* 论文部分

摘要：

传统的SGD对于大规模优化问题解决时候具有收敛速度慢的问题，这是由于固有方差影响的。

本文引入，SVRG，在光滑和强凸函数下与SAD和SDCA的收敛速度相同。

但与其不同的是，方法简单直观，而且对于结构预测和NNs比较适用。

结论：介绍了一种随机梯度下降法的显式方差缩减方法，与SDCA和SAD收敛速度相同，但证明更加简单直观。

introduction:

对于优化问题

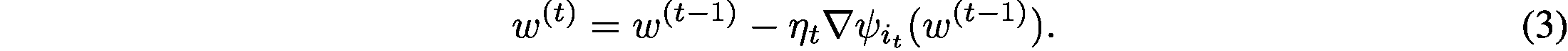
1. 不需要存储完整的梯度
2. 提供了更简单的线性收敛结果
3. 也适用于非凸优化问题。

* 论文中提到的性质：

强凸性：多了二次项，使得凸性更加适用于优化问题

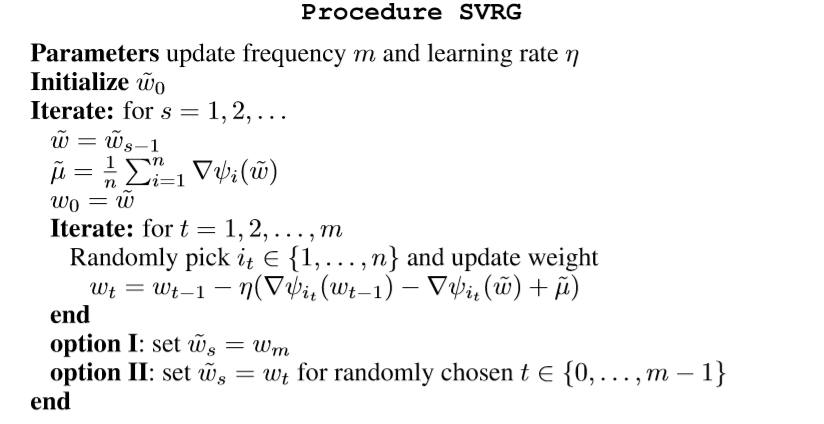
* 论文中提到的方法：

SGD：随机梯度下降：



优势：每次参数迭代只需要计算一个随机的梯度

劣势：每次选择的随意性导致大的方差？（忽大忽小，导致学习率如果为常数，则步长忽大忽小），导致收敛速度变慢

SVRG: 

1. 文献研究的是什么问题？
   1. 文献研究的是在机器学习领域中的优化算法的改进（随机方差梯度下降算法）。
2. 本文用到了什么研究方法？
   1. 用到了类比，对比的研究方法。
3. 文章的结论有哪些？

利用方差预测可以实现加速随机梯度下降的收敛速度。其收敛速度和SAG,SDCA相同。

主要的贡献有三点：1，不需要存储完整的梯度

1. 提供了更简单的线性收敛结果证明
2. 也适用于非凸优化问题。

* 代码复现：

Fig2. Multiclass logistic regression (convex) on MNIST



1，Training loss（Cost） comparison with SGD with fixed learning rates.可以看到随着固定步长的增大，SGD下降得更快，但是随之而来的是收敛的不稳定性。这是由于固定步长太大造成的。对于SVRG:0.05的固定步长来说，收敛得更快也更加稳定。

2，Training loss Residual with SGD0.001,SGD\_best and SVRG,可以看到SVRG的收敛速度很快，而SGD\_best和SGD0.001 收敛得很慢。

3，Variance of W,可以看到w更新的方差是在不断缩减的，尤其时SVRG的速度更加快。

Fig3. More convex-case results.Loss residual P(w)-P(w\*), and test error rates .L2-regularized logistic regression (10-class for CIFAR-10 and binary for the rest)



1, the SGD-best and SVRG training loss on rcv1 convex,其实这张图中的SGD-best,并没有采用步长缩减的策略，如果采用，则不会出现波动，会很光滑。这是与论文中不同的一处。SGD-best收敛很慢，而SVRG收敛速度很快。

2，3，4分别是protein convex ,covertype ,CIFAR-10.的情况，大抵与第一幅图相同。



1. test error rate of rcv1 convex ，可以看到,SGD-best波动性很大，而SVRG的波动性则很小，都达到了选取的样本量的最佳test error rate.

Fig4.全连接神经网络（非凸）上SVRG与SGD在Training loss ，test error rates 的表现





1,Training loss on MNIST data set nonconvex with lambda with lambda=1e-3, 可以看到

2, Training loss on CIFAR-10 data set nonconvex with lambda = 1e-4,可以看到SGD收敛过程中有波动，而SVRG很光滑。

3，Test error rate on CIFAR-10 nonconvex with lambda=1e-4,可以看到收敛过程中，本应是SVRG更加光滑，而SGD-best则有波动，这与凸集情况类似。但是我的实验结果恰恰相反。令我百思不得其解