homework-sto

黄晃 数院 1701210098

2018 年 4 月 23 日

1 问题

原问题为

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

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

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

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

Covertype 一共581012个样本,其中每个x^i是54维向量

数据归一化 对于$X = [x_1, x_2, \cdots, x_N] = [a_1, a_2, \cdots, a_p]^T$,我们对其进行行归一化,即取$\hat{X} = [\hat{a_1}, \hat{a_2}, \cdots, \hat{a_p}]^T$,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

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

其中

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

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

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

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.

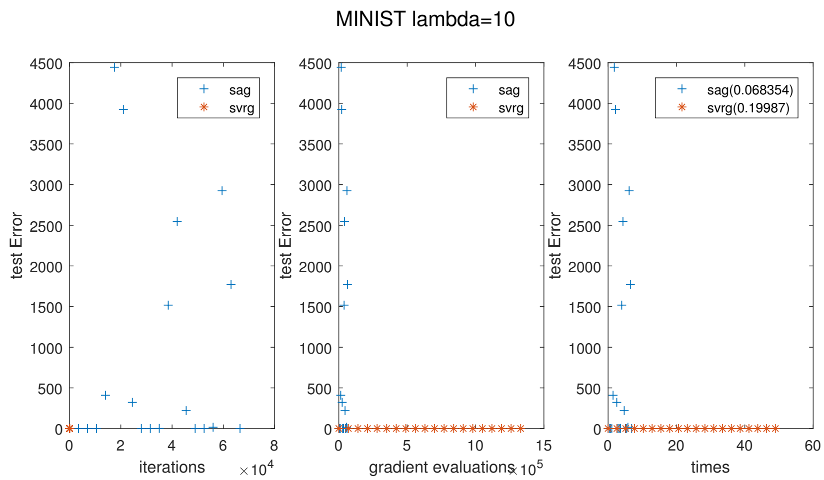
图注 在每一组图的图注位置,在相应算法后记录最后时刻的R(w)

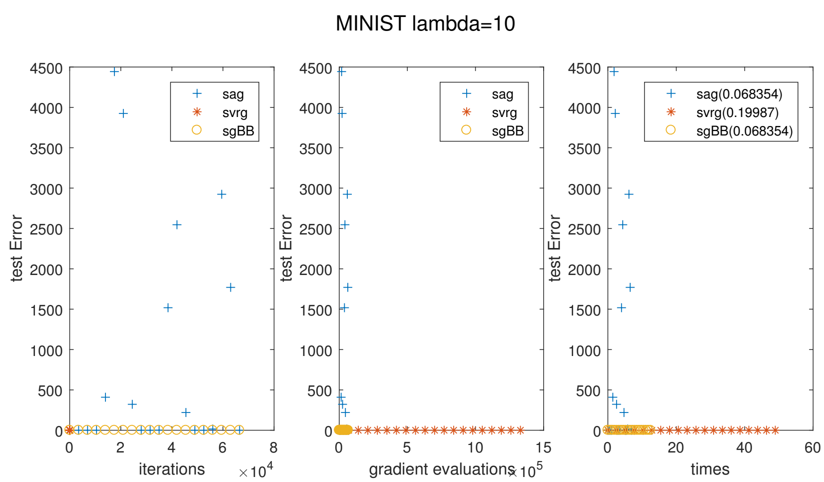
2.1 MINIST

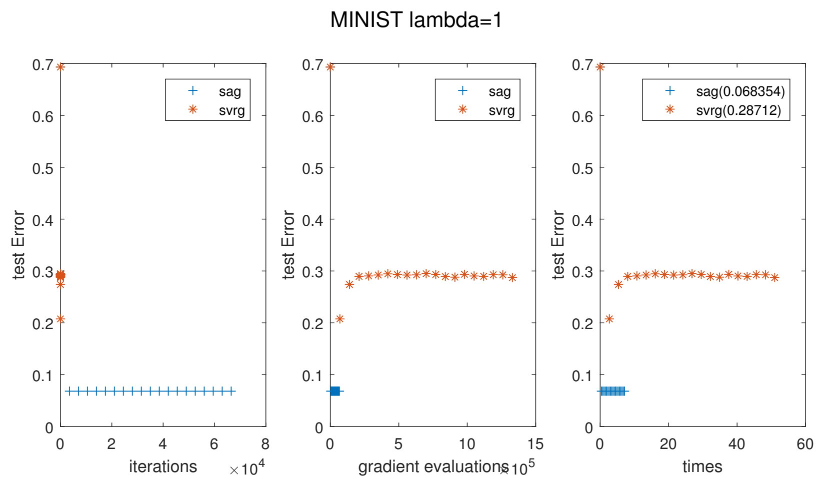
其中最后两幅图与其他略有区别:1 是去掉初始点后,SAG与SVRG的比较;2是去掉初始点后SAG的R(w).以上两幅图的单独列出是为了排除掉初始点,让函数值的变化看的更清晰(初始点值太大产生比例尺的干扰)

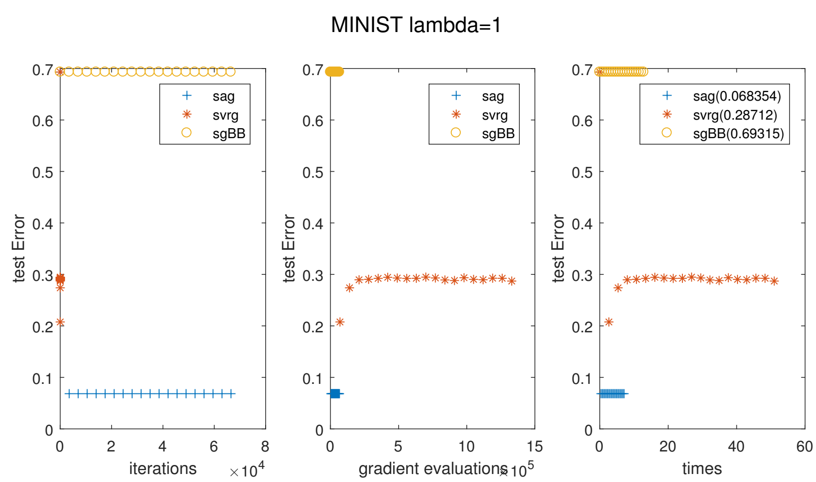
2.2 COVERTYPE

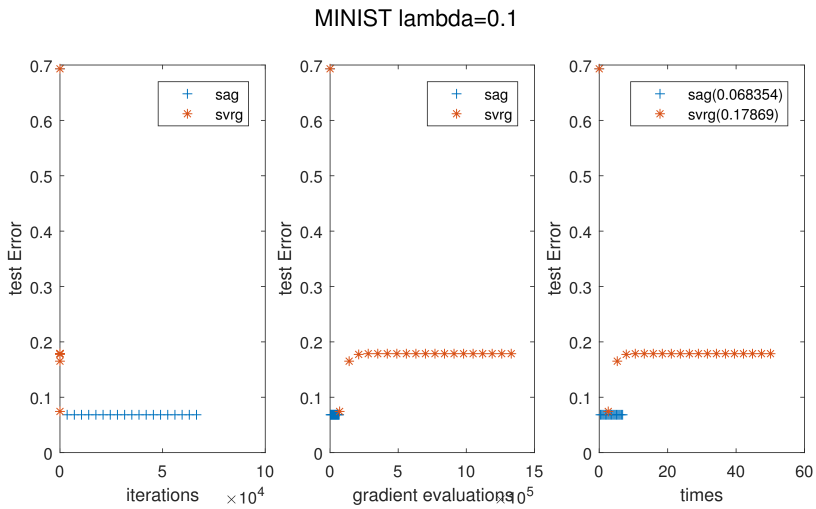
与之前一样,最后两幅图与其他略有区别:3 是去掉初始点后,SAG与SVRG的比较;4是去掉初始点后SAG的R(w).

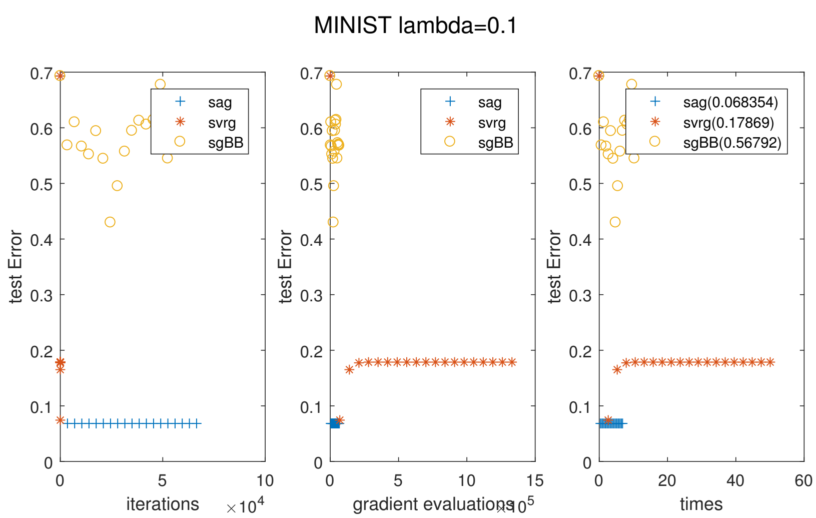

MINIST lambda=10
test Error
iterations
gradient evaluations
times
sag
svrg
sag(0.068354)
svrg(0.19987)

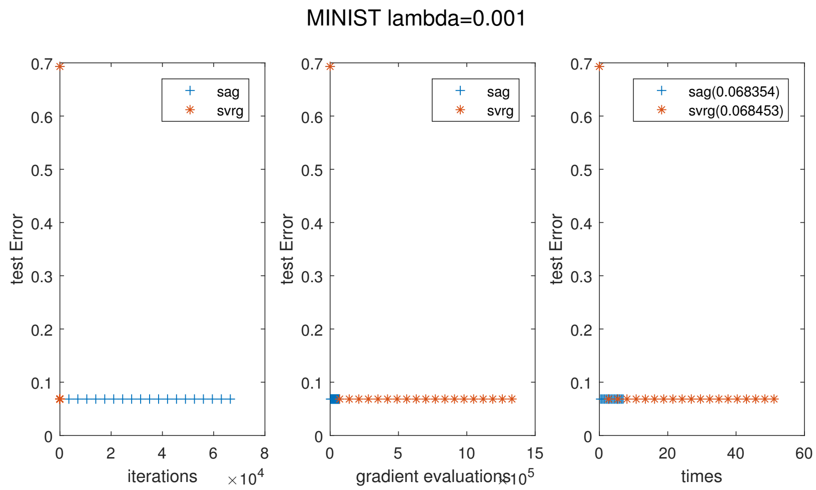

MINIST lambda=10
test Error
iterations
gradient evaluations
times
sag
svrg
sgBB
sag(0.068354)
svrg(0.19987)
sgBB(0.068354)

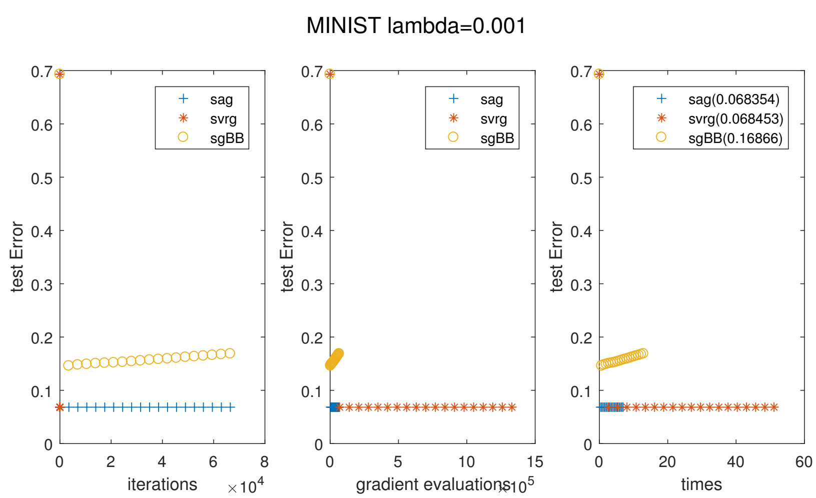

MINIST lambda=1
test Error
iterations
×10^4
gradient evaluations
×10^5
times
sag
svrg
sag(0.068354)
svrg(0.28712)

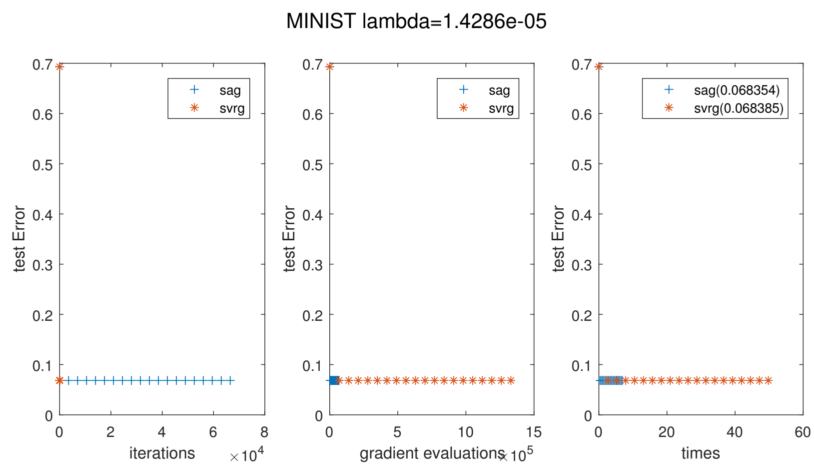

MINIST lambda=1
test Error
iterations
×10^4
gradient evaluations
×10^5
times
sag
svrg
sgBB
sag(0.068354)
svrg(0.28712)
sgBB(0.69315)

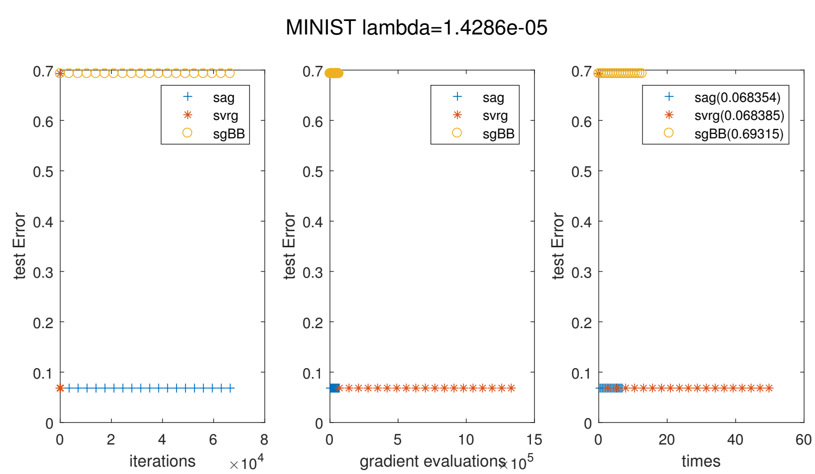

MINIST lambda=0.1
test Error
iterations
gradient evaluations
times
sag
svrg
sag(0.068354)
svrg(0.17869)


MINIST lambda=0.1
test Error
iterations
gradient evaluations
times
sag
svrg
sgBB
sag(0.068354)
svrg(0.17869)
sgBB(0.56792)


MINIST lambda=0.001
sag
svrg
sag(0.068354)
svrg(0.068453)
test Error
iterations
gradient evaluations
times
×10^4
×10^5


MINIST lambda=0.001
sag
svrg
sgBB
sag(0.068354)
svrg(0.068453)
sgBB(0.16866)
test Error
iterations
gradient evaluations
times
×10^4
×10^5


MINIST lambda=1.4286e-05
test Error
iterations
×10^4
sag
svrg
gradient evaluations
×10^5
times
sag(0.068354)
svrg(0.068385)


MINIST lambda=1.4286e-05
test Error
iterations
×10^4
sag
svrg
sgBB
gradient evaluations
×10^5
times
sag(0.068354)
svrg(0.068385)
sgBB(0.69315)

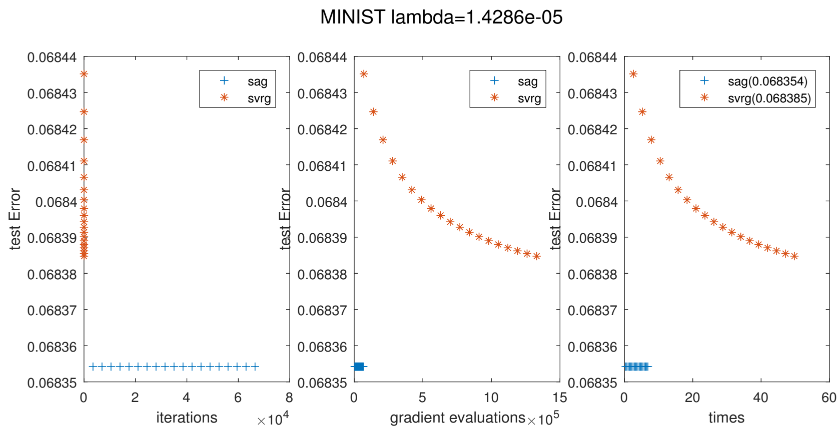

MINIST lambda=1.4286e-05
test Error
0.06844
0.06843
0.06842
0.06841
0.0684
0.06839
0.06838
0.06837
0.06836
0.06835
sag
svrg
sag(0.068354)
svrg(0.068385)
iterations
$\times 10^4$
gradient evaluations
$\times 10^5$
times

图 1:

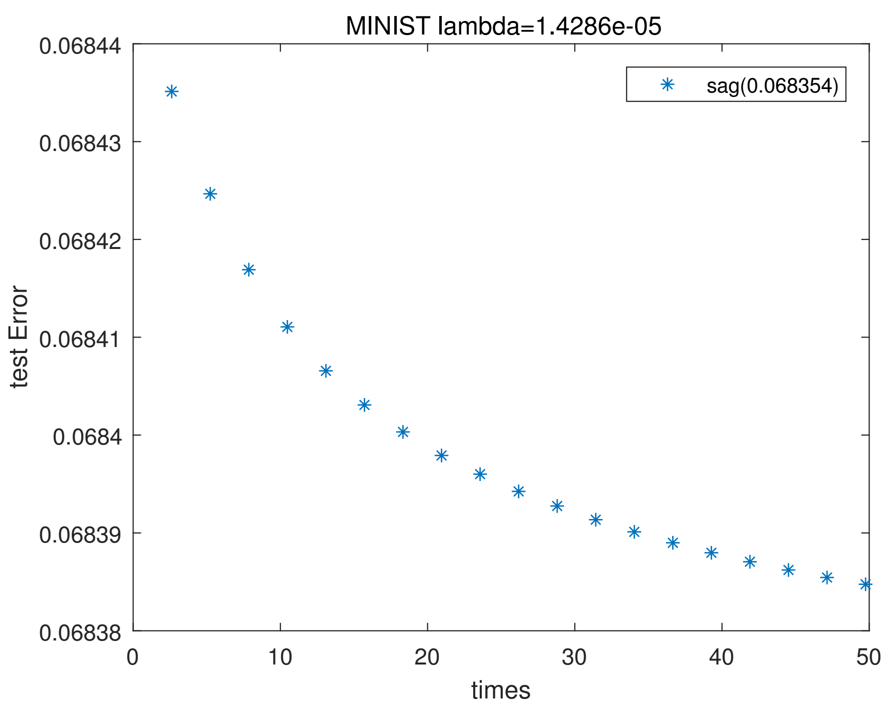
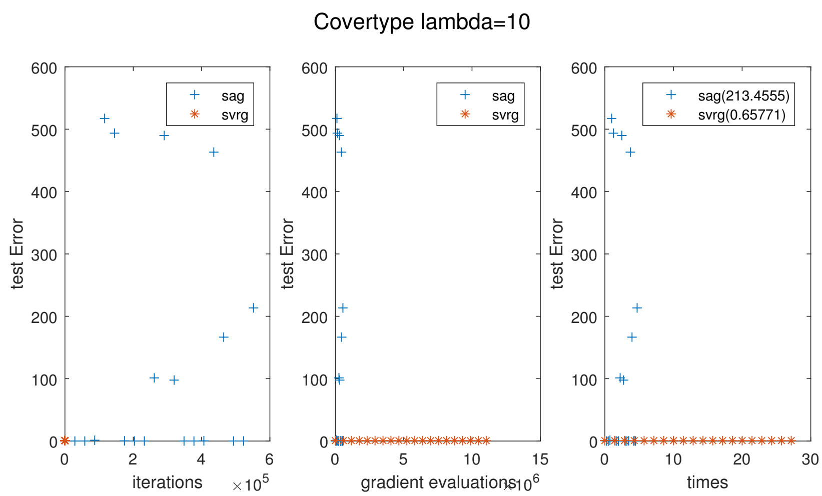
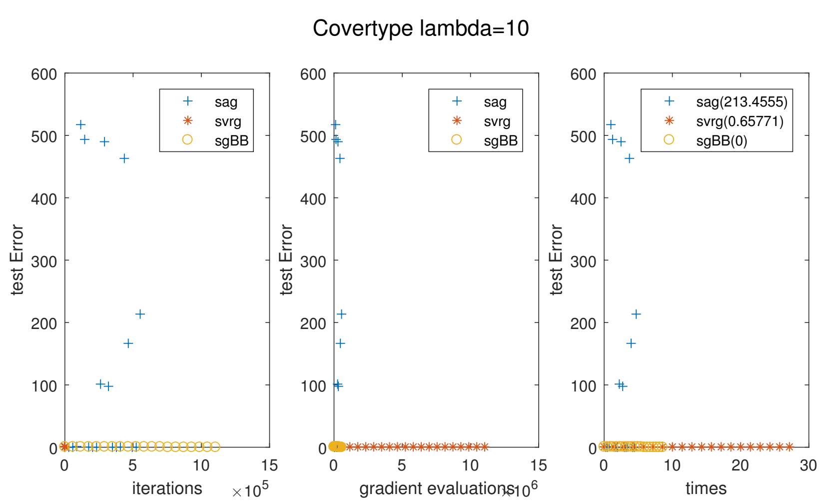
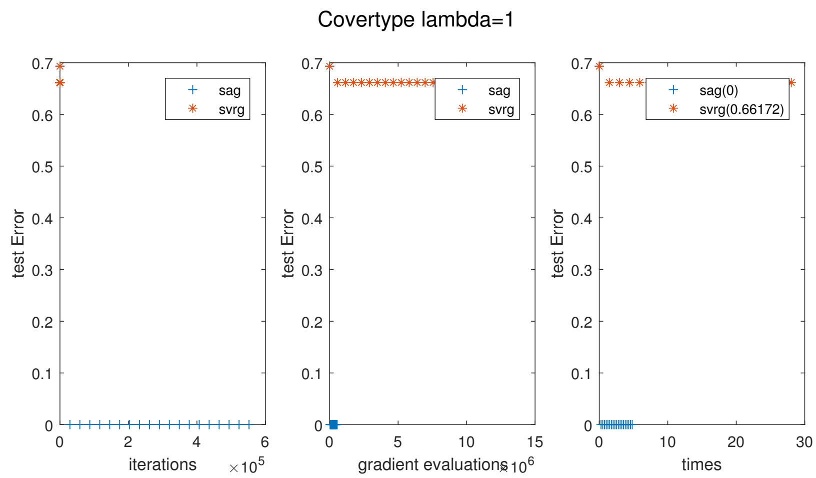
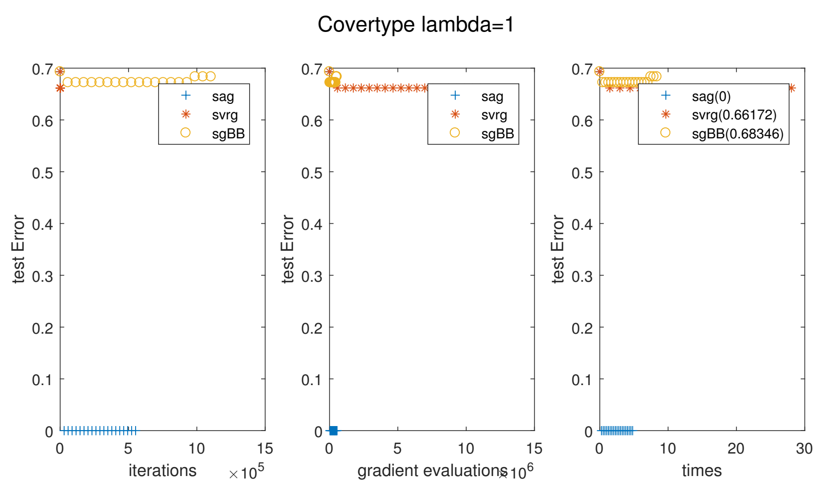

MINIST lambda=1.4286e-05
sag(0.068354)
test Error
0.06844
0.06843
0.06842
0.06841
0.0684
0.06839
0.06838
0 10 20 30 40 50
times

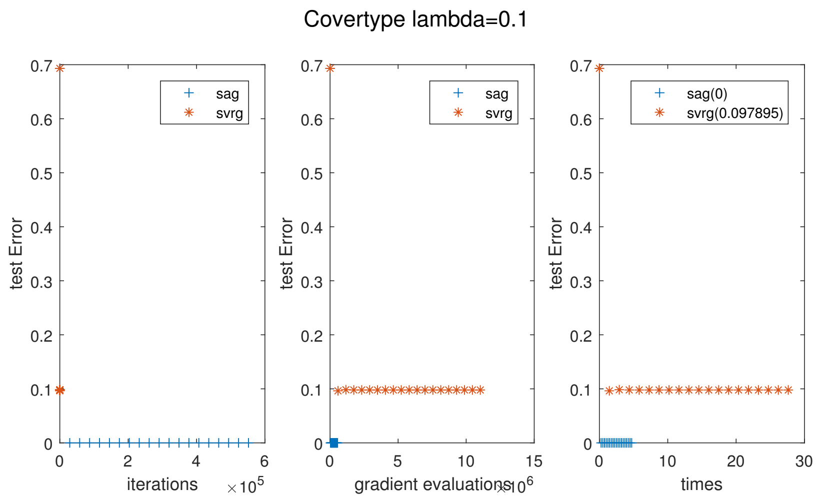
图 2:

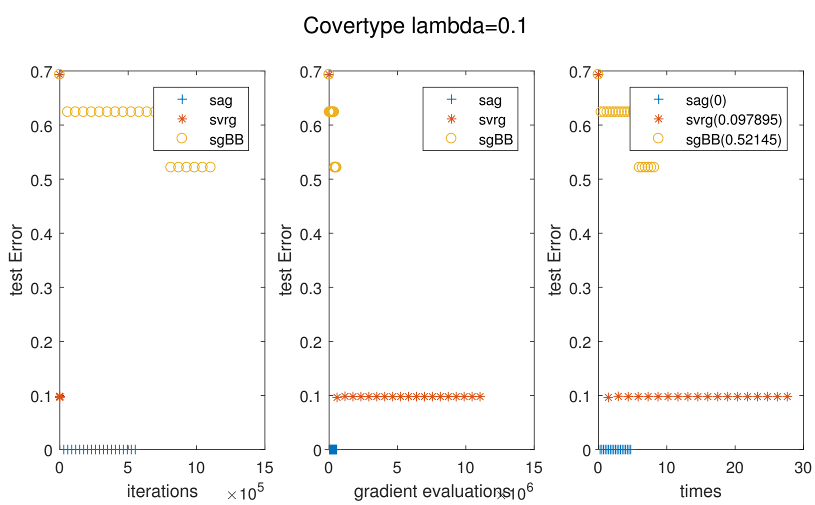

Covertype lambda=10
test Error
iterations
×10^5
sag
svrg
gradient evaluations
×10^6
times
sag(213.4555)
svrg(0.65771)

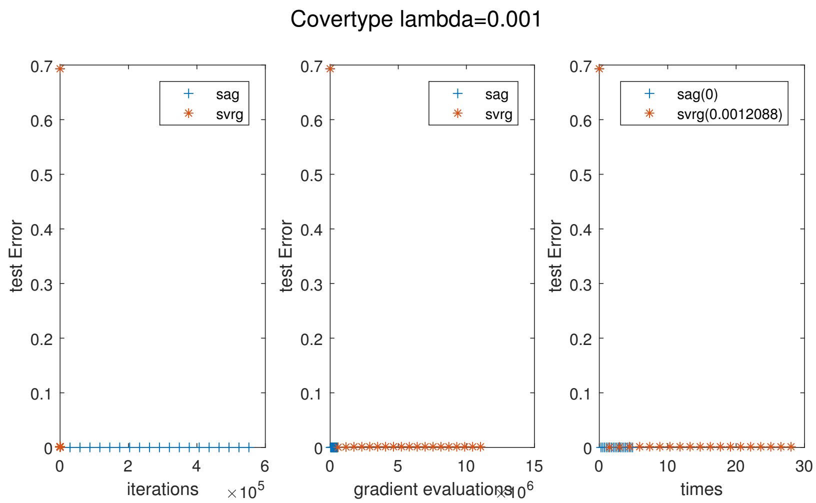

Covertype lambda=10
test Error
iterations
×10^5
sag
svrg
sgBB
gradient evaluations
×10^6
times
sag(213.4555)
svrg(0.65771)
sgBB(0)

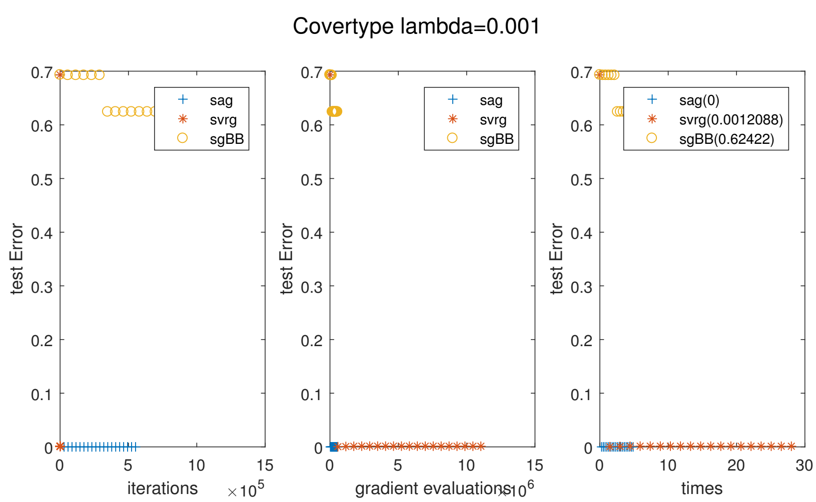

Covertype lambda=1
test Error
iterations
×10^5
gradient evaluations
×10^6
times
sag
svrg
sag(0)
svrg(0.66172)

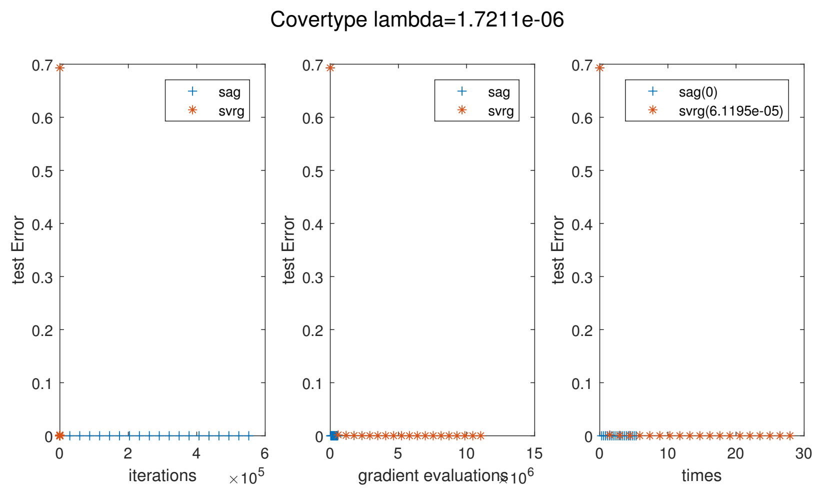

Covertype lambda=1
test Error
iterations
×10^5
gradient evaluations
×10^6
times
sag
svrg
sgBB
sag(0)
svrg(0.66172)
sgBB(0.68346)

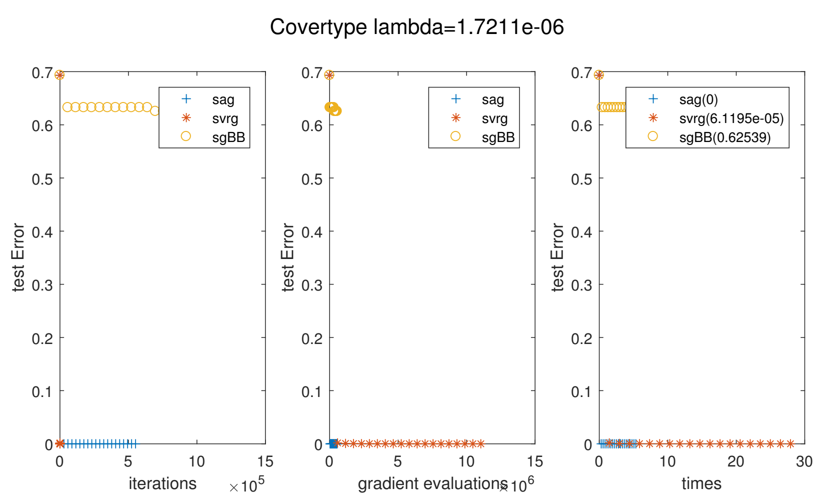

Covertype lambda=0.1
test Error
iterations
×10^5
gradient evaluations
×10^6
times
sag
svrg
sag(0)
svrg(0.097895)


Covertype lambda=0.1
test Error
iterations
×10^5
gradient evaluations
×10^6
times
sag
svrg
sgBB
sag(0)
svrg(0.097895)
sgBB(0.52145)


Covertype lambda=0.001
test Error
iterations
$\times 10^5$
sag
svrg
gradient evaluations
$\times 10^6$
times
sag(0)
svrg(0.0012088)


Covertype lambda=0.001
test Error
iterations
$\times 10^5$
sag
svrg
sgBB
gradient evaluations
$\times 10^6$
times
sag(0)
svrg(0.0012088)
sgBB(0.62422)


Covertype lambda=1.7211e-06
test Error
iterations
gradient evaluations
times
sag
svrg
sag(0)
svrg(6.1195e-05)


Covertype lambda=1.7211e-06
test Error
iterations
gradient evaluations
times
sag
svrg
sgBB
sag(0)
svrg(6.1195e-05)
sgBB(0.62539)

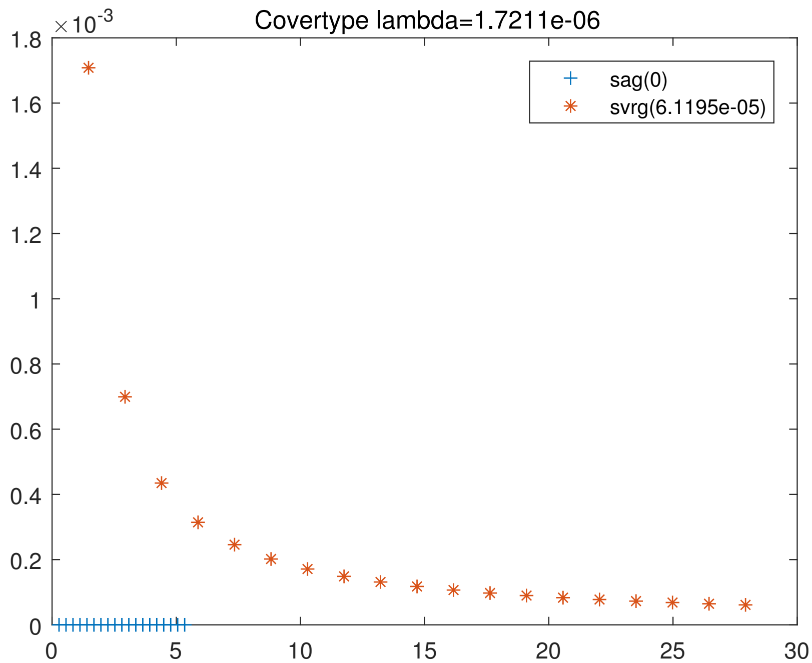

Covertype lambda=1.7211e-06
×10-3
sag(0)
svrg(6.1195e-05)
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
0
5
10
15
20
25
30

图 3:

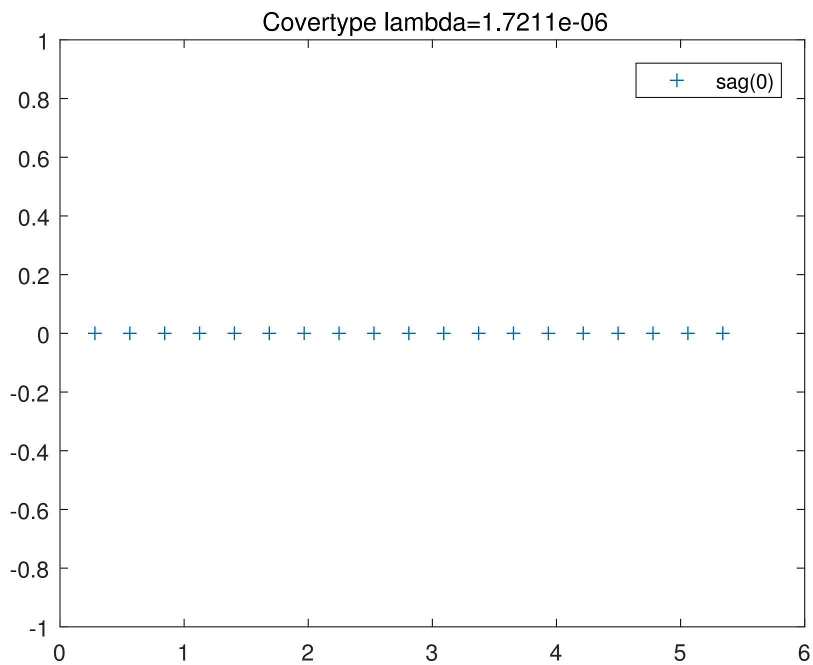

Covertype lambda=1.7211e-06
sag(0)
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
0
1
2
3
4
5
6

图 4:

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
- 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