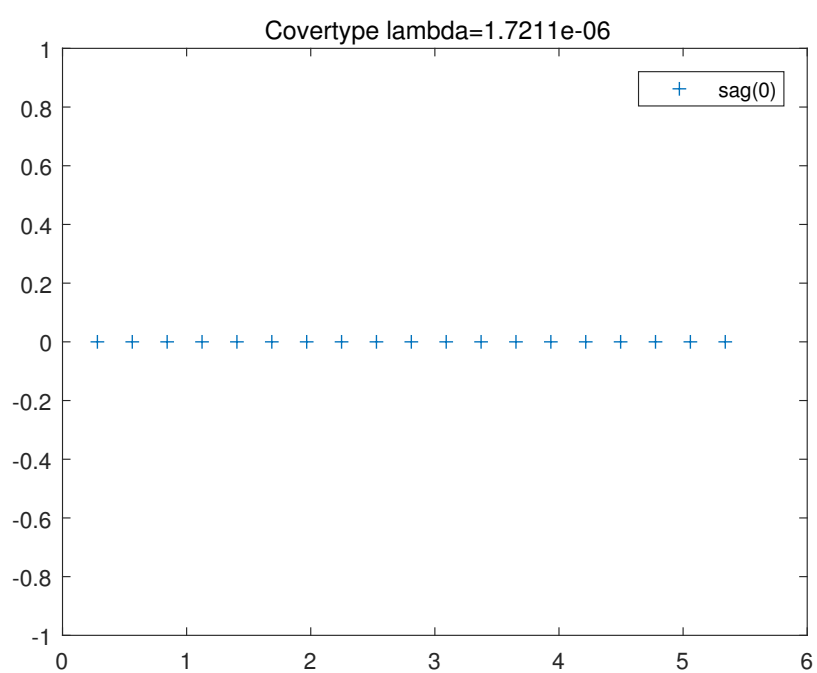


图 4:

- λ 较小时,SAG的结果比SVRG的结果要好
- 对于数据集MINIST,如图2所示,SAG的TESTING ERROR随计算时间单调减小
- 对于EXTRA-CREDIT中的BB步长SG方法,仅当 $\lambda = 10$ 时有最好的结果,其余的情况下ERROR较大