Comparison of Modern Stochastic Optimization Algorithms

THE UNIVERSITY of EDINBURGH

George Papamakarios
g.papamakarios@sms.ed.ac.uk

Peter Richtárik
peter.richtarik@ed.ac.uk

Problem Description

In machine learning, we are often faced with an optimization problem of the form

$$\min_{\mathbf{w}} f(\mathbf{w}) \equiv \frac{1}{N} \sum_{n=1}^{N} f_n(\mathbf{w})$$

where $\mathbf{w} \in \mathbb{R}^D$ is a classifier, $f_n(\mathbf{w})$ the cost on the n^{th} data point and N may be large.

Algorithms

Gradient Descent

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \alpha \nabla f(\mathbf{w}_k)$$

Stochastic Gradient Descent

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \alpha_k \nabla f_n(\mathbf{w}_k)$$

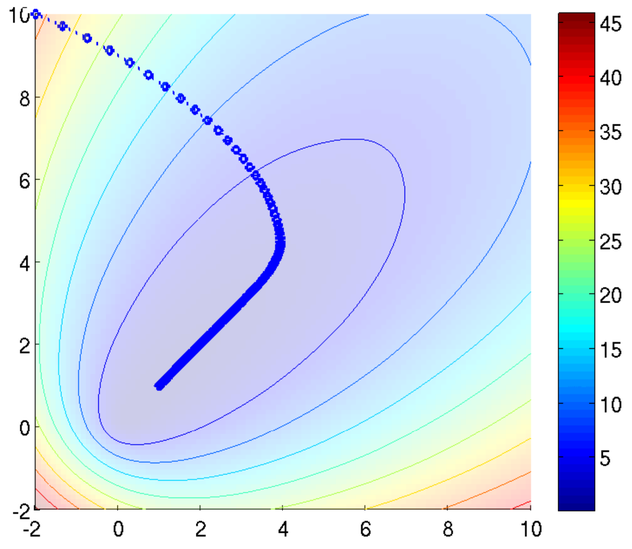
Semi-Stochastic Gradient Descent [1]

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \alpha \Big[\nabla f(\mathbf{y}) - \nabla f_n(\mathbf{y}) + \nabla f_n(\mathbf{w}_k) \Big]$$

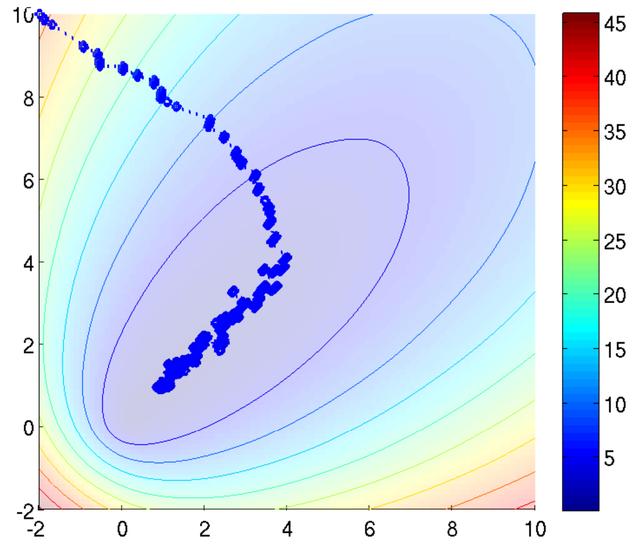
Stochastic Average Gradient [2]

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \frac{\alpha}{N} \left[\sum_{n' \neq n} \nabla f_{n'}(\mathbf{w}_{k'}) + \nabla f_n(\mathbf{w}_k) \right]$$

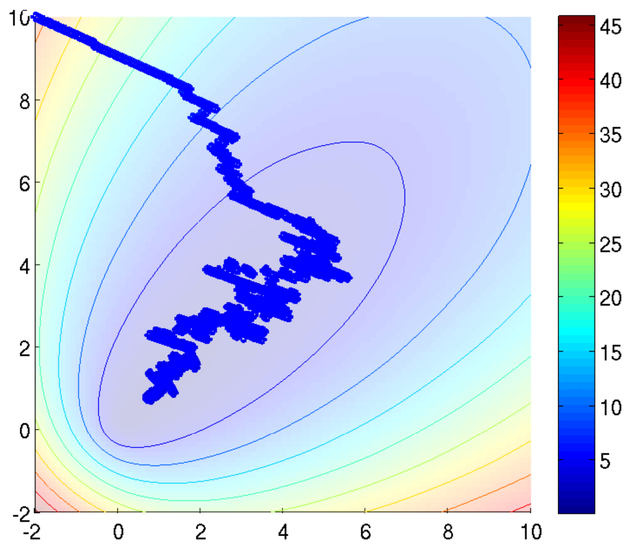
Behaviour of Gradient-Based Optimization



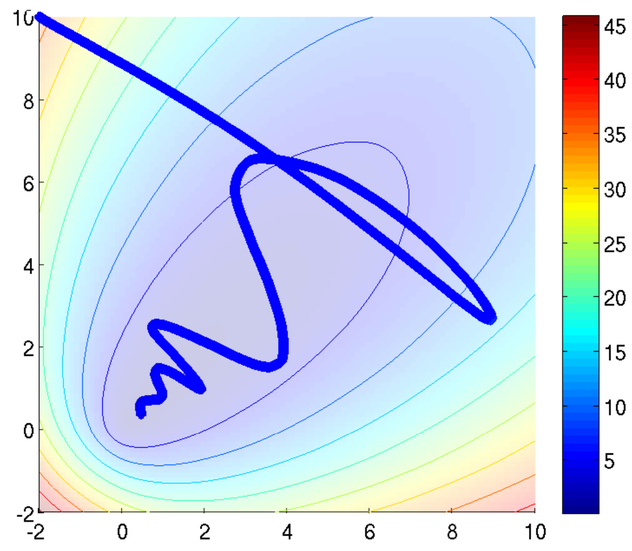
Gradient Descent (GD)



Stochastic Gradient Descent (SGD)

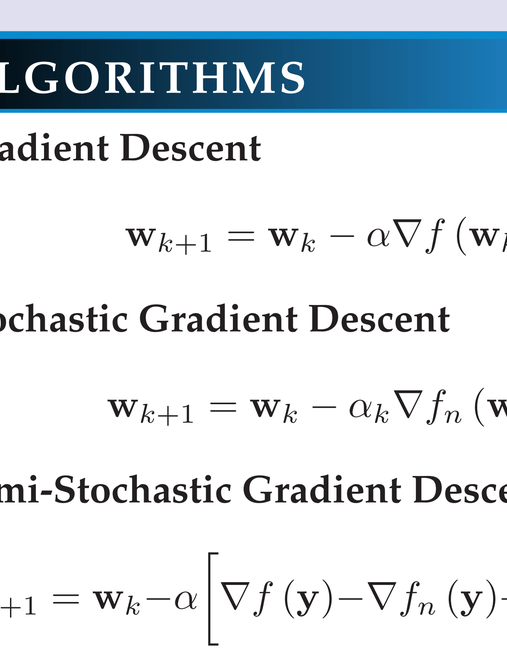


Semi-Stochastic Gradient Descent (S2GD)

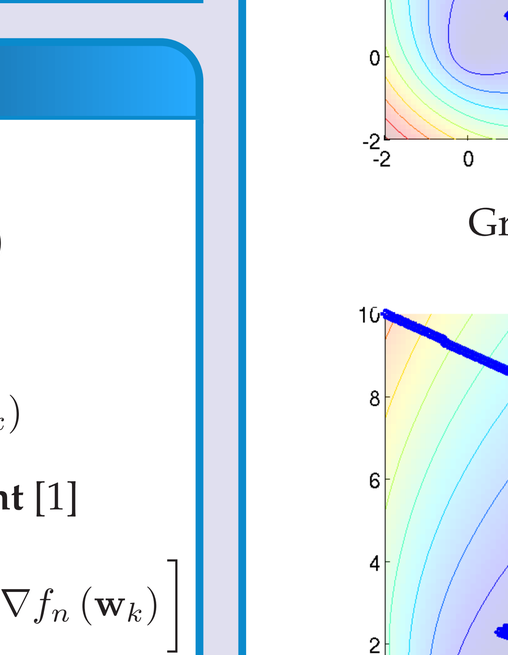


Stochastic Average Gradient (SAG)

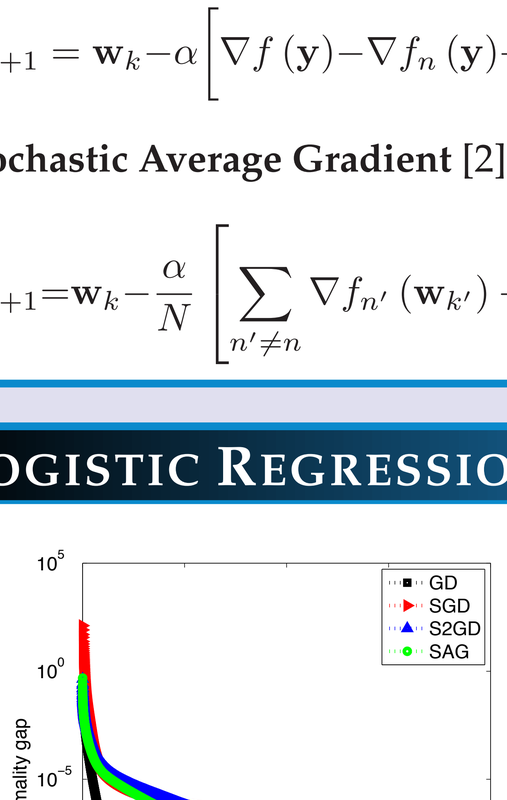
Logistic Regression on Synthetic Data



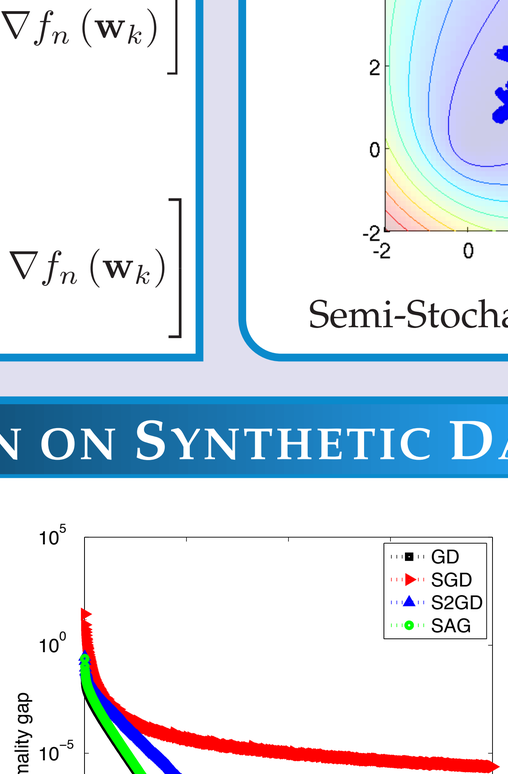
$D = 100$, $N = 10^2$



$D = 100$, $N = 10^3$

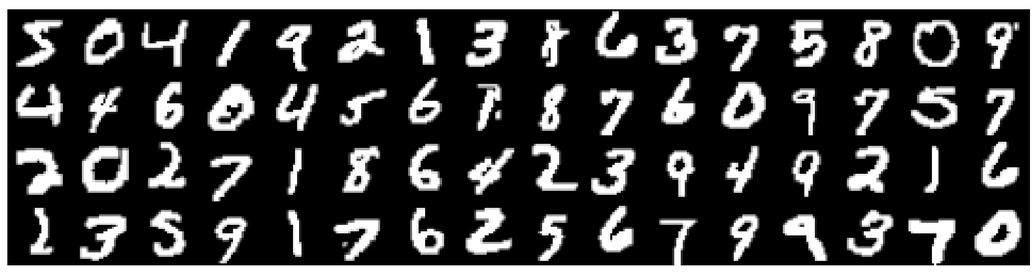


$D = 100$, $N = 10^4$

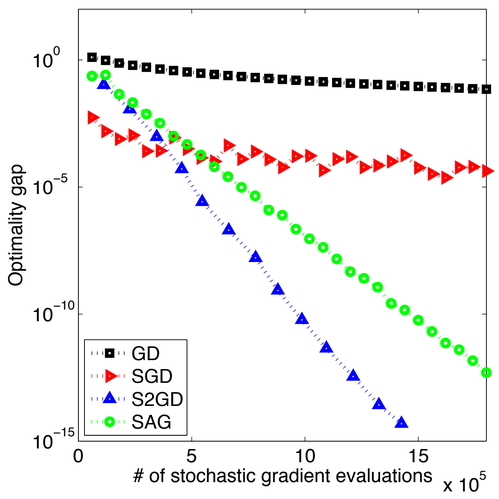


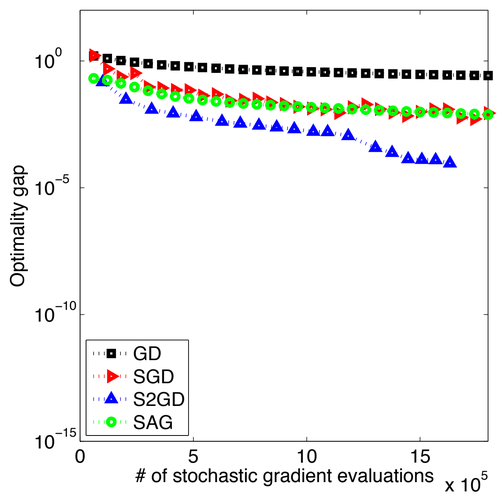
$D = 100$, $N = 10^5$

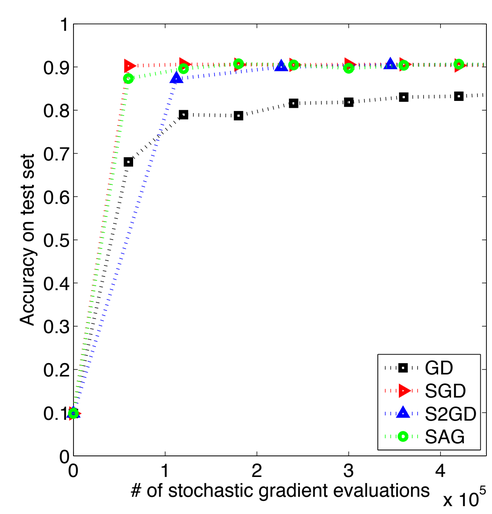
Softmax Regression on MNIST



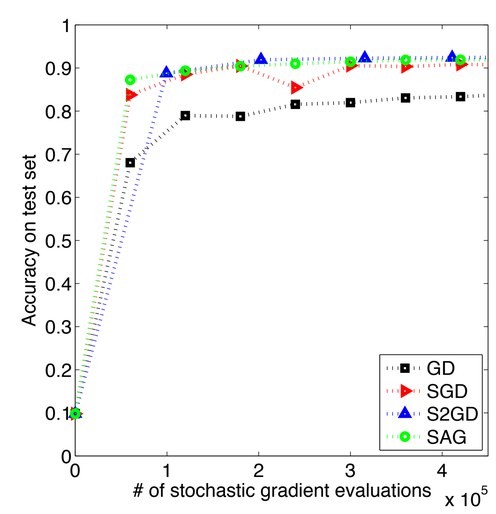
Examples of MNIST handwritten digits
$D = 784$, $N = 60000$, sparsity level $= 80.9\%$







regularizer $\lambda = 10^{-2}$



regularizer $\lambda = 10^{-4}$

References

[1] J. Konečný and P. Richtárik. Semi-Stochastic Gradient Descent Methods. *ArXiv e-prints*, Dec. 2013.

[2] N. L. Roux, M. Schmidt and F. R. Bach. A stochastic gradient method with an exponential convergence rate for finite training sets. In *Advances in Neural Information Processing Systems 25*, pages 2663–2671. 2012.