B6.2 Optimisation for Data Science

Lecture Notes, Lectures 9–16 (Part I)
Oxford Mathematical Institute, HT 2022

Prof. Coralia Cartis

February 15, 2022

7 Stochastic gradient methods

7.1 Introduction

Many optimization problems arising in data science applications can be written as an (unconstrained) sum of functions, namely,

$$\min_{x\in\mathbb{R}^n} f(x) = \frac{1}{m}\sum_{j=1}^{m} f_j(x), \tag{43}$$

where each $f_j : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable for $j \in \{1, \ldots, m\}$.

In particular, recall the general formulation of the Data fitting problem in Section 1.4 (Lectures 1–8), where each f_j measures the error/misfit/loss at a given data point; for example, the (squared) error between predicted model output at a data point and its observed value. Many of the examples given at the start of these lectures fall in this framework, provided the component functions are smooth[5]: Example 2 (Lectures 1–8) of regression problems (apart from Example 2 iv)-Lasso); smooth formulations of the matrix completion problem; Example 7 (Lectures 1–8) of multi-class classification and training deep learning models.

It is typical in these applications that both n - the number of variables or unknowns and m - the number of data points are very large. This implies that calculating even one entire gradient ∇f of f is extremely computationally expensive or even impossible. The issue then arises on how to scale up first-order/gradient descent methods to the huge scale and requirements of these modern applications, in a robust and rigorous way?

We will investigate this question in the remaining lectures. We will first address the large number m of data points or observations and how to 'reduce' this algorithmically – this will lead to stochastic first-order algorithms. Then we will look at reducing the size n of the variables – this will lead to coordinate (or more generally, subspace) variants of gradient descent. Both these classes of algorithms can be viewed as approximate/inexact gradient descent variants, and so we build on the knowledge you gained in Section 4.

7.2 Stochastic gradient methods

The celebrated Stochastic Gradient Descent (SGD) method, proposed by Robbins and Monro in 1951 for problem (43), avoids the calculation of the full gradient of f in (43) by calculating only the gradient(s) of one or a a small(er) number of component functions f_j, chosen uniformly at random from the terms $(f_1, \ldots, f_m)$. It replaces the negative gradient direction in the method of steepest descent with the negative (averaged) gradient of this subset of functions.

[5] A nonsmooth regularization term may be allowed, but we avoid it for now for simplicity.

At iteration $k \geq 0$, given the (current) iterate x^k, the SGD algorithm constructs an update x^{k+1} to x^k as follows,

$$x^{k+1} = x^k + \alpha_k g^k, \tag{44}$$

where $\alpha_k > 0$ is as in earlier lectures, the step-length (can be constant, or varying, and typically pre-defined at the start of the algorithm); and g^k is

$$g^k = \nabla f_{\mathcal{S}_k}(x^k) = \frac{1}{m_k} \sum_{j \in \mathcal{S}_k} \nabla f_j(x^k), \tag{45}$$

where $\mathcal{S}_k \subset \{1, 2, \ldots, m\}$ chosen uniformly at random, and where the cardinality of $\mathcal{S}_k$ is m_k (so $|\mathcal{S}_k| = m_k$), which we also refer to as *batch size*.

Remark: We can see that if $m_k = m$ then (44) coincides with the steepest descent iteration (with general stepsize) when applied to (43). If $m_k = 1$, then only one term f_j (only one data point) from the sum of functions in (43) is chosen uniformly at random and its gradient is used to construct the next iterate. □

A summary of (a realization) of SGD is given next.

Stochastic gradient descent (SGD)

Algorithm 2 (SGD). *Given $x^0 \in \mathbb{R}^n$ (deterministic), for $k = 0, 1, 2, \ldots$ repeat:*

sample $\mathcal{S}_k$ i.i.d. $\sim U(\{1, \ldots, m\})$ → (induces randomness)

calculate $g^k = \nabla f_{\mathcal{S}_k}(x^k)$ according to (45)

form $x^{k+1} = x^k - \alpha_k g^k$ → (random vector)

We denote random/stochastic variables by capital letters, and their realisations by usual letters: $x^k \to X^k$, $g^k \to G^k$. Thus SGD algorithm is a stochastic algorithm/process.

While for gradient descent (steepest descent) method we have guaranteed sufficient descent at each iteration, here we have expected descent only. Indeed, (a realization of) $-G^k$ may not be a descent direction: $\nabla f(X^k)^{\mathrm{T}}(-G^k) < 0$ cannot be guaranteed, but is guaranteed in expectation. Therefore, to ensure convergence, we must analyse the expected descent of the random iterates (X^k).

A numerical illustration. Binary classification tasks are similar/part of the 'Multiclass classification' (Example 7), namely, we would like to classify/label data points into two classes, by means of finding a separating hyperplane that 'best' separates two given classes of already labelled data points; the notion of 'best' is quantified by means of minimizing the error between the predicted classification and the true one in a regression sense (either logistic regression or mean squared regression). We then use the 'optimal' model/hyperplane that

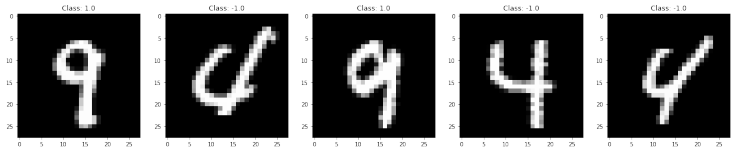

Figure 11: Training data of handwritten digits 4 and 9 from the MNIST data set.

we found to decide/predict which side of it (or which class) unseen/new data points must lie.

Optimization formulation and methods are used in the training phase of finding the optimal model/hyperplane. We aim to find $x \in \mathbb{R}^n$ such that on the training data, (a_i, y_i), $i \in \{1, \ldots, m\}$, where $a_i \in \mathbb{R}^n$ and $y_i \in \{-1, 1\}$ denotes the class to which the data a_i belongs to, we classify correctly on average using squared error loss[6], namely,

$$\min_{x \in \mathbb{R}^n} \sum_{i=1}^{m} (a_i^T x - y_i)^2. \tag{46}$$

We may solve this problem using the SGD algorithm. Here, we focus on constructing a dataset (from the MNIST data set) for the task of distinguishing between images of handwritten digits: 4 and 9, see figure below. Solving the ensuing (46) using the SGD algorithm with $\alpha^k = 0.001$ for all k and $|\mathcal{S}_k| = 1$, we obtain the following results given in Figure **??**. Note the stochastic nature of the objective decrease and its (large) variance and seeming oscillatory stagnation asymptotically.

A more general set up The SGD algorithm can be extended to a more general stochastic framework, namely, when $f(x) := \mathbb{E}_\xi(F(x, \xi)))$. Then the single sampled gradient $\nabla_j f(X^k)$ is replaced by a stochastic estimate $g(X^k, \xi_k) \approx \nabla f(X^k)$; see for example, [5] and Problem Sheet 3 for an illustration. In fact, this makes it easy to run SGD on stochastic objectives and compare it with gradient descent methods in order to understand their respective behaviours; we illustrate this here.

Numerical Illustration. We consider the scaled quadratic function on Problem Sheet 2 (Problem 3): $f(x) = \frac{1}{2}(\kappa x_1^2 + x_2^2)$, where $\kappa = 20$ here. We can transform this into a stochastic objective by adding Gaussian perturbations ('noise') of magnitude 0.02. We then apply SGD (in the more general setting) with $|\mathcal{S}_k| = 1$ when noise is present (which becomes) gradient descent when no noise is present (and then GD is applied to f as is.) The linesearch is adaptive

[6] Logistic loss is also possible/allowed.

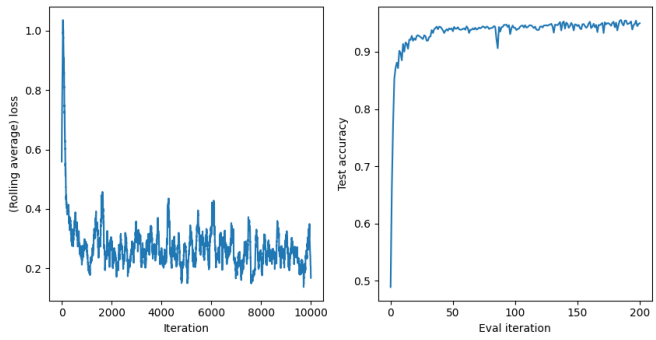

Figure 12: Regression loss across the iterations of SGD method (left) and the test accuracy over evaluations (right) [@Ilan Price, Oxford].

(not fixed or dependent on the Lipschitz constant) but following conditions similar to Wolfe linesearch. The following results are obtained. The numerical illustrations here can be found on Google Colab (Please feel free to copy this and run it and modify it yourself):

https://colab.research.google.com/drive/1wcLFWkiiUABzWU9hhZRgTbPRjoBsx9Gf?usp=sharing.

7.3 Global convergence of the SGD algorithm

Conditions and assumptions needed to show convergence of SGD applied to (43): If $|\mathcal{S}_k| = 1$ (one data element), the expected value of the gradient with respect to the selected data point is an unbiased estimator of the true gradient :

$$\mathbb{E}_{\mathcal{S}_k}[G^k] = \mathbb{E}[G^k|\mathcal{S}_k] = \sum_{j=1}^{m} \mathbb{E}[G^k|\mathcal{S}_k = j]\cdot\mathbb{P}[\mathcal{S}_k = j] = \sum_{j=1}^{m} \nabla f_j(X^k)\cdot\frac{1}{m} = \nabla f(X^k). \tag{47}$$

- Similarly for larger sets $\mathcal{S}_k$ drawn uniformly from $\binom{m}{|\mathcal{S}_k|}$ possible configurations; referred to as mini-batches. See Problem Sheet 3.
- Above, we used $\mathbb{E}[G^k|\mathcal{S}_k = j] = \nabla f_j(X^k)$ (true due to iid choice of $\mathcal{S}_k$ and G^k). More generally, we require an unbiased estimator of the true gradient: $\mathbb{E}_{\mathcal{S}_k}[G^k] = \nabla f(X^k)$.

In addition to the above underlying assumption of our analysis that G^k conditioned on current batch is an unbiased estimator of the true gradient[7], we

[7]Namely, (47) holds. This property is true here (and when $|\mathcal{S}_k| > 1$), but it would have to be assumed/enforced in a more general stochastic framework.

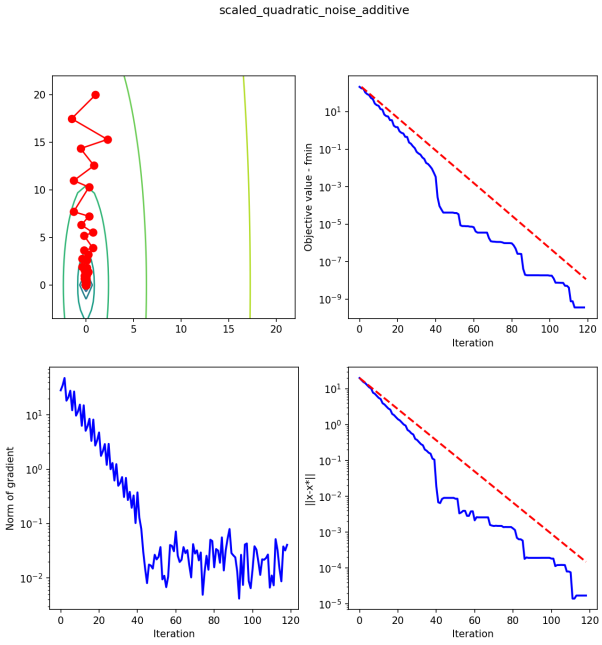

Figure 13: SGD applied to the scaled quadratic objective perturbed by additive noise [@Ilan Price, Oxford].

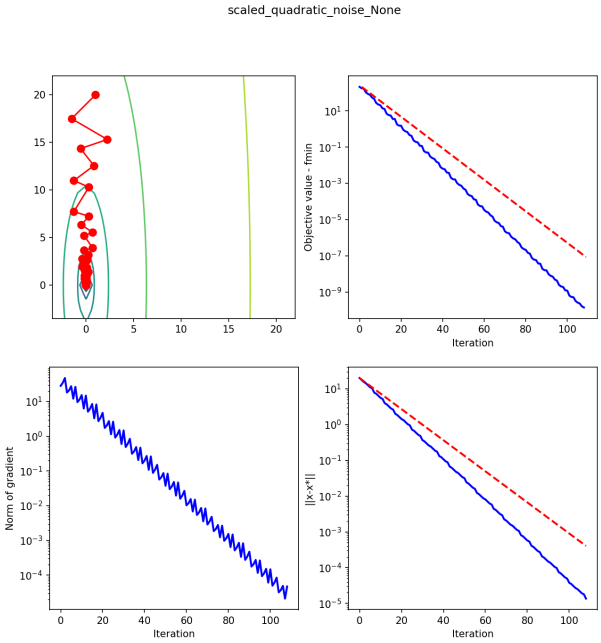

Figure 14: GD applied to the scaled quadratic objective (no noise) [@Ilan Price, Oxford].

require the following ($|\mathcal{S}_k| = 1$):

(1) for all $j \in \{1, \ldots, m\}$, ∇f_j is L-smooth (i.e, ∇f_j is Lipschitz continuous is constant L). Note that this implies that f in (43) is L-smooth (i.e, ∇f is Lipschitz continuous is constant L).[8]

(2) there exists $M > 0$ such that

$$\mathrm{VAR}(G^k|\mathcal{S}_k) := \mathbb{E}[(G^k - \nabla f(X^k))^{\mathrm{T}}(G^k - \nabla f(X^k))|\mathcal{S}_k] \leq M$$

for all k [9].

A useful property - in expectation

Lemma 5. *[An overestimation property - in expectation] Assume Assumption (1) holds. When applying SGD to f with $|\mathcal{S}_k| = 1$, we have*

$$\mathbb{E}_{\mathcal{S}_k}\left[f(X^{k+1})\right] \leq f(X^k) - \alpha \nabla f(X^k)^{\mathrm{T}}\, \mathbb{E}_{\mathcal{S}_k}\left[G^k\right] + \frac{L\alpha^2}{2}\, \mathbb{E}_{\mathcal{S}_k}\left[\|G^k\|^2\right]. \quad (48)$$

If Assumption (2) also holds, then

$$\mathbb{E}_{\mathcal{S}_k}\left[f(X^{k+1})\right] \leq f(X^k) - \alpha^k \left(\frac{L\alpha^k}{2} - 1\right) \|\nabla f(X^k)\|^2 + \frac{ML(\alpha^k)^2}{2}. \quad (49)$$

Proof. Since Assumption (1) implies that f is L-smooth, Proposition 2(iv) (Lectures 1–8)[10] applies to give the following (deterministic) 'overestimation' property,

$$f(x + \alpha d) \leq f(x) + \alpha \nabla f(x)^T d + \frac{1}{2}\alpha^2 L \|d\|^2, \quad (50)$$

for all x, $d \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$. Letting $x = X^k$, $d = G^k$ and $\alpha = \alpha^k$ in (50) and using $X^{k+1} = X^k + \alpha^k G^k$, we deduce

$$f(X^{k+1}) \leq f(X^k) - \alpha^k \nabla f(X^k)^T G^k + \frac{L}{2}(\alpha^k)^2 \|G^k\|^2.$$

Applying expectation on both sides with respect to $\mathcal{S}_k$,

$$\mathbb{E}_{\mathcal{S}_k}[f(X^{k+1}] \leq f(X^k) - \alpha^k \nabla f(X^k)^T\, \mathbb{E}_{\mathcal{S}_k}[G^k] + \frac{L}{2}(\alpha^k)^2\, \mathbb{E}_{\mathcal{S}_k}[\|G^k\|^2],$$

where we used that $f(X^k)$ and $\nabla f(X^k)$ do not depend on $\mathcal{S}_k$. Thus (48) follows.

[8] The latter assumption is common to GD methods when f is deterministic.

[9] Bounded total variance can usually be guaranteed in a neighbourhood of x^* but not globally for strongly convex f.

[10] Or equation (11) (Lectures 1–8)

We already showed[11] that $\mathbb{E}_{\mathcal{S}_k}[G^k] = \nabla f(X^k)$. Thus

$$\begin{aligned} \mathrm{VAR}(G^k|\mathcal{S}_k) &= \mathbb{E}_{\mathcal{S}_k}\left[\|G^k\|^2\right] - 2\nabla f(X^k)^{\mathrm{T}}\,\mathbb{E}_{\mathcal{S}_k}[G^k] + \|\nabla f(X^k)\|^2 \\ &= \mathbb{E}_{\mathcal{S}_k}\left[\|G^k\|^2\right] - \|\nabla f(X^k)\|^2. \end{aligned}$$

which together with Assumption (2), gives $\mathbb{E}_{\mathcal{S}_k}\left[\|G^k\|^2\right] \leq M + \|\nabla f(X^k)\|^2$. This and (48) gives (49). □

Global convergence of SGD: the general case

Theorem 9. *[SGD with fixed stepsize: general case] Let f in (43) be bounded below by f_{low}, satisfying Assumptions (1) and (2). Apply the SGD method with $|\mathcal{S}|_k = 1$ and fixed stepsize $\alpha = \eta/L$, where $\eta \in (0, 1]$ to (43). Then, for $k \geq 1$,*

$$\min_{0\leq i\leq k-1} \mathbb{E}[\|\nabla f(X^i)\|^2] \leq \alpha LM + \frac{2(f(x^0) - f_{\mathrm{low}})}{k\alpha} = \eta M + \frac{2L(f(x^0) - f_{\mathrm{low}})}{k\eta}.$$

and so the SGD method takes at most $k \leq \frac{2L(f(x^0)-f_{\mathrm{low}})}{\eta\epsilon}$ iterations/evaluations to generate $\mathbb{E}[\|\nabla f(X^{k-1})\|^2] \leq \epsilon + \eta M$.

- Theorem 9 implies that $\min_{0\leq i\leq k} \mathbb{E}[\|\nabla f(X^i)\|^2] - \eta M \leq \mathcal{O}\left(\frac{1}{k}\right)$ and so $\liminf_{k\to\infty} \mathbb{E}[\|\nabla f(X^k)\|^2] \leq \eta M$. With more work, and an additional smoothness assumption on $\|\nabla f\|^2$, one can show that $\lim_{k\to\infty} \mathbb{E}[\|\nabla f(X^k)\|^2] \leq \eta M$, not just liminf. Thus (some form of) global convergence of SGD is obtained up to level ηM - 'noise level' or 'noise floor'! This limits the accuracy to which convergence can be achieved.
- Compare this result to the gradient descent method and its general convergence rate; see Theorem 1 (Lectures 1–8) (both have sublinear rate of convergence).

Proof. (Theorem 9) Lemma 5 and $\frac{L\alpha}{2} - 1 = \frac{\eta}{2} - 1 < -\frac{1}{2}$ give

$$\mathbb{E}_{\mathcal{S}_k}\left[f(X^{k+1})\right] \leq f(X^k) - \frac{\alpha}{2}\|\nabla f(X^k)\|^2 + \frac{ML\alpha^2}{2}.$$

Taking expectation $\mathbb{E}$ with respect to the past, namely, $\mathcal{S}_0, \ldots, \mathcal{S}_{k-1}$ on both sides of the above, we note that we have a memoryless property so current iterate only depends on previous sample size,

$$\mathbb{E} = \mathbb{E}_k := \mathbb{E}(\cdot|\mathcal{S}_0, \ldots, \mathcal{S}_k) = \mathbb{E}_{\mathcal{S}_k},$$

and so,

$$\mathbb{E}_k\left[f(X^{k+1})\right] \leq \mathbb{E}_{k-1}\left[f(X^k)\right] - \frac{\alpha}{2}\,\mathbb{E}_{k-1}\left[\|\nabla f(X^k)\|^2\right] + \frac{ML\alpha^2}{2}.$$

[11] Or may need to assume in general stochastic cases

(Recall the techniques in the proof of convergence for GD method, we need to connect the per iteration decrease with the gradient.) We have for all $k \geq 0$, that the expected decrease satisfies

$$\mathbb{E}_{k-1}\left[f(X^k)\right] - \mathbb{E}_k\left[f(X^{k+1})\right] \geq \frac{\alpha}{2}\mathbb{E}_{k-1}\left[\|\nabla f(X^k)\|^2\right] - \frac{ML\alpha^2}{2}. \tag{51}$$

Summing up (51) from $i = 0$ to k, and using $f(X^{k+1}) \geq f_{\text{low}}$ we deduce

$$\begin{aligned} f(x^0) - f_{\text{low}} &\geq f(x^0) - \mathbb{E}_k\left[f(X^{k+1})\right] \\ &\geq \tfrac{\alpha}{2}\textstyle\sum_{i=0}^{k}\mathbb{E}_{k-1}\left[\|\nabla f(X^i)\|^2\right] - (k+1)\frac{ML\alpha^2}{2}. \\ &\geq \tfrac{\alpha}{2}(k+1)\left[\min_{0\leq i\leq k}\mathbb{E}[\|\nabla f(X^i)\|^2] - ML\alpha\right] \end{aligned}$$

□

We are able to modify the SGD algorithm to ensure improved convergence, beyond the 'noise floor'. More on this in a moment. Until then, we show that the behaviour of the SGD algorithm improves compared to the sublinear rate in Theorem 9.

Global convergence of SGD: the strongly convex case Let f be γ-strongly convex conform Defition 2(v) (Lectures 1–8).

Theorem 10. *[SGD with fixed stepsize: strongly convex case] Let f be γ-strongly convex and satisfying Assumptions (1) and (2). Let SGD with fixed stepsize be applied to minimize f, where $\alpha^k = \alpha = \frac{\eta}{L}$ and $\eta \in (0, 1]$. Then SGD converges linearly to a residual error in the following sense: for all $k \geq 0$,*

$$\mathbb{E}[f(X^k)] - f(x^*) - \frac{\eta M}{2\gamma} \leq \left(1 - \frac{\eta\gamma}{L}\right)^k \cdot \left[f(x^0) - f(x^*) - \frac{\eta M}{2\gamma}\right]. \tag{52}$$

Remarks:

- Theorem 10 implies that $\lim_{k\to\infty}(\mathbb{E}[f(X^k)] - f(x^*)) \leq \frac{\alpha ML}{2\gamma} = \frac{\eta M}{2\gamma}$. Thus global convergence of SGD is obtained, in expectation, up to the level $\frac{\eta M}{2\gamma}$ (noise level !). This level can be decreased in various ways, as we discuss shortly.
- The ratio $\frac{L}{\gamma}$ in (52) is a condition number of f (connect to second derivatives). Recall rate of GD!

Proof. (Theorem 10)[12] Lemma 5 and $\frac{L\alpha}{2} - 1 = \frac{\eta}{2} - 1 < -\frac{1}{2}$ give

$$\mathbb{E}_{\mathcal{S}_k}\left[f(X^{k+1})\right] \leq f(X^k) - \frac{\alpha}{2}\|\nabla f(X^k)\|^2 + \frac{ML\alpha^2}{2}.$$

[12] The beginning of this proof is similar to that of Theorem 9.

Taking expectation $\mathbb{E}$ with respect to the past, namely, $\mathcal{S}_0, \ldots, \mathcal{S}_{k-1}$ on both sides of the above, we note that we have a memoryless property so current iterate only depends on previous sample size,

$$\mathbb{E} = \mathbb{E}_k := \mathbb{E}(\cdot|\mathcal{S}_0, \ldots, \mathcal{S}_k) = \mathbb{E}_{\mathcal{S}_k},$$

and so,

$$\mathbb{E}_k\left[f(X^{k+1})\right] - f(x^*) \leq \mathbb{E}_{k-1}\left[f(X^k)\right] - f(x^*) - \frac{\alpha}{2}\,\mathbb{E}_{k-1}\left[\|\nabla f(X^k)\|^2\right] + \frac{ML\alpha^2}{2}.$$

A consequence of the strong convexity property is that the global minimizer x^* is unique and $f(X^k) - f(x^*) \leq \frac{1}{2\gamma}\|\nabla f(X^k)\|^2$ (conform the proof of Theorem 3, page 16 of Lectures 1–8); thus $2\gamma\,\mathbb{E}_{k-1}(f[(X^k] - f(x^*)) \leq \mathbb{E}_{k-1}(\|\nabla f(X^k)\|^2)$. We deduce

$$\mathbb{E}_k\left[f(X^{k+1})\right] - f(x^*) \leq (1-\gamma\alpha)\left(\mathbb{E}_{k-1}\left[f(X^k)\right] - f(x^*)\right) + \frac{ML\alpha^2}{2}, \tag{53}$$

or equivalently,

$$\mathbb{E}_k\left[f(X^{k+1})\right] - f(x^*) - \frac{\alpha ML}{2\gamma} \leq (1-\gamma\alpha)\left(\mathbb{E}_{k-1}\left[f(X^k)\right] - f(x^*) - \frac{\alpha ML}{2\gamma}\right). \tag{54}$$

Note that $\alpha = \eta/L \leq 1/L \leq 1/\gamma$. Replacing α gives

$$\mathbb{E}_k\left[f(X^{k+1})\right] - f(x^*) - \frac{M\eta}{2\gamma} \leq \left(1 - \frac{\eta\gamma}{L}\right)\left(\mathbb{E}_{k-1}\left[f(X^k)\right] - f(x^*) - \frac{M\eta}{2\gamma}\right),$$

The claim now follows by induction. □

Decreasing the SGD "noise floor": technique 1 Though not always desirable (due to the need for small 'generalization error'/prediction on unseen data set), the SGD "floor" (noise level) of $\frac{\eta M}{2\gamma}$ can be removed so that $\lim_{k\to\infty}\mathbb{E}[f(X^k)] = f(x^*)$.

Technique 1: Dynamically reduce $\alpha^k = \frac{\eta_k}{L}$ Note that $\eta_k \to 0$ makes the residual $\frac{\eta_k M}{2\gamma} \to 0$ but it also means that $\left(1 - \frac{\eta_k}{L}\right) \to 1$, so the price is that we lose linear convergence!

Theorem 11. *[Dynamic stepsize stochastic gradient descent (DS-SGD)] Let f be γ-strongly convex and satisfying Assumptions (1) and (2). In the SGD algorithm, let $\alpha^k = \dfrac{2}{2L + k\gamma}$, for all $k \geq 0$. Then SGD satisfies, for all $k \geq 0$,*

$$0 \leq \mathbb{E}[f(X^k)] - f(x^*) \leq \frac{\nu}{2\frac{L}{\gamma} + k} \tag{55}$$

where $\nu := 2\frac{L}{\gamma} \times \max\left\{\frac{M}{\gamma}, f(x^0) - f(x^*)\right\}$.
Thus $\lim_{k\to\infty}\mathbb{E}[f(X^k)] = f(x^*)$. But rate is $\mathcal{O}\left(\frac{1}{k}\right)$ - sublinear !

Proof. (Theorem 11) (Similar to the proof of Theorem 10) Note that all arguments in the proof of Theorem 10 continue to hold with α replaced by α^k, until and including (54). Thus, using $\alpha^k \leq 1/L \leq 1/\gamma$, for all $k \geq 0$, we have

$$\mathbb{E}_k\left[f(X^{k+1})\right] - f(x^*) - \frac{\alpha^k ML}{2\gamma} \leq \left(1-\gamma\alpha^k\right)\left(\mathbb{E}_{k-1}\left[f(X^k)\right] - f(x^*) - \frac{\alpha^k ML}{2\gamma}\right).$$

We are now going to prove the desired conclusion (55) by induction. Clearly at $k=0$, (55) holds. Assume (55) holds at $k>0$, and substitute (55) into the above displayed equation. We obtain

$$\mathbb{E}_k\left[f(X^{k+1})\right] - f(x^*) - \frac{\alpha^k ML}{2\gamma} \leq \left(1-\gamma\alpha^k\right)\left(\frac{\nu}{2\frac{L}{\gamma}+k} - \frac{\alpha^k ML}{2\gamma}\right).$$

Using the expression of α^k in the above and simplifying the expressions provides (55) with k replaced by $(k+1)$.

Decreasing the SGD "noise floor": technique 2 In this approach, increase mini-batch sizes from $|\mathcal{S}_k| = 1$ to $|\mathcal{S}_k| = p \geq 1$.

In SGD, use $G^k = \frac{1}{p}\sum_{j\in\mathcal{S}_k} \nabla f_j(X^k)$, where $j \in \mathcal{S}_k$ i.i.d. $\sim \mathcal{U}(\{1,\ldots,m\})$. Then we have,

$$\begin{aligned}
\mathrm{VAR}(G^k|\mathcal{S}_k) &= \sum_{j\in\mathcal{S}_k}\frac{1}{p^2}\mathbb{E}_{\mathcal{S}_k}\left[\|\nabla f_j(X^k) - \nabla f(X^k)\|^2\right] \\
&\quad +2\sum_{j<i}\frac{1}{p^2}\mathbb{E}_{\mathcal{S}_k}\left[\nabla f_j(X^k) - \nabla f(X^k)\right]^{\mathrm{T}}\mathbb{E}_{\mathcal{S}_k}\left[\nabla f_i(X^k) - \nabla f(X^k)\right] \\
&= \frac{1}{p^2}\sum_{j\in\mathcal{S}_k}\mathrm{VAR}(\nabla f_j(X^k)) + 0 \leq \frac{M}{p},
\end{aligned}$$

where we have used $|\mathcal{S}_k| = p$ and the independence of i and j indices in $\mathcal{S}_k$ in the first equality as well as the lack of bias $\mathbb{E}_{\mathcal{S}_k}\left[\nabla f_j(X^k)\right] = \nabla f(X^k)$. We also have $\mathbb{E}_{\mathcal{S}_k}\left[G^k\right] = \nabla f(X^k)$ - unbiased batch gradient.

Then, as in Theorem 10, we deduce, under the same assumptions,

$$\mathbb{E}[f(X^k)] - f(x^*) - \frac{\eta M}{2\gamma p} \leq \left(1-\frac{\eta\gamma}{L}\right)^k \cdot \left[f(x^0) - f(x^*) - \frac{\eta M}{2\gamma p}\right]. \tag{56}$$

Thus the noise level is decreased by batch size p, without impacting the convergence factor. The noise floor is not necessarily removed.
(Compare and contrast Techniques 1 and 2.)

Decreasing the SGD "noise floor": technique 3 This approach adds acceleration to SGD.

Technique 3: Acceleration for gradient variance reduction is used to reduce $\mathrm{VAR}(G^k|\mathcal{S}_k)$. This yields $E[f(X^k)] \to f(x^*)$ with linear convergence rate, with a much smaller cost per iteration than mini-batching (see the 'Katyusha' paper). Other techniques (earlier than Katyusha): variance reduction (SVRG), SAG (Schmidt, Le Roux, Bach'15: restores linear rate for SGD), SAGA (Defazio et al'14). Some of these techniques discussed later and some in Part C (Theories of Deep Learning).

Conclusions:

- Each of the three approaches for improving SGD have merit and are often all used at once. In particular, once SGD appears to stagnate one both reduces the stepsize and increases the batch-size; though this is stopped once validation error begins to increase.
- For the general case, when f may not be (known to be) strongly convex, we can improve on Theorem 9 by using Technique 1 and Technique 2 above (Technique 3 using acceleration is difficult in general, it needs convexity). Regarding decreasing stepsize (Technique 1), let $\alpha^k = \eta_k/L$ where $\eta_k \in (0, 1]$. Similarly to the proof of Theorem 9, we obtain

$$\sum_{i=0}^{k} \alpha^i \, \mathbb{E}_{i-1} \left[\|\nabla f(X^i)\|^2 \right] \leq 2(f(x^0) - f_{\text{low}}) + ML \sum_{i=0}^{k} (\alpha^i)^2.$$

  And so to reduce the noise term, assume that $\sum_{i=0}^{\infty} \alpha^i = \infty$ and $\sum_{i=0}^{\infty} (\alpha^i)^2 < \infty$ in SGD. Various choices of α^k are thus possible.

Global convergence of SGD: additional comments

- Compare the obtained bounds with GD results: convergence in expectation/with positive probability.
- Rates comparison with GD
- Ill-conditioning present, just like in any first-order method. See Part C Continuous Optimization.

7.4 Stochastic variance reduction methods

Reducing the variance in SGD by increasing batch size (again) Recall Technique 2 above, namely (56): when f is L-smooth and γ-strongly convex, SGD with fixed stepsize converges - in expectation - linearly up to the level $\frac{\eta M}{2\gamma}$ (noise level !), where M related to bound on variance of $\|G^k\|^2$. Instead of Assumption (2), assume now that for all $k \geq 0$,

$$\mathrm{VAR}(G^k|\mathcal{S}_k) := \mathbb{E}_{\mathcal{S}_k}[\|G^k - \nabla f(X^k)\|^2] \leq \frac{M}{\xi^k} \tag{57}$$

where $\xi > 1$.

Theorem 12. *Let f in (43) be γ-strongly convex and satisfying Assumption 1 and (57). Let SGD with fixed stepsize be applied to minimize f, where $\alpha^k = \alpha = \frac{\eta}{L}$ where $\eta \in (0, 1]$. Then SGD satisfies the linear rate:*

$$\mathbb{E}[f(X^k)] - f(x^*) \leq \nu_{vr}\rho^k, \quad \text{for all } k \geq 0, \tag{58}$$

where $\nu_{vr} := \max\left\{\frac{\alpha LM}{\gamma}, f(x^0) - f(x^*)\right\}$ and $\rho := \max\{1 - \alpha\gamma/2, 1/\xi\} < 1$. Thus $\lim_{k\to\infty} \mathbb{E}[f(X^k)] = f(x^*)$ linearly.

Proof. (Theorem 12) We use similar techniques to earlier SGD results (see proof of Theorem 10). Namely, the following holds, which we borrow from the proof of Theorem 10 (see equation (53)) and in which we replace M by M/ξ^k:

$$\mathbb{E}_k\left[f(X^{k+1})\right] - f(x^*) \leq (1 - \gamma\alpha)\left(\mathbb{E}_{k-1}\left[f(X^k)\right] - f(x^*)\right) + \frac{ML\alpha^2}{2\xi^k}, \; k \geq 0. \tag{59}$$

We prove (58) by induction on k. For $k = 0$ clearly (58) holds due to definition of ν_{vr}. Assume (58) holds for k. Replacing (58) into the above decrease we obtain:

$$\begin{aligned}
\mathbb{E}_k\left[f(X^{k+1})\right] - f(x^*) &\leq (1 - \gamma\alpha)\,\nu_{vr}\rho^k + \tfrac{ML\alpha^2}{2\xi^k} \\
&\leq \nu_{vr}\rho^k\left[1 - \gamma\alpha + \tfrac{ML\alpha^2}{2\nu_{vr}}\tfrac{1}{(\rho\xi)^k}\right] \\
&\leq \nu_{vr}\rho^k\left[1 - \gamma\alpha + \tfrac{ML\alpha^2}{2\nu_{vr}}\right] \text{ as } \rho \geq 1/\xi \\
&\leq \nu_{vr}\rho^k\left[1 - \gamma\alpha + \tfrac{\gamma\alpha}{2}\right] \text{ due to def } \nu_{vr} \\
&= \nu_{vr}\rho^k\left[1 - \tfrac{\gamma\alpha}{2}\right] = \nu_{vr}\rho^{k+1}
\end{aligned}$$

□

Reducing the variance in SGD by increasing batch size: to achieve (58), let $|\mathcal{S}_k| = p > 1 \longrightarrow |\mathcal{S}_k| = p^k$ (p to power k). Disadvantage: potentially expensive.

Reducing the variance by gradient aggregation Assume $|\mathcal{S}_k| = 1$ (for simplicity). SVRG (Johnson and Zhang'13):

$$\nabla f(x^k) \approx g^k = \nabla f_{\mathcal{S}_k}(x^k) - \nabla f_{\mathcal{S}_k}(\bar{x}^k) + \nabla f(\bar{x}^k)$$

for some $\bar{x}^k$ close to x^k. Requires calculation of full gradient from time to time ! (Too) Expensive !

Some intuition: Assume we want to estimate $\mathbb{E}(X)$ for some random variable $X (= G^k)$. Let $Z = X - Y + \mathbb{E}(Y)$; then $\mathbb{E}(Z) = \mathbb{E}(X)$ and $\mathrm{VAR}(Z) = var(X) + var(Y) - 2cov(X, Y)$. Thus measuring Z instead of X could give an estimate of mean with lower variance if X and Y are positively correlated.

If x^k and $\bar{x}^k$ are close, likely that $\nabla f_{\mathcal{S}_k}(x^k)$ and $\nabla f_{\mathcal{S}_k}(\bar{x}^k)$ are positively correlated.

Stochastic variance-reduced gradient (SVRG)

Algorithm 3 (SVRG). *Given $x^0 \in \mathbb{R}^n$, for $k = 0, 1, 2, \ldots$ repeat:*

calculate $\nabla f(x^k)$ and let $\bar{x}^{k,0} := x^k$.

for $j \in \{0, \ldots, p-1\}$, do:

- *select $\mathcal{S}_{k,j}$ i.i.d. $\sim U(\{1, \ldots, m\})$ with replacement, $|\mathcal{S}_{k,j}| = 1$.*
- $g^{k,j} := \nabla f_{\mathcal{S}_{k,j}}(\bar{x}^{k,j}) - \nabla f_{\mathcal{S}_{k,j}}(x^k) + \nabla f(x^k)$
- $\bar{x}^{k,j+1} = \bar{x}^{k,j} - \alpha^{k,j} g^{k,j}$

New iterate: $x^{k+1} \in \{\bar{x}^{k,j} : j \in \{0, \ldots, p-1\}\}$, chosen uniformly at random.

The following convergence result holds.

Theorem 13. *Let f_j for each $j \in \{1, \ldots, m\}$ be convex and satisfy Assumption 1. Let f in (43) be γ-strongly convex with minimizer x^*. Apply the SVRG algorithm with constant stepsize α and p such that*

$$0 < \alpha < \frac{1}{4L} \quad \text{and} \quad p > \frac{1}{\gamma\alpha(1 - 4L\alpha)}. \tag{60}$$

Then

$$\mathbb{E}[f(x^k)] - f(x^*) \leq \rho^k (f(x^0) - f(x^*)), \quad k \geq 0, \tag{61}$$

where $\rho := \dfrac{1}{\gamma\alpha(1 - 2L\alpha)p} + \dfrac{2L}{1 - 2L\alpha} < 1.$
Thus $\lim_{k\to\infty} \mathbb{E}[f(X^k)] = f(x^*)$ *linearly.*

Before we prove Theorem 13, we need a useful lemma that concerns the smoothness properties of f_j and f.

Lemma 6. *Let f_i for each $i \in \{1, \ldots, m\}$ be convex and satisfy Assumption 1. Let f in (43) be γ-strongly convex with minimizer x^*. Then, for all $x \in \mathbb{R}^n$,*

$$\|\nabla f_i(x) - \nabla f_i(x^*)\| \leq 2L[f_i(x) - f_i(x^*) - \nabla f_i(x^*)^T(x - x^*)]. \tag{62}$$

Furthermore, if $\mathcal{S} = \{j\}$ is chosen uniformly at random from $\{1, \ldots, m\}$ (with $|\mathcal{S}| = 1$), then for all $x \in \mathbb{R}^n$,

$$\mathbb{E}_{\mathcal{S}}[\|\nabla f_j(x) - \nabla f_j(x^*)\|^2] \leq 2L[f(x) - f(x^*)]. \tag{63}$$

Proof. (Lemma 6) To show (62)[13], define $h(x) = f_i(x) - f_i(x^*) - \nabla f_i(x^*)^T(x - x^*)$, $x \in \mathbb{R}^n$. Since f_i is a convex function, and x^* is fixed, h is also a convex function, since h is the sum of a convex function $f_i - f_i(x^*)$ and a linear function in x, namely $(-\nabla f_i(x^*))^T(x - x^*)$; recall that any linear function is also convex

[13] Note that (62) is true for any points x and x^* and we do not use that x^* is the minimizer of f (but not necessarily f_i).

and the sum of convex functions is also a convex function. Since h is a convex function and $\nabla h(x^*) = \nabla f_i(x^*) - \nabla f(x^*) = 0$, it follows that x^* is a global minimizer of h and so $h(x) \geq h(x^*)$ for all $x \in \mathbb{R}^n$. In particular,

$$h\left(x - \frac{1}{L}\nabla h(x)\right) \geq h(x^*) = 0, \quad x \in \mathbb{R}^n. \tag{64}$$

Note now that $\nabla h(x) = \nabla f_i(x) - \nabla f(x^*)$ and so h is also L-smooth. It follows that h satisfies the 'Foundational Inequality of Steepest Descent' (Lemma 3, Lectures 1–8) namely, for all $x \in \mathbb{R}^n$,

$$h\left(x - \frac{1}{L}\nabla h(x)\right) \leq h(x) - \frac{1}{2L}\|\nabla h(x)\|^2.$$

This and (64) imply that $h(x) \geq \frac{1}{2L}\|\nabla h(x)\|^2$, which together with the definition of h and ∇h, provides (62).

Let us now prove (63). [Recall the derivation (47); similar derivation here for the first two equalities.]

$$\begin{array}{rcl}
\mathbb{E}_{\mathcal{S}}[\|\nabla f_j(x) - \nabla f_j(x^*)\|^2] & = & \sum_{i=1}^m \mathbb{E}(\|\nabla f_i(x) - \nabla f_i(x^*)\|^2 | \mathcal{S} = i)\, \mathbb{P}(\mathcal{S} = i) \\
& = & \frac{1}{m}\sum_{i=1}^m \|\nabla f_i(x) - \nabla f_i(x^*)\|^2 \\
& \leq & \frac{2L}{m}\sum_{i=1}^m [f_i(x) - f_i(x^*) - \nabla f_i(x^*)^T(x - x^*)] \\
& = & 2L[f(x) - f(x^*) - \nabla f(x^*)^T(x - x^*)],
\end{array}$$

where in the first inequality we used (62) for each f_i, and in the last equality we used the definition of f in (43) and its gradient ∇f. Now (63) follows by using that $\nabla f(x^*) = 0$ as x^* is the minimizer of f. □

Proof. (Theorem 13) Recall the simple identity: $\|a + b\|^2 \leq 2(\|a\|^2 + \|b\|^2)$ for any a and b in $\mathbb{R}^n$. Recall the construction of the SVRG algorithm and $G_{k,j} = \nabla f_{\mathcal{S}_{k,j}}(\bar{X}^{k,j}) - \nabla f_{\mathcal{S}_{k,j}}(X^k) + \nabla f(X^k)$. Adding and subtracting $\nabla f_{\mathcal{S}_{k,j}}(x^*)$ to $G_{k,j}$ and using the simple inequality, we obtain

$$\begin{array}{l}
\mathbb{E}_{\mathcal{S}_{k,j}}[\|G_{k,j}\|^2] \\
\leq 2\,\mathbb{E}[\|\nabla f_{\mathcal{S}_{k,j}}(\bar{X}^{k,j}) - \nabla f_{\mathcal{S}_{k,j}}(x^*)\|^2] + 2\,\mathbb{E}[\|\nabla f_{\mathcal{S}_{k,j}}(x^*) - \nabla f_{\mathcal{S}_{k,j}}(X^k) + \nabla f(X^k)\|^2] \\
\leq 2\,\mathbb{E}[\|\nabla f_{\mathcal{S}_{k,j}}(\bar{X}^{k,j}) - \nabla f_{\mathcal{S}_{k,j}}(x^*)\|^2] + 2\,\mathbb{E}[\|\nabla f_{\mathcal{S}_{k,j}}(X^k) - \nabla f_{\mathcal{S}_{k,j}}(x^*) - (\nabla f(X^k) - \nabla f(x^*))\|^2].
\end{array}$$

where in the last equality we used that $\nabla f(x^*) = 0$ in the second term; note that $\mathbb{E}$ above is with respect to $\mathcal{S}_{k,j}$. A standard property is that $\mathbb{E}[\|Y - \mathbb{E}[Y]\|^2] = \mathbb{E}[\|Y\|^2] - \|\,\mathbb{E}[Y]\|^2 \leq \mathbb{E}[\|Y\|^2]$. Using this, and that we have unbiased gradient estimates, $\mathbb{E}_{\mathcal{S}_{k,j}}[\nabla f_{\mathcal{S}_{k,j}}(X^k) - \nabla f_{\mathcal{S}_{k,j}}(x^*)] = \nabla f(X^k) - \nabla f(x^*)$, we deduce

$$\mathbb{E}_{\mathcal{S}_{k,j}}[\|G_{k,j}\|^2] \leq 2\,\mathbb{E}[\|\nabla f_{\mathcal{S}_{k,j}}(\bar{X}^{k,j}) - \nabla f_{\mathcal{S}_{k,j}}(x^*)\|^2] + 2\,\mathbb{E}[\|\nabla f_{\mathcal{S}_{k,j}}(X^k) - \nabla f_{\mathcal{S}_{k,j}}(x^*)\|^2].$$

Apply (63) twice – to both terms in the last inequality, and deduce

$$\mathbb{E}_{\mathcal{S}_{k,j}}[\|G_{k,j}\|^2] \leq 4L[f(\bar{X}^{k,j}) - f(x^*) + f(X^k) - f(x^*)]. \tag{65}$$

This will help us evaluate the distance to the solution, using the definition of the inner iterates and $\mathbb{E}_{\mathcal{S}_{k,j}}[G_{k,j}] = \nabla f(\bar{X}^{k,j})$,

$$\begin{aligned}
&\mathbb{E}_{\mathcal{S}_{k,j}}[\|\bar{X}^{k,j+1} - x^*\|^2] \\
&= \|\bar{X}^{k,j} - x^*\|^2 - 2\alpha(\bar{X}^{k,j} - x^*)^T \mathbb{E}_{\mathcal{S}_{k,j}}[G_{k,j}] + \alpha^2 \mathbb{E}_{\mathcal{S}_{k,j}}[\|G_{k,j}\|^2] \\
&= \|\bar{X}^{k,j} - x^*\|^2 - 2\alpha(\bar{X}^{k,j} - x^*)^T \nabla f(\bar{X}^{k,j}) + \alpha^2 \mathbb{E}_{\mathcal{S}_{k,j}}[\|G_{k,j}\|^2]
\end{aligned}$$

which further becomes, given the convexity of f, $(x - x^*)^T \nabla f(x) \geq f(x) - f(x^*)$, that

$$\begin{aligned}
&\mathbb{E}_{\mathcal{S}_{k,j}}[\|\bar{X}^{k,j+1} - x^*\|^2] \\
&\leq \|\bar{X}^{k,j} - x^*\|^2 - 2\alpha[f(\bar{X}^{k,j}) - f(x^*)] + \alpha^2 \mathbb{E}_{\mathcal{S}_{k,j}}[\|G_{k,j}\|^2] \\
&\leq \|\bar{X}^{k,j} - x^*\|^2 - 2\alpha[f(\bar{X}^{k,j}) - f(x^*)] + \alpha^2 4L[f(\bar{X}^{k,j}) - f(x^*) + f(X^k) - f(x^*)],
\end{aligned}$$

where in the last inequality, we applied (65). And so we conclude for now that

$$\mathbb{E}_{\mathcal{S}_{k,j}}[\|\bar{X}^{k,j+1} - x^*\|^2] \leq \|\bar{X}^{k,j} - x^*\|^2 - 2\alpha(1 - 2L\alpha)[f(\bar{X}^{k,j}) - f(x^*)] + 4L\alpha^2[f(X^k) - f(x^*)].$$

Taking total expectations (with respect to all iterates - inner and outer) and, we deduce,

$$2\alpha(1 - 2L\alpha)\,\mathbb{E}[f(\bar{X}^{k,j}) - f(x^*)] \leq \mathbb{E}[\|\bar{X}^{k,j} - x^*\|^2 - \|\bar{X}^{k,j+1} - x^*\|^2] + 4L\alpha^2\,\mathbb{E}[f(X^k) - f(x^*)].$$

Summing this over j in $\{0, \ldots, p-1\}$, and noting that $\frac{1}{p}\sum_{j=0}^{p-1} \mathbb{E}[f(\bar{X}^{k,j})] = \mathbb{E}[f(X^{k+1})]$, we obtain

$$\begin{aligned}
&2\alpha(1 - 2L\alpha)p\,\mathbb{E}[f(X^{k+1}) - f(x^*)] \\
&\quad \leq \mathbb{E}[\|\bar{X}^{k,0} - x^*\|^2 - \|\bar{X}^{k,p} - x^*\|^2] + 4L\alpha^2 p\,\mathbb{E}[f(X^k) - f(x^*)] \\
&\quad \leq \mathbb{E}[\|X^k - x^*\|^2] + 4L\alpha^2 p\,\mathbb{E}[f(X^k) - f(x^*)] \\
&\quad \leq \left(\tfrac{2}{\gamma} + 4L\alpha^2 p\right)\mathbb{E}[f(X^k) - f(x^*)]
\end{aligned}$$

where we used $\bar{X}^{k,0} = X^k$ in the second inequality, and $f(x) - f(x^*) \geq \frac{\gamma}{2}\|x - x^*\|^2$ in the last inequality (this property is a consequence of Proposition 2(iii)(Lectures 1–8) with $x = x^*$ and $\nabla f(x^*) = 0$). The required conclusion now follows. □

Many variants of SVRG algorithms. Prominently, SAGA is an SVRG variant that does not require the calculation of full gradient.

References

[1] A. Beck. First Order Methods in Optimization. MOS-SIAM Series on Optimization, 2017.

[2] S.J. Wright. Optimization Algorithms for Data Analysis. http://www.optimization-online.org/DB FILE/2016/12/5748.pdf

[3] S.J. Wright. Coordinate descent algorithms. Mathematical Programming, 151:3–34, 2015. https://arxiv.org/abs/1502.04759

[4] D.P. Woodruff. Sketching as a Tool for Numerical Linear Algebra" https://arxiv.org/abs/1411.4357

[5] L. Bottou, F.E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. SIAM Review, 59(1): 65-98, 2017.

[6] Z. Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. The Journal of Machine Learning Research, 18(1): 8194-8244, 2017.

[7] J. Wright and Y. Ma. High-dimensional data analysis with low-dimensional models. CUP, 2021.

[8] A.S. Nemirovskii and D.B. Yudin. Complexity of problems and efficiency of optimization methods. "Nauka" Moskow, 1979, (Russian).

[9] N. Shor. Minimization methods for Non-Differentiable Functions. (Springer-Verlag, Berlin, 1985)

[10] B. Polyak. Introduction to Optimization. (Optimization Software Inc., Publications Division, New York, 1987)

[11] Y. Nesterov. A Method of Solving a Convex Programming Problem with Convergence Rate $O(1/k^2)$. Soviet Math. Dokl., Vol 27 (1983), No 2.

[12] Y. Nesterov. Smooth Minimization of Non-Smooth Functions. Math. Program., Ser. A 103, 127-152 (2005).