# Momentum and Second order methods

June 7, 2023

Stochastic Gradient Descent (SGD) first introduced in [1] cast as a stochastic estimation problem. Enjoys nice properties with big data setup, where computing the full gradient is expansive and redundant. It is widely used today in machine learning, with multiple variants (accelerations, preconditioning, second-order methods, etc.)

These notes follow and summarize the presentation one found in [2]. Today, we will focus on

1. (Recap) Expression of SGD
2. Acceleration of the method with momentum
3. Second order methods
4. Adam, Adagrad, RMSProp, etc.

We first give the general framework of study.

## 1 Recall SGD

### 1.1 Notations

The samples are in the space $\mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$. The notations that will be used will allow to study different forms of SGD (sample based, mini-batch, full-batch).

The **estimator** is denoted by $h\colon \mathbb{R}^{d_x} \times \mathbb{R}^{d} \to \mathbb{R}^{d_y}$, where $\mathbb{R}^d$ is the *parameter space*. The **loss function** is denoted by $\ell\colon \mathbb{R}^{d_y} \times \mathbb{R}^{d_y} \to \mathbb{R}$, and is at least assumed to be differentiable (with possible extension to the case where sub-differentiability alone is assumed).

The composition of the prediction function and the loss function is denoted by $f\colon \mathbb{R}^{d} \times \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \to \mathbb{R}$, $(w; x, y) \mapsto f(w, x, y) \coloneqq \ell(y, h(x, w))$.

In addition, a **random variable** $\xi$ with values in $\mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$ will encode how the samples are drawn. For instance, if $\xi$ has $n$ different realizations $\{(x_i, y_i)\}_{i\in[n]}$ with equal probability, the expectation

$$\mathbb{E}_\xi[f(w;\xi)] = \frac{1}{n}\sum_{i\in[n]} f(w; , x_i, y_i)$$

will be the empirical risk on the training dataset $\{(x_i, y_i)\}_{i\in[n]}$. If $\xi$ draws samples from $\mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$ according to the true data distribution, $\mathbb{E}_\xi[f(w;\xi)]$ will be the expected (or true) risk.

The **objective** $F\colon \mathbb{R}^d \to \mathbb{R}$ is defined as

$$F(w) \coloneqq \mathbb{E}_\xi[f(w;\xi)]$$

and, depending on $\xi$, can encode the expected or empirical risk. The value $\nabla F(w_k)$ is named the **true gradient**, and will usually not be available.

For instance, batch gradient (regular gradient descent) can be recovered when $\xi$ is the uniform sampling of a training dataset $\{(x_i, y_i)\}_{i\in[n]}$, in which case $F(w) = \frac{1}{n}\sum_{i\in[n]} f(w; x_i, y_i)$ is the empirical risk and the update $w_{k+1} = w_k - \alpha\nabla F(w_k)$ is the gradient descent update with step size $\alpha$. We know that, when $F$ is $c$-strongly convex and $L$-smooth, with $\alpha < 1/L$, the gradient descent converges linearly. We will try to understand what happens in the case of a SGD update.

## 1.2 Algorithms

In what follows, we assumed that three tools exist:

1. a mechanism for generating the realization of a **random variable** $\xi_k$, where $\{\xi_k\}_{k\in\mathbb{N}}$ is a sequence of jointly independent random variables
2. given $w_k \in \mathbb{R}^d$ and the realization of $\xi_k$, a way to compute a **stochastic vector** $g(w_k, \xi_k) \in \mathbb{R}^d$.
3. given an iteration $k \in \mathbb{N}$, a way to compute the **step size** $\alpha_k > 0$.

The general SGD algorithm is spelled out in Algorithm 1. It relies on the three previous assumptions. We give the update rule of the full gradient method to comparison (in which case $F = R_n$).

$$w_{k+1} = w_k - \alpha \nabla R_n(w_k) \tag{FG}$$

$$w_{k+1} = w_k - \alpha_k g(w_k, \xi_k) \tag{SG}$$

$$w_{k+1} = w_k - \alpha_k g(w_k, \xi_k) + \beta_k (w_k - w_{k+1}) \tag{SGMnt}$$

$$w_{k+1} = w_k - \alpha_k g(w_k + \beta_k(w_k - w_{k+1}), \xi_k) + \beta_k (w_k - w_{k+1}) \tag{NAG}$$

**Algorithm 1** Stochastic Gradient Descent algorithm [2, Algorithm 4.1].

```
Choose an initial iterate $w_0$
for $k = 0, 1, \dots$ do
    Generate a realization of the random variable $\xi_k$
    Compute a stochastic vector $g(w_k, \xi_k)$
    Choose a step size $\alpha_k > 0$
    Set the new iterate as $w_{k+1} \leftarrow w_k - \alpha_k g(w_k, \xi_k)$.
end for
```

The step size sequence $\{\alpha_k\}_{k\in\mathbb{N}}$ is key; the convergence of the algorithm will rely partly on it.

The random variable $\xi$ can account for single-sample expression or a mini-batch of samples.

The stochastic estimation $g(w_k, \xi_k)$ is relatively generic. In different situations, it will correspond to a given value (e.g. unbiased estimator of the true gradient $\nabla F(w_k)$). For instance,

$$g(w_k, \xi_k) = \begin{cases} \nabla f(w_k; \xi_k) & \text{simple SGD} \\ {n_k}^{-1} \displaystyle\sum_{i\in[n_k]} \nabla f(w_k; \xi_{k,i}) & \text{mini-batch SGD} \\ H_k {n_k}^{-1} \displaystyle\sum_{i\in[n_k]} \nabla f(w_k; \xi_{k,i}) & \text{preconditioned SGD} \end{cases}$$

where one can choose a positive definite scaling matrix $H_k$ and a batch size $n_k$.

The *angle* between $g(w_k, \xi_k)$ and $\nabla F(w_k) = \mathbb{E}_\xi[f(w_k; \xi)]$ will also be relevant in the convergence analysis.

Recall the expression of the (full) batch gradient descent:

$$w_{k+1} \leftarrow w_k - \alpha_k R_n(w_k) = w_k - \frac{\alpha_k}{n} \sum_{i=1}^{n} \nabla f_i(w_k) \tag{1}$$

In comparison, the (simple) stochastic gradient descent is defined as

$$w_{k+1} \leftarrow w_k - \alpha_k \nabla f_{i_k}(w_k) \tag{2}$$

$i_k$ is chosen randomly from $\{1, \dots, n\}$.

**Theoretical Motivation** Different complexity for the two algorithms.

- For strongly convex $R_n$, batch gradient achieves linear convergence:
$$R_n(w_k) - R_n^* = O(\rho^k), \quad \rho \in (0,1).$$
For $\varepsilon$-approximation, the number of iterations required is therefore $O(\log \varepsilon^{-1})$. Since evaluate of $R_n(w_k)$ is in $O(n)$, this totals to complexity in $O(n \log \varepsilon^{-1})$.

- As we will see in Theorem 2, in the same setting, SGD achieves
$$\mathbb{E}[R_n(w_k) - R_n^*] = O(k^{-1}),$$
which does not depend on $n$. Therefore, the total complexity for an $\varepsilon$-approximation is $O(\varepsilon^{-1})$. For large $n$, this can be more efficient than batch gradient.

**Mini-batch setting** In order to reduce the noise stemming from the random selection of one sample, the mini-batch setting is widely used.

In this case, $\xi_k$ is a subset $\mathcal{S}_k \subset \{1, \dots, n\}$.

$$w_{k+1} \leftarrow w_k - \frac{\alpha_k}{|\mathcal{S}_k|} \sum_{i \in \mathcal{S}_k} \nabla f_i(w_k)$$

Reducing the variance of the gradients estimate, this helps choosing the step sizes $\{\alpha_k\}$.

# 2 Analysis of SGD

The analysis will assume some properties on the objective $F$, which is either the expected or the empirical risk.

## 2.1 Expression of the loss

The objective will be $F \colon \mathbb{R}^d \to \mathbb{R}$, which represents either the expected or empirical loss, depending on the law of $\xi$. In the following $P$ stands for the true data distribution, whereas $P_n$ stands for the uniform distribution over a training dataset of $n$ points $\{(x_i, y_i)\}_{i \in [n]}$.

$$F(w) = \begin{cases} R(w) = \mathbb{E}[f(w;\xi)] & \text{law of } \xi \text{ is } P \\ R_n(w) = n^{-1} \sum_{i \in [n]} f_i(w) & \text{law of } \xi \text{ is } P_n \end{cases}$$

The analysis is valid for any of the two value for $F$, so that in what follows, $F$ can be either the true or the empirical risk.

## 2.2 Lipschitz-smoothness assumption

The first key assumption on $F$ is its $L$-smoothness.

**Assumption 1.** *$F$ is $L$-smooth with $L > 0$, i.e. for any $(w, \overline{w}) \in (\mathbb{R}^d)^2$,*

$$\|\nabla F(w) - \nabla F(\overline{w})\|_2 \leqslant L \|w - \overline{w}\|_2 \implies F(w) \leqslant F(\overline{w}) + \langle \nabla F(\overline{w}), w - \overline{w} \rangle + \frac{L}{2} \|w - \overline{w}\|_2^2. \tag{3}$$

*Proof of the implication.* Use the function $e(t) = E(x + t(y - x))$ to write $e(1) = e(0) + \int_0^1 e'(t)\,\mathrm{d}t$ and conclude with Cauchy-Schwarz and $L$-smoothness (3) inequalities. □

Assumption 1 alone lead to the following lemma.

**Lemma 1** ([2, Lemma 4.2])**.** *Assume that $F$ is $L$-smooth. Then, at any iteration $k \in \mathbb{N}$, writing $\mathbb{E}_{\xi_k}[\cdot]$ the expectation with respect to $\xi_k$ given the iterate $w_k$ and under the update of Algorithm 1,*

$$\mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k) \leqslant -\alpha_k \nabla F(w_k)^\top \mathbb{E}_{\xi_k}[g(w_k, \xi_k)] + \frac{\alpha_k^2 L}{2} \mathbb{E}_{\xi_k}[\|g(w_k, \xi_k)\|_2^2]. \tag{4}$$

*Proof.* Write $F(w_{k+1}) - F(w_k) \leqslant \nabla F(w_k)^\top (w_{k+1} - w_k) + \frac{L}{2}\|w_{k+1} - w_k\|_2^2$ and SG update $w_{k+1} - w_k = -\alpha_k g(w_k, \xi_k)$. Take expectation with respect to $\xi_k$ given $w_k$, for which $F(w_k)$ is constant (only $w_{k+1}$ depends on $\xi_k$). □

At any time step $k \in \mathbb{N}$ (i.e. irrelevant of how we arrive at $w_k$), Lemma 1 provides a quantification of the objective decrease but subject to the stochastic estimation $g(w_k, \xi_k)$. A more detailed control over $g(w_k, \xi_k)$ is therefore necessary in order to progress.

**Assumption 2** ([2, Assumption 4.3])**.** *1. All the iterates $\{w_k\}_{k \in \mathbb{N}}$ are in an open set $\Omega$ such that $\inf_{w \in \Omega} F(w) > -\infty$ (i.e. $F|_\Omega \not\equiv -\infty$).*

*2. There exist scalars $\mu_G \geqslant \mu > 0$ such that, for all $k \in \mathbb{N}$,*

$$\langle \nabla F(w_k), \mathbb{E}_{\xi_k}[g(w_k, \xi_k)] \rangle \geqslant \mu \|\nabla F(w_k)\|_2^2, \tag{5}$$

*and*

$$\|\mathbb{E}_{\xi_k}[g(w_k, \xi_k)]\|_2 \leqslant \mu_G \|\nabla F(w_k)\|_2. \tag{6}$$

*3. There exist scalars $M \geqslant 0$ and $M_V \geqslant 0$ such that, for all $k \in \mathbb{N}$,*

$$\mathbb{V}_{\xi_k}[g(w_k, \xi_k)] \coloneqq \mathbb{E}_{\xi_k}[\|g(w_k, \xi_k)\|_2^2] - \|\mathbb{E}_{\xi_k}[g(w_k, \xi_k)]\|_2^2 \leqslant M + M_V \|\nabla F(w_k)\|_2^2.$$

Assumption 2.1 is a technical assumption in order for the problem to be well-posed.

Assumption 2.2 enforces that, in expectation, the stochastic $g(w_k, \xi_k)$ has a large enough angle with the true gradient $\nabla F(w_k)$ (5), without having a too large norm (6).

Assumption 2.3 ensures that the variance of the estimation is bounded. Notice how the variance might not vanish even though $\|\nabla F(w_k)\|$ does (in the case where $M > 0$).

All together, Assumption 2 requires that the second moment of $g(w_k, \xi_k)$ satisfies

$$\mathbb{E}_{\xi_k}[\|g(w_k, \xi_k)\|_2^2] \leqslant M + M_G \|\nabla F(w_k)\|_2^2 \qquad \text{with} \quad M_G \coloneqq M_V + \mu_G^2 \geqslant \mu > 0. \tag{7}$$

**Lemma 2** ([2, Lemma 4.4])**.** *Under Assumptions 1 and 2, the iterates of Algorithm 1 satisfy*

$$\begin{align} \mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k) &\leqslant -\mu \alpha_k \|\nabla F(w_k)\|_2^2 + \frac{\alpha_k^2 L}{2} \mathbb{E}_{\xi_k}[\|g(w_k, \xi_k)\|_2^2] \tag{8} \\ &\leqslant -\alpha_k \left( \mu - \frac{\alpha_k L M_G}{2} \right) \|\nabla F(w_k)\|_2^2 + \frac{\alpha_k^2 L M}{2} \tag{9} \end{align}$$

*Proof.* Combine Equation (4) with Equations (5) and (6). □

Lemma 2 ensures that, given any iterate $k$, the next expected value of the objective will decrease if $\mu - \frac{\alpha_k L M_G}{2}$ is positive, at least until $\|\nabla F(w_k)\|_2^2$ is not too small. After this point, the non-vanishing term $\frac{\alpha_k^2 L M}{2}$ renders the bound loose to show decrease of the (expected) objective.

We require one more assumption in order to show linear convergence of the iterates.

### 2.3 Strongly convex objectives

The following assumption will be critical to establish the convergence rate with a fixed step size.

**Assumption 3.** *$F$ is $c$-strongly convex, for $c > 0$.*

This assumption comes with its inequalities.

**Lemma 3.** *Under Assumption 3, $F_* := \min_w F(w)$ exists and is unique, and we have*

$$\forall w \in \mathbb{R}^d, \quad 2c(F(w) - F_*) \leqslant \|\nabla F(w)\|^2 \tag{10}$$

*Proof.* Existence and uniqueness of the minimizer is implied by the strict convexity of $F$, implied itself by Assumption 3.

For the inequality, one studies the quadratic $q(\overline{w}) := F(w) + \langle \nabla F(w), \overline{w} - w\rangle + \frac{c}{2}\|\overline{w} - w\|_2^2$ which lowerbounds $F(\overline{w})$ for all $(w, \overline{w}) \in (\mathbb{R}^d)^2$. □

### 2.3.1 Constant step size

We are now ready to state the result of SGD convergence for $L$-smooth and $c$-strongly convex objective with constant step size $\bar{\alpha}$.

**Theorem 1** ([2, Theorem 4.6]). *Assume Assumptions 1 to 3. Let $F_* = \min_w F(w)$. Then, by running Algorithm 1 with a fixed step size $0 \leqslant \bar{\alpha} \leqslant \frac{\mu}{LM_G}$, each iterate $k \in \mathbb{N} \setminus \{0\}$ satisfies*

$$\begin{aligned} \mathbb{E}[F(w_k) - F_*] &\leqslant \frac{\bar{\alpha}LM}{2c\mu} + (1 - \bar{\alpha}c\mu)^{k-1}\left(F(w_0) - F_* - \frac{\bar{\alpha}LM}{2c\mu}\right) \\ &\xrightarrow[k\to\infty]{} \frac{\bar{\alpha}LM}{2c\mu}, \end{aligned} \tag{11}$$

*where $\mathbb{E}[\cdot]$ denotes the total expected value, i.e. with respect to the law of $\{\xi_k\}_{k\in\mathbb{N}}$.*

*Proof.* At one time-step $k \in \mathbb{N}$, the previous bound (9) gives

$$\mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k) \leqslant -\bar{\alpha}\left(\mu - \frac{\bar{\alpha}LM_G}{2}\right)\|\nabla F(w_k)\|^2 + \frac{\bar{\alpha}^2 LM}{2},$$

which combined with $-\|\nabla F(w_k)\|^2 \leqslant -2c(F(w_k) - F_*)$ further gives

$$\mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k) \leqslant -2c\bar{\alpha}\left(\mu - \frac{\bar{\alpha}LM_G}{2}\right)(F(w_k) - F_*) + \frac{\bar{\alpha}^2 LM}{2}.$$

Now, since $0 < \bar{\alpha} \leqslant \frac{\mu}{LM_G}$, $\mu - \frac{\bar{\alpha}LM_G}{2} \geqslant \frac{\mu}{2}$ and

$$\mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k) \leqslant -\bar{\alpha}c\mu(F(w_k) - F_*) + \frac{\bar{\alpha}^2 LM}{2}.$$

Subtracting $F_*$ from both sides and factorizing by $F(w_k) - F_*$ leads to

$$\mathbb{E}_{\xi_k}[F(w_{k+1})] - F_* \leqslant (1 - \bar{\alpha}c\mu)(F(w_k) - F_*) + \frac{\bar{\alpha}^2 LM}{2}.$$

Taking the total expected value (with respect to $\xi_1 \otimes \cdots \otimes \xi_k$, denoted by $\mathbb{E}[\cdot]$) gives

$$\mathbb{E}[F(w_{k+1}) - F_*] \leqslant (1 - \bar{\alpha}c\mu)\mathbb{E}[F(w_k) - F_*] + \frac{\bar{\alpha}^2 LM}{2}.$$

This is an arithmetico-geometric sequence $u_{k+1} = au_k + b$, with ratio $a := 1 - \bar{\alpha}c\mu \neq 1$ and with fixed-point $b(1-a)^{-1} = \bar{\alpha}^2 LM(2\bar{\alpha}c\mu)^{-1} = \bar{\alpha}LM(2c\mu)^{-1}$, so that upper-bounding by the geometric sequence $v_k := u_k - \frac{b}{1-a}$ of ratio $a$ gives

$$\mathbb{E}[F(w_{k+1}) - F_*] - \frac{\bar{\alpha}LM}{2c\mu} \leqslant (1 - \bar{\alpha}c\mu)\left(\mathbb{E}[F(w_k) - F_*] - \frac{\bar{\alpha}LM}{2c\mu}\right) \tag{12}$$

Note that $0 \leqslant 1 - \bar{\alpha}c\mu < 1$, since $1 \geqslant \frac{L}{c} \geqslant \bar{\alpha}c\mu > 0$ (we have a *contraction*: at each time-step $k$, the value of the sequence shrinks).

The desired inequality (11) is obtained by applying (12) inductively for $k = 0, 1, \ldots$. □

Theorem 1 ensures a convergence to a region of minimizer. The fixed step size prevents from showing the convergence to a minimizer, in case $g(w_k, \xi_k)$ is noisy ($\mu \neq 1$). This is the topic of the next section.

### 2.3.2 Diminishing step size

We loose the assumption on the fixed step size, but keep Assumptions 1 to 3. Moreover, it will be required for the step size sequence $\{\alpha_k\}_{k\in\mathbb{N}}$ to satisfy

$$\sum_{k=0}^{\infty} \alpha_k = \infty \quad \text{and} \quad \sum_{k=0}^{\infty} \alpha_k^2 < \infty. \tag{13}$$

**Theorem 2.** *Under Assumptions 1 to 3, suppose Algorithm 1 is run with a step size sequence $\{\alpha_k\}_{k\in\mathbb{N}}$ satisfying*

$$\alpha_k = \frac{\beta}{\gamma + k} \quad \textit{for some } \beta > \frac{1}{c\mu} \textit{ and } \gamma > 0 \textit{ such that } \alpha_0 \leqslant \frac{\mu}{LM_G}.$$

*Then, denoting $F_* = \min_w F(w)$, the expected optimality gap satisfies*

$$\forall k \in \mathbb{N}, \qquad \mathbb{E}[F(w_k) - F_*] \leqslant \frac{\nu}{\gamma + k}, \tag{14}$$

*where*

$$\nu := \max\left\{\frac{\beta^2 LM}{2(\beta c\mu - 1)}, \gamma(F(w_0) - F_*)\right\} \tag{15}$$

*Proof.* The proof is performed by first bounding at one time-step $k \in \mathbb{N}$ the value $\mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k)$. The definition of $\{\alpha_k\}_{k\in\mathbb{N}}$ ensures that, for all $k \in \mathbb{N}$, $\alpha_k LM_G \leqslant \alpha_0 LM_G \leqslant \mu$. Lemma 2 and (10) then give

$$\begin{aligned} & \mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k) \leqslant -\alpha_k c\mu(F(w_k) - F_*) + \frac{\alpha_k^2 LM}{2} \\ \implies \qquad & \mathbb{E}[F(w_{k+1}) - F_*] \leqslant (1 - \alpha_k c\mu)\mathbb{E}[F(w_k) - F_*] + \frac{\alpha_k^2 LM}{2}, \end{aligned} \tag{16}$$

where the second equation is obtained by subtracting $F_*$ from both sides and taking the total expectation, with respect to $\{\xi_j\}_{j\in[k]}$.

Now, the inequality (14) can be showed by induction on $k \in \mathbb{N}$.

By definition of $\nu$, $\gamma(F(w_0) - F_*) \leqslant \nu$, and (14) is satisfied for $k = 0$. Assuming (14) is true for some $k \in \mathbb{N}$, Equation (16) and the definition of $\{\alpha_j\}_{j\in\mathbb{N}}$ allow to write

$$\begin{aligned} \mathbb{E}[F(w_{k+1}) - F_*] &\leqslant \left(1 - \frac{\beta c\mu}{\hat{k}}\right)\frac{\nu}{\hat{k}} + \frac{\beta^2 LM}{2\hat{k}^2} \quad \text{with } \hat{k} := \gamma + k \\ &= \left(\frac{\hat{k} - \beta c\mu}{\hat{k}^2}\right)\nu + \frac{\beta^2 LM}{2\hat{k}^2} \\ &= \left(\frac{\hat{k} - 1}{\hat{k}^2}\right)\nu - \left(\frac{2(\beta c\mu - 1)}{2\hat{k}^2}\right)\nu + \frac{\beta^2 LM}{2\hat{k}^2} \\ &= \left(\frac{\hat{k} - 1}{\hat{k}^2}\right)\nu - (2\hat{k}^2)^{-1}\underbrace{(2(\beta c\mu - 1)\nu - \beta^2 LM)}_{\geqslant 0 \text{ by def. of } \nu} \\ &\leqslant \frac{\nu}{\hat{k} + 1}, \end{aligned}$$

where the last inequality follows since $\hat{k}^2 \geqslant (\hat{k} + 1)(\hat{k} - 1)$. □

Notice how the condition $\beta > (c\mu)^{-1}$ requires that the **step sizes are still "big enough"** in order to obtain a $O(1/k)$ convergence rate. This bound on $\beta$ is a function of $c$, the strong convexity parameter of $F$: the less $F$ is convex, the greater the step sizes should be.

The role of the distance from initialization to optimality, $F(w_0) - F_*$ does not appear with an exponential decay, as it was the case with a fixed step size. Instead, if greater than $\frac{\beta^2 LM}{2\gamma(\beta c\mu - 1)}$, it controls $\nu$ and can slow down the convergence. With appropriate initialization, one can lessen the role played by the gap: start with a constant step size $\bar{\alpha}$ until one obtains the point, called $w_0$, such that $F(w_0) - F_* \leqslant \bar{\alpha} LM/(2c\mu)$ (cf. (11)). Then, taking this new initialization, the first term in (15) dominates.

### 2.4 Non-convex objective

We relax here Assumption 3, and only assume Assumptions 1 and 2. The results are weaker than Theorems 1 and 2.

**Theorem 3** ([2, Theorem 4.8])**.** *Fixed step size $\alpha_k \equiv \bar{\alpha}$ such that $0 \leqslant \bar{\alpha} \leqslant \frac{\mu}{LM_G}$.*
*Then, the expected sum-of-squares and average-squared gradients of $F$ corresponding to the SGD iterates from Algorithm 1 satisfy the following inequalities for all $K \in \mathbb{N}$:*

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=0}^{K}\|\nabla F(w_k)\|_2^2\right] \leqslant \frac{\bar{\alpha} LM}{\mu} + \frac{2(F(w_0) - F_{\text{inf}})}{K\mu\bar{\alpha}} \quad \xrightarrow[K\to\infty]{} \quad \frac{\bar{\alpha} LM}{\mu}.$$

*Proof.* Take the expectation of (9), and telescopic sum for the different time steps. □

The case $M = 0$ recovers the full gradient descent for nonconvex objectives: the sum of squared gradient norms remain finite, implying that $\|\nabla F(w_k)\|_2 \xrightarrow[k\to\infty]{} 0$.

#### 2.4.1 Diminishing step size

There is also a result equivalent to Theorem 2 for non-convex objectives.

**Theorem 4** ([2, Theorem 4.10])**.** *Under Assumptions 1 and 2, assuming that the step size sequence $\{\alpha_k\}_{k\in\mathbb{N}}$ satisfies (13), and with $A_K := \sum_{k=0}^{K} \alpha_k$, the iterates of Algorithm 1 satisfy*

$$\begin{aligned} &\lim_{K\to\infty} \mathbb{E}\left[\sum_{k=0}^{K}\alpha_k\|\nabla F(w_k)\|_2^2\right] < \infty \\ \implies \quad &\mathbb{E}\left[\frac{1}{A_K}\sum_{k=0}^{K}\alpha_k\|\nabla F(w_k)\|_2^2\right] \xrightarrow[K\to\infty]{} 0. \end{aligned}$$

*Proof.* The proof is again obtained by upper-bounding $\mathbb{E}[F(w_{k+1}) - F(w_k)]$ and summing the inequality for $k \in \{0, \ldots, K\}$. □

The behaviour of the algorithm depends on the ratio $\frac{\sum_k \alpha_k^2}{\sum_k \alpha_k}$ like so: taking the total expectation of (9) gives

$$\begin{aligned} \mathbb{E}[F(w_{k+1})] - \mathbb{E}[F(w_k)] &\leqslant -(\mu - \frac{\alpha_k LM_G}{2})\alpha_k \mathbb{E}[\|\nabla F(w_k)\|_2^2] + \frac{\alpha_k^2 LM}{2} \\ &\leqslant -\frac{\mu\alpha_k}{2}\mathbb{E}[\|\nabla F(w_k)\|_2^2] + \frac{\alpha_k^2 LM}{2} \end{aligned}$$

Summing both sides for $k \in \{0, \ldots, K\}$ gives

$$F_{\text{inf}} - \mathbb{E}[F(w_0)] \leqslant \mathbb{E}[F(w_{k+1})] - \mathbb{E}[F(w_0)] \leqslant -\frac{\mu}{2}\sum_{k=0}^{K}\alpha_k \mathbb{E}[\|\nabla F(w_k)\|_2^2] + \frac{LM}{2}\sum_{k=0}^{K}\alpha_k^2$$

Rearranging and dividing by $\mu/2$ gives

$$\sum_{k=0}^{K}\alpha_k \mathbb{E}[\|\nabla F(w_k)\|_2^2] \leqslant \frac{2(\mathbb{E}[F(w_0)] - F_{\text{inf}})}{\mu} + \frac{LM}{\mu}\sum_{k=0}^{K}\alpha_k^2,$$

and

$$\min_k \mathbb{E}[\|\nabla F(w_k)\|_2^2] \leqslant \frac{2(\mathbb{E}[F(w_0)] - F_{\inf})}{\mu \sum_{k=0}^K \alpha_k} + \frac{LM}{\mu} \frac{\sum_{k=0}^K \alpha_k^2}{\sum_{k=0}^K \alpha_k}. \tag{17}$$

The term $\frac{\sum_k \alpha_k^2}{\sum_k \alpha_k}$ is crucial and dictates the rate of convergence of the algorithm.

Example: $\alpha_k \propto \frac{1}{k} \implies \sum_k \alpha_k = O(\log k)$ and $\sum_k \alpha_k^2 = O(1)$, therefore $\min_k \mathbb{E}[\|\nabla F(w_k)\|_2^2] \leqslant O(1/\log k)$.

# 3 Momentum methods

## 3.1 General Momentum

Keep a momentum value during the update rule, appeared in [3]. Additional sequence of momentums $\{\beta_k\}_{k\in\mathbb{N}}$.

Modified update rule:

$$w_{k+1} = w_k - \alpha_k g(w_k, \xi_k) + \beta_k(w_k - w_{k-1}) \tag{18}$$

Motivation: discretization of a certain second-order ODE with friction

Heavy ball method corresponds to constant sequences $\alpha_k \equiv \alpha$, $\beta_k \equiv \beta$. This leads to

$$w_{k+1} = w_k - \alpha \sum_{j=1}^k \beta^{k-j} g(w_j, \xi_j)$$

Intuition: accumulates persistent directions of stochastic gradients

## 3.2 Accelerated Gradient Method

First proposed by Nesterov [4], further studied extensively (see e.g. [5]). It's a two-steps procedure written as

$$\widetilde{w}_k \leftarrow w_k + \beta_k(w_k - w_{k+1}) \tag{19a}$$

$$w_{k+1} \leftarrow \widetilde{w}_k - \alpha_k g(\widetilde{w}_k, \xi_k) \tag{19b}$$

which is equivalent to

$$w_{k+1} \leftarrow w_k - \alpha_k g\big(w_k + \beta_k(w_k - w_{k+1}), \xi_k\big) + \beta_k(w_k - w_{k+1}).$$

This can be interpreted as looking ahead (19a) before estimating the gradient and applying a regular SGD update (19b).

With correct $(\alpha_k, \beta_k)$, the overall update corresponds to a convergence rate of $O(k^{-2})$ vs. $O(k^{-1})$ for steepest descent on convex objectives.

See also [6] for additional expressions.

# 4 Second oder methods

## 4.1 Intuition

Second order development around an iterate $w_k$:

$$q(w) = F(w_k) + \langle \nabla F(w_k), w - w_k \rangle + \frac{1}{2}\langle w - w_k, \nabla^2 F(w_k)(w - w_k) \rangle \tag{20}$$

Find the minimizer of $q$, assuming $\det(\nabla^2 F(w_k)) \neq 0$:

$$\nabla q(w^*) = 0 \implies w^* = w_k - \underbrace{\left(\nabla^2 F(w_k)\right)^{-1}}_{:=P} \nabla F(w_k) \tag{21}$$

Preconditioning by $\left(\nabla^2 F(w_k)\right)^{-1}$.

- regular gradient descent corresponds to taking $\alpha I$ as preconditioner
- only intuition, computation is too expensive in practice
- only comes from linear scaling invariance $Bw_k$ for some $B \succ 0$

Hence, only estimation of $P$, easiest case: $P \equiv D$ is diagonal, which corresponds to scaling the different dimensions of $w_k \in \mathbb{R}^d$.

### 4.2 Adagrad

Stands for Adaptive Gradient. Appeared in [7].

Running sum:

$$[R_k]_i = [R_{k-1}]_i + [g(w_k, \xi_k)]_i^2$$

Update rule:

$$[w_{k+1}]_i = [w_k]_i - \frac{\alpha}{\sqrt{[A_k]_i + \mu}}[g(w_k, \xi_k)]_i \tag{Adagrad}$$

### 4.3 RMSProp (Root Mean Square Propagation)

Estimate the average magnitude of each element of stochastic gradient vector $g(w_k, \xi_k)$ with running average, for some $\lambda \in (0, 1)$:

$$[R_k]_i = (1 - \lambda)[R_{k-1}]_i + \lambda[g(w_k, \xi_k)]_i^2$$

Update of the parameters:

$$[w_{k+1}]_i = [w_k]_i - \frac{\alpha}{\sqrt{[R_k]_i + \mu}}[g(w_k, \xi_k)]_i \tag{RMSProp}$$

### 4.4 Adam (Adaptive moments algorithm)

Appeared in [8].

Combine both Adagrad and RMSProp. Keep first order and second order moments adapted. Update:

$$\begin{aligned}
g_k &:= g(w_k, \xi_k) && \text{Stochastic gradient} && (22)\\
m_{k+1} &= \beta_1 m_k + (1 - \beta_1) g(w_k, \xi_k) && \text{biased first moment estimate} && (23)\\
v_{k+1} &= \beta_2 v_k + (1 - \beta_2) g_k \odot g_k && \text{biased second moment estimate} && (24)\\
\hat{m}_{k+1} &= \frac{m_{k+1}}{1 - \beta_1^k} && && (25)\\
\hat{v}_{k+1} &= \frac{v_k}{1 - \beta_2^k} && && (26)\\
w_{k+1} &= w_k - \alpha \frac{\hat{m}_{k+1}}{\sqrt{\hat{v}_{k+1}} + \varepsilon} && && (27)
\end{aligned}$$

### 4.5 Caveats

It has been observed that the generalization properties of Adam, etc. were sometimes not as good as the one by SGD.

| Algorithm | Function | Step size | Rate | Cost for $\varepsilon$-approximation |
|---|---|---|---|---|
| FG (1) | $c$-strongly convex, $L$-Lipschitz | $\alpha < 1/L$ | $O((1-c/L)^k)$ | $O(n \log 1/\varepsilon)$ |
| | $L$-Lipschitz | $\alpha \leqslant 1/L$ | $O(1/\sqrt{k})$ | $O(n/\varepsilon^2)$ |
| SG (2) | $c$-strongly convex, $L$-smooth | $\alpha_k \propto 1/k$ | $O(1/k)$ | $O(1/\varepsilon)$ |
| | $L$-smooth | $\alpha_k \propto 1/k$ | $O(1/\log k)$ | $O(\exp(1/\varepsilon))$ |
| NAG (19) | Convex, $L$-Lipschitz | $\alpha \leqslant 1/L$ | $O(1/k^2)$ | $O(1/\sqrt{\varepsilon})$ |

Table 1: Summary of the different algorithms presented. FG stands for full gradient, SG for stochastic gradient, NAG for Nesterov Accelerated Gradient.

## 5 Conclusion

We summarize the different setups in Section 5, with a comparison to the full gradient descent.

This presentation allowed to consider a general framework to study different stochastic gradient methods. As in deterministic analyses, the assumptions on the target function brings better convergence rates. Moreover, the step size schedule is key in the convergence results. Overall, there exists a trade-off between the cost per iteration (e.g the cost of estimating the gradient) and the convergence rate for the method. Mini-batch is a mixed method which aims at gaining from noise reduction method (reducing the variance of the estimation) while keeping the cost per iteration low.

The second order methods, and more details on the first order ones, are available in [2].

## References

[1] Herbert Robbins and Sutton Monro. "A Stochastic Approximation Method". In: *The Annals of Mathematical Statistics* 22.3 (Sept. 1951), pp. 400–407. ISSN: 0003-4851, 2168-8990. DOI: 10.1214/aoms/1177729586. (Visited on 02/11/2022).

[2] Léon Bottou, Frank E. Curtis, and Jorge Nocedal. "Optimization Methods for Large-Scale Machine Learning". 2016. DOI: 10.1137/16M1080173. arXiv: 1606.04838.

[3] B.T. Polyak. "Some Methods of Speeding up the Convergence of Iteration Methods". In: *USSR Computational Mathematics and Mathematical Physics* 4.5 (Jan. 1964), pp. 1–17. ISSN: 00415553. DOI: 10.1016/0041-5553(64)90137-5. (Visited on 03/21/2022).

[4] Y. Nesterov. "A Method for Solving the Convex Programming Problem with Convergence Rate O(1/K$\hat{2}$)". In: *Proceedings of the USSR Academy of Sciences* (1983). URL: https://www.semanticscholar.org/paper/A-method-for-solving-the-convex-programming-problem-Nesterov/8d3a318b62d2e970122da35b2a2e70a5d12cc16f (visited on 03/21/2022).

[5] Alexandre d'Aspremont, Damien Scieur, and Adrien Taylor. "Acceleration Methods". In: *Foundations and Trends® in Optimization* 5.1-2 (2021), pp. 1–245. ISSN: 2167-3888, 2167-3918. DOI: 10.1561/2400000036. arXiv: 2101.09545. (Visited on 03/21/2022).

[6] Ilya Sutskever et al. "On the Importance of Initialization and Momentum in Deep Learning". In: (2013).

[7] John Duchi, Elad Hazan, and Yoram Singer. "Adaptive Subgradient Methods for Online Learning and Stochastic Optimization". In: *Journal of Machine Learning Research* 12.61 (2011), pp. 2121–2159. ISSN: 1533-7928. URL: http://jmlr.org/papers/v12/duchi11a.html (visited on 06/07/2023).

[8] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". Dec. 22, 2014. arXiv: 1412.6980 [cs]. URL: http://arxiv.org/abs/1412.6980 (visited on 03/27/2019).