Accelerating Stochastic Gradient Descent

Kamyar Salahi
Department of Electrical Engineering and Computer Science
UC Berkeley

Abstract

Despite the increased popularity of stochastic gradient methods for large-scale machine learning, until recently, these approaches were largely understudied in the theoretical setting. Although stochastic gradient methods have much lower iteration complexity compared to their deterministic full gradient counterparts, their high variance gradient estimates can hinder convergence. As a consequence, several variance reduction methods have been proposed, often injecting bias into the gradient estimates in order to reduce gradient variance and yield better convergence rates. Unlike deterministic gradient descent, the concept of acceleration or momentum in the stochastic setting has been poorly understood from a theoretical perspective. Recently, it has been shown that stochastic gradient methods can achieve provably optimal convergence by utilizing both variance reduction and an adapted momentum term.

1 Introduction

In the paradigm of large-scale machine learning, calculating the Hessian or higher level derivatives for a given function defined over a high-dimensional domain can be computationally expensive. Although limited memory quasi-newton methods such as L-BFGS Liu and Nocedal (1989) and non-linear conjugate gradient methods such as Fletcher-Reeves Fletcher and Reeves (1964) can eliminate the need to calculate the Hessian, the computational complexity of such methods remains high where computations over a large dataset can be prohibitively slow. Nocedal and Wright (2006)

While some work has shown that on smaller datasets conjugate gradient methods and quasi-newton methods can yield better convergence rates, for larger datasets (such as ImageNet with over 14 million images) runtime complexity remains an issue Le et al. (2011). Overall, the time complexity dependence on the dataset size n fordeterministic methods has caused the machine learning community to utilize stochastic gradient methods wherein a single or few example(s) are utilized estimate the gradient. Although in deterministic gradient descent an optimal convergence rate of $\frac{1}{t^2}$ can be achieved by Nesterov Acceleration E. (1983), the theory for the convergence and acceleration of stochastic gradient methods is considerably less well understood.

In this review, we will explore the limitations of stochastic gradient methods and the proposed solutions. We will begin by introducing the formulation of stochastic gradient descent and its convergence guarantees. We will then proceed to describe how gradient estimate variance can negatively impact method convergence and introduce approaches that address this issue to yield better convergence guarantees. Finally, we discuss the use of momentum for stochastic gradient methods.

2 Stochastic Gradient Descent

Consider the following convex optimization problem

$$\min_{x\in\mathbb{R}^d} F(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x). \tag{1}$$

We define each f_i as being a smooth, convex function.
In stochastic gradient descent Bottou (2010), the update rule is defined as

$$x_{k+1} \leftarrow x_k - \eta g_k$$

where η is the step size or learning rate and g_k is an estimator for the gradient.

Typically, an index i is sampled uniformly at random from $[1, n]$ and $g_k = \nabla f_i(x_k)$. In particular, the function F can represent the average loss over a set of data points with $\nabla f_i(x_k)$ representing the gradient of the loss at the i-th data point. In this scheme, our estimator is unbiased, $E[g_k] = \nabla f(x_k)$.

We will now consider several convergence results for stochastic gradient descent.

Theorem 1. *Let f_i be smooth, convex, and Lipschitz defined over the domain $\mathcal{X}$ such that* $\sup_{x \in \mathcal{X}} E[\|g(x)\|^2] \leq \sigma^2$.

1. *Let $\bar{x} = \frac{1}{T+1} \sum_{k=0}^{T} x_k$ with a fixed $\eta > 0$. Then after T steps, we obtain a bound on the expected suboptimality*

$$E[f(\bar{x})] - f^* \leq \frac{\|x_0 - x_*\|^2}{2\eta(T+1)} + \frac{\eta\sigma^2}{2} \tag{2}$$

2. *Let $\bar{x} = \frac{1}{T+1} \sum_{k=0}^{T} x_k$ with a decreasing $\eta = \frac{\|x_0 - x^*\|}{\sigma\sqrt{T+1}}$. Then after T steps, we obtain a bound on the expected suboptimality*

$$E[f(\bar{x})] - f^* \leq \frac{\|x_0 - x_*\| \, \sigma}{\sqrt{T+1}} \tag{3}$$

From the above theorem, it appears that the convergence rate for a fixed learning rate can be split into two parts. In the first part, we see linear convergence to 0. In the second part, we see a term containing the variance of the gradient independent from the number of iterates. This term will not converge to 0. This means that SGD has linear convergence up to some tolerance after which additional iterates make no further progress.

$$\begin{aligned}
\|x_{k+1} - x_*\| &= \|x_k - x_* - \eta g_k\|^2 && (4)\\
&= \|x_k - x_*\|^2 - \langle x_k - x_*, g_k \rangle + \eta^2 \|g_k\|^2 && (5)\\
E[\|x_{k+1} - x_*\|] &= E[\|x_k - x_*\|^2] - \langle x_k - x_*, \nabla f(x_k) \rangle + \eta^2 E[\|g_k\|^2] && (6)
\end{aligned}$$

With a fixed learning rate, the expected distance between the iterate and the optimizer remains lower bounded by the variance at the given iterate. Since this variance component does not converge in the standard stochastic gradient descent, the method may converge away from the optimizer. For this reason, decreasing step sizes are generally required in order to reduce this variance component over iterates. With a linearly decreasing step size, convergence is no longer hindered by the variance term, but the rate of convergence is now $\frac{1}{\sqrt{t}}$ instead of the $\frac{1}{t}$ that one would obtain in standard gradient descent. Nonetheless, under the assumption of strong convexity and Lipschitz gradients, it is possible to show that one can attain a convergence rate of $\frac{1}{t}$ Nemirovski et al. (2009) which is still worse than the linear convergence rate of deterministic gradient descent in the same case.

Under the assumption that the method only has access to an unbiased measurement of the objective function and its gradient, it turns out that these convergence rates are optimal Blair (1985); Nemirovski et al. (2009). Nonetheless, additional assumptions, such as the fact that functions are sampled from a finite dataset, allow us to show that faster convergence rates of stochastic gradient methods may be possible while preserving iteration complexity. This motivates the need for methods that are able to inherently reduce the estimator variance to enable more rapid convergence without linearly decreasing step sizes.

In other words, we would like to have a stochastic gradient method in which

$$\lim_{x_k \longrightarrow x_*} E[\|g_k\|^2] = 0.$$

3 Variance Reduction

Rather than reducing our learning rate linearly with iterates, it would be more convenient to have the variance of our gradients decay to 0 over iterations. To motivate the approaches that are utilized to reduce variance, we will consider two random variables X and Z such that $\mathrm{Cov}(X, Z) > 0$. We define

$$\hat{X} = X - Z + E[Z].$$

Notice that by linearity of expectation

$$E[\hat{X}] = E[X]$$

and

$$\mathrm{Var}(\hat{X}) = \mathrm{Var}(X) + \mathrm{Var}(Z) - 2\,\mathrm{Cov}(X, Z).$$

Therefore, we can use a Z that is highly correlated with our gradient estimator to reduce the variance of stochastic gradient descent and enable larger step sizes.

3.1 Stochastic Average Gradient Method

The Stochastic Average Gradient Method Schmidt et al. (2013) was one of the first variance reduction stochastic gradient method proposed, extending incremented aggregated gradients Blatt et al. (2007) to the stochastic case. One way of implementing SAG is to maintain a table of i entries each of which contains the last computed gradient of f_i. We define the i-th entry at the k-th iteration as y_i^k. At each iteration, an index i_k is selected uniformly at random from $[1, n]$. We define

$$y_i^k = \begin{cases} \nabla f_i(x^k) & \text{if } i = i_k \\ y_i^{k-1} & \text{otherwise} \end{cases}$$

The next iterate is therefore defined as

$$x_{k+1} \leftarrow x_k - \frac{\eta}{n} \sum_{i=1}^{n} y_i^k.$$

Like the deterministic full gradient descent, this method utilizes gradients calculated with respect to all functions. However, it only computes one gradient in any given iteration meaning that it continues to preserve the iteration complexity of stochastic gradient descent.

In practice, a complete table of all gradients is unnecessary and only a few variables need to be stored.

Algorithm 1 Basic SAG with step size η

```
$d \leftarrow 0$
$y_i \leftarrow 0$ for $i \in [1, n]$
for k=0,1,... do
    Sample $i$ from $\{1, 2, ..., n\}$
    $d \leftarrow d - y_i + \nabla f_i(x)$
    $y_i \leftarrow \nabla f_i(x)$
    $x \leftarrow x - \frac{\eta}{n} d$
end for
```

It is important to recognize that SAG's gradient estimates are no longer unbiased. We can rewrite the iterate step as

$$x_{k+1} \leftarrow x_k - \frac{\eta}{n}(\nabla f_i(x_k) - (y_i^{k-1} - \sum_{i=1}^{n} y_i^{k-1})).$$

Let us define

$$X = \nabla f_i(x_k)$$

and

$$Z = y_i^{k-1} - \sum_{i=1}^{n} y_i^{k-1}.$$

It is evident that $E[X] = f_i(x_k)$ and $E[Z] \neq 0$. This means that $E[\hat{X}] \neq E[X]$ for $\hat{X} = X - Z$. As a consequence, our SAG estimates for the gradient are no longer unbiased.

However, since X and Y are correlated, the variance of our gradient estimate is greatly reduced. In particular, note that as $x_k \to x_*$,

$$\nabla f_i(x_k) - y_i^{k-1} \to 0$$

and

$$\sum_{i=1}^{n} y_i^{k-1} \to \nabla f(x_*) = 0.$$

It turns out that this variance reduction trick enables the use of constant step sizes and improves the convergence rate from $\frac{1}{\sqrt{t}}$ to $\frac{1}{t}$ in the general case and achieves linear convergence in the case of f_i strongly convex.

Assume that the gradients of f_i are Lipschitz continuous with constant L. We define $\bar{x}_k = \frac{1}{k}\sum_{i=1}^{k} x_i$.

Theorem 2. *With a constant step size of $\eta_k = \frac{1}{16L}$, SAG iterations for $k \geq 1$ satisfy:*

$$E[f(\bar{x}_k)] - f(x_*) \leq \frac{32n}{k} C_0$$

where if $y_i^0 = 0$,

$$C_0 = f(x_0) - f(x_*) + \frac{4L}{n} \|x_0 - x_*\|^2 + \frac{\sigma^2}{16L},$$

and if $y_i^0 = \nabla f_i(x_0) - \nabla f(x_0)$,

$$C_0 = \frac{3}{2}(f(x_0) - f(x_*)) + \frac{4L}{n} \|x_0 - x_*\|^2 .$$

If f is μ-strongly convex, we have that

$$E[f(\bar{x}_k)] - f(x_*) \leq \left(1 - \min\left\{\frac{\mu}{16L}, \frac{1}{8n}\right\}\right)^k C_0.$$

Theorem 2 shows that in later iterations, SAG is able to obtain a much faster convergence rate than SGD methods. However, notice that due to the dependence on n, convergence for small k can be slower than that of SGD methods. In practice, one can initialize x_0 using n iterations of SGD and set $y_i^0 = \nabla f_i(x_0) - \nabla f(x_0)$ to attain convergence rates of $\frac{\sqrt{n}}{k}$ and $\frac{\rho^k}{n}$ in the convex and strongly-convex cases respectively.

3.2 Stochastic Variance Reduced Gradient Method

Instead of utilizing a sum of gradients calculated at various different points, the Stochastic Variance Reduced Gradient Method Johnson and Zhang (2013) calculates the full gradient at a snapshot point and the stochastic gradient at both the current iterate and the snapshot. In doing so, unlike SAG, the SVRG is able to achieve variance reduction while retaining an unbiased gradient estimator. In practice, there is some trade off with bias and variance wherein SVRG's gradient estimates can retain higher variance than SAG since they are unbiased.

In particular, consider the gradient estimator

$$g_k = \nabla f_i(x_k) - \nabla f_i(\tilde{x}_k) + \nabla f(\tilde{x}_k)$$

where i is sampled uniformly at random.

It is evident that this estimator is unbiased

$$E[\nabla f_i(x_k) - \nabla f_i(\tilde{x}_k) + \nabla f(\tilde{x}_k)] = \nabla f(x_k).$$

However, for x_k close to $\tilde{x}_k$, assuming f_i have Lipschitz-continuous gradients, $\nabla f_i(x_k)$ and $\nabla f_i(\tilde{x}_k)$ should be correlated. If we define $X = \nabla f_i(x_k)$ and $Z = \nabla f_i(\tilde{x}_k)$, we can see that $\hat{X} = X - Z +$

$E[Z]$ will have reduced variance following our previous discussion. In total, this means that SVRGs gradient estimates are unbiased gradient estimates with reduced variance.

In particular, we note that as $\tilde{x} \to x_*$,

$$\nabla f(\tilde{x}) \to 0.$$

This means that if $\nabla f_i(\tilde{x}_k) \to \nabla f_i(x_*)$

$$\nabla f_i(x_k) - \nabla f_i(\tilde{x}_k) + \nabla f(\tilde{x}_k) \to \nabla f_i(x_k) - \nabla f_i(x_*) \to 0.$$

Algorithm 2 SVRG with update frequency m and step size η

```
Intialize $\tilde{x}_0$
for s=0,1,... do
    $\tilde{x} \leftarrow \tilde{x}_{s-1}$
    $\tilde{\mu} \leftarrow \nabla f(\tilde{x})$
    $x_0 = \tilde{x}$
    for t=0,1,...,m do
        Sample $i$ from $\{1, 2, ..., n\}$
        $x_t \leftarrow x_{t-1} - \eta(\nabla f_i(x_{t-1}) - \nabla f_i(\tilde{x}) + \tilde{\mu})$
    end for
    $\tilde{x}_s \leftarrow x_m$
end for
```

Similar to SAG, this variance reduction trick allows for the use of constant step sizes and can improve the convergence rate from $\frac{1}{\sqrt{t}}$ to $\frac{1}{t}$.

Theorem 3. *Consider the SVRG algorithm. Assume that f_i are convex and smooth and f is strongly convex. Assume that m is large enough such that*

$$\alpha = \frac{1}{\gamma\eta(1-2L\eta)m} + \frac{2L\eta}{1-2L\eta} < 1$$

then we have geometric convergence of SVRG

$$E[f(\tilde{x}_s)] - f(x_*) \leq \alpha^s (f(\tilde{x}_0) - f(x_*))$$

Much like in SAG, we see geometric convergence for strongly convex functions, yielding considerably faster convergence rates than SGD methods.

4 Acceleration

Recently, some papers have attempted to demonstrate that approaches analogous to Nesterov's momentum trick can be applied for stochastic gradient descent Frostig et al. (2015); Lin et al. (2015); Shalev-Shwartz and Zhang (2014). However, many of these approaches either yield suboptimal convergence rates or have practicality concerns with regard to storage or hyperparameter tuning. As we will discuss in the next section, one such method, Katyusha Momentum Allen-Zhu (2017) is able to tractably achieve optimal convergence.

4.1 Katyusha Momentum

Katyusha momentum presented a means of accelerating stochastic gradient descent with low storage overhead to reach provably optimal convergence rates. The authors present two separate optimization algorithms with similar underlying principles for the strongly-convex and general cases.

We modify the optimization problem from before such that

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x) + \phi(x)$$

In our discussion of the approach, we will refer to $\kappa = \frac{L}{\sigma}$ where f_i are L-Lipschitz and ϕ is σ-strongly convex.

For the strongly convex case, we will see that $(n+\sqrt{n\kappa})\log(\frac{1}{\epsilon})$ SGD iterations are necessary for ϵ convergence. For the general case, we will see that $n\log(\frac{1}{\epsilon})+\sqrt{\frac{nL}{\epsilon}}$ SGD iterations are necessary for ϵ convergence.

Much like SVRG, Katyusha momentum utilizes $\tilde{x}$ snapshots that are updated every m iterations. $\tilde{\nabla}_{k+1}$ is the gradient estimator defined in the same way as in SVRG. τ_1 and τ_2 represent two separate momentum parameters. We define $\alpha = \frac{1}{3\tau_1 L}$.

Algorithm 3 Katyusha for σ-Strongly Convex ϕ and L-Lipschitz Functions

```
$m \leftarrow 2n$
$\tau_2 \leftarrow \frac{1}{2},\ \tau_1 \leftarrow \min\{\frac{\sqrt{m\sigma}\sqrt{3L}}{,}\frac{1}{2}\}, \alpha \leftarrow \frac{1}{3\tau_1 L}$
$y_0 = z_0 = \tilde{x}_0 \leftarrow x_0$
for $s = 0, ..., S-1$ do
    $\mu^s \leftarrow \nabla f(\tilde{x}_s)$
    for $j = 0, ..., m-1$ do
        $k \leftarrow (sm) + j$
        $x_{k+1} \leftarrow \tau_1 z_k + \tau_2 \tilde{x}_s + (1-\tau_1-\tau_2)y_k$
        Sample $i$ from $\{1, 2, ..., n\}$
        $\tilde{\nabla}_{k+1} \leftarrow \mu_s + \nabla f_i(x_{k+1}) - \nabla f_i(\tilde{x}_s)$
        $z_{k+1} = \arg\min_z(\frac{1}{2\alpha}\|z-z_k\|^2 + \langle\tilde{\nabla}_{k+1}, z\rangle + \phi(z))$
        $y_{k+1} = \arg\min_z(\frac{3L}{2}\|y-x_{k+1}\|^2 + \langle\tilde{\nabla}_{k+1}, y\rangle + \phi(y))$
    end for
    $\tilde{x}_{s+1} \leftarrow (\sum_{j=1}^{m-1}(1+\alpha\sigma)^j)^{-1} \cdot (\sum_{j=0}^{m-1}(1+\alpha\sigma)^j \cdot y_{sm+j+1})$
end for
```

As seen above, Katyusha momentum defines the next iterate as a convex combination of three vectors. The z_k vector can be seen as a weighted sum of the previous gradients. In traditional Nesterov acceleration, the next iterate can be defined as a convex combination of the vectors y_k and z_k where y_k represents the gradient step and z_k represents the momentum term. At an intuitive level, the addition of the third term ensures that x_k does not move too far away from the snapshot $\tilde{x}_k$. This contributes in two ways. Firstly, the SVRG gradient estimator will more effectively reduce variance when the gradient is more correlated with the gradient of the snapshot. As the x_k move further away from $\tilde{x}_k$, the gradients at the snapshot may not be as correlated with the gradients at x_k. Secondly, it can be seen as a negative momentum term to counteract some momentum from early iterations that are no longer contributing beneficially to convergence.

One can note that when $\tau_2 = 0$, the classical Nesterov Acceleration Method is obtained. For $\tau_1, \tau_2 = 0$, Katyusha momentum reduces to SVRG. The authors found that both in theory $\tau_1 = \min\{\sqrt{\frac{n\sigma}{L}}, 0.5\}$ and $\tau_2 = 0.5$ are optimal and in practice these parameters work well.

Theorem 4. *If each of $f_i(x)$ is convex, L-smooth and $\phi(x)$ is strongly convex then Katyusha satisfies*

$$E[f(\tilde{x}_s)] - f(x_*) \leq \begin{cases} O\left((1+\sqrt{\frac{\sigma}{3Lm}})^{-Sm}\right) \cdot (f(x_0) - f(x_*)), & \text{if } \frac{m\sigma}{L} \leq \frac{3}{4} \\ O\left((1.5)^{-S}\right) \cdot (f(x_0) - f(x_*)), & \text{if } \frac{m\sigma}{L} > \frac{3}{4}. \end{cases}$$

This means that with $m = O(n)$, Katyusha is able to obtain ϵ error in $O((n+\sqrt{\frac{nL}{\sigma}}) \cdot \log(\frac{f(x_0)-f(x_*)}{\epsilon}))$ iterations.

The primary modification to Katyusha momentum in the general case is the dependence of $\tau_{1,s}$ on the index s. A decaying τ_1 and α are also seen in the accelerated gradient methods in the non-strongly convex case.

Theorem 5. *If each of $f_i(x)$ is convex, L-smooth and $\phi(x)$ is strongly convex then our general Katyusha algorithm satisfies*

$$E[f(\tilde{x}_s)] - f(x_*) \leq O\left(\frac{f(x_0)-f(x_*)}{S^2} + \frac{L\|x_0 - x_*\|^2}{mS^2}\right)$$

Algorithm 4 Katyusha for General Case

```
$m \leftarrow 2n$
$\tau_2 \leftarrow \frac{1}{2}$
$y_0 = z_0 = \tilde{x}_0 \leftarrow x_0$
for $s = 0, ..., S - 1$ do
    $\tau_{1,s} \leftarrow \frac{2}{s+4}, \alpha_s \leftarrow \frac{1}{3\tau_{1,s}L}$
    $\mu^s \leftarrow \nabla f(\tilde{x}_s)$
    for $j = 0, ..., m - 1$ do
        $k \leftarrow (sm) + j$
        $x_{k+1} \leftarrow \tau_{1,s} z_k + \tau_2 \tilde{x}_s + (1 - \tau_{1,s} - \tau_2) y_k$
        Sample $i$ from $\{1, 2, ..., n\}$
        $\tilde{\nabla}_{k+1} \leftarrow \mu_s + \nabla f_i(x_{k+1}) - \nabla f_i(\tilde{x}_s)$
        $z_{k+1} = \arg\min_z(\frac{1}{2\alpha_s} \|z - z_k\|^2 + \langle \tilde{\nabla}_{k+1}, z \rangle + \phi(z))$
        $y_{k+1} = \arg\min_z(\frac{3L}{2} \|y - x_{k+1}\|^2 + \langle \tilde{\nabla}_{k+1}, y \rangle + \phi(y))$
    end for
    $\tilde{x}_{s+1} \leftarrow \frac{1}{m} \sum_{j=1}^{m} y_{sm+j}$
end for
```

This means that with $m = O(n)$, Katyusha is able to obtain ϵ error in $O\left(\frac{\sqrt[n]{f(x_0) - f(x_*)}}{\epsilon} + \frac{\sqrt{nL}\|x_0 - x_*\|}{\sqrt{\epsilon}}\right)$ iterations.

References

Allen-Zhu, Z. (2017). Katyusha: The first direct acceleration of stochastic gradient methods. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2017, page 1200–1205, New York, NY, USA. Association for Computing Machinery.

Blair, C. (1985). Problem complexity and method efficiency in optimization (a. s. nemirovsky and d. b. yudin). *SIAM Review*, 27(2):264–265.

Blatt, D., Hero, A. O., and Gauchman, H. (2007). A convergent incremental gradient method with a constant step size. *SIAM Journal on Optimization*, 18(1):29–51.

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In Lechevallier, Y. and Saporta, G., editors, *Proceedings of COMPSTAT'2010*, pages 177–186, Heidelberg. Physica-Verlag HD.

E., N. Y. (1983). A method for solving the convex programming problem with convergence rate o(1/k^2)$. Dokl.Akad.NaukSSSR, 269 : 543 - -547$.

Fletcher, R. and Reeves, C. M. (1964). Function minimization by conjugate gradients. *The Computer Journal*, 7(2):149–154.

Frostig, R., Ge, R., Kakade, S. M., and Sidford, A. (2015). Un-regularizing: Approximate proximal point and faster stochastic algorithms for empirical risk minimization. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 2540–2548. JMLR.org.

Johnson, R. and Zhang, T. (2013). Accelerating stochastic gradient descent using predictive variance reduction. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Le, Q. V., Ngiam, J., Coates, A., Lahiri, A., Prochnow, B., and Ng, A. Y. (2011). On optimization methods for deep learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, page 265–272, Madison, WI, USA. Omnipress.

Lin, H., Mairal, J., and Harchaoui, Z. (2015). A universal catalyst for first-order optimization. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Liu, D. C. and Nocedal, J. (1989). On the limited memory bfgs method for large scale optimization. *MATHEMATICAL PROGRAMMING*, 45:503–528.

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. (2009). Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609.

Nocedal, J. and Wright, S. J. (2006). *Numerical Optimization*. Springer, New York, NY, USA, 2e edition.

Schmidt, M., Roux, N. L., and Bach, F. (2013). Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*.

Shalev-Shwartz, S. and Zhang, T. (2014). Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. In Xing, E. P. and Jebara, T., editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 64–72, Bejing, China. PMLR.