Chapter 2.

Stochastic gradient descent

In this section we will take a closer look at the stochastic gradient method for strongly convex objective functions. We will analyse its behaviour regarding convergence and its convergence rate. Our goal is to determine convergence properties and give inequalities for worst case complexity bounds. The structure of this chapter is based on [BCN18, Sec. 4]. Two main results including some additional information will be shown below. Additional proofs can be found in [BCN18, Sec. 4].

2.1. Generic stochastic gradient descent algorithm

To make the analysis of the algorithm more general, we will define a generic objective function.

Definition 2.1.1 (Generic objective function)**.** For analysing both, expected and empirical risk, let us define the **generic objective function** F as

$$F : \mathbb{R}^d \to \mathbb{R}, w \mapsto \begin{cases} R(w) & = \mathbb{E}\left[f(w;\xi)\right] \\ \text{or} & \\ R_n(w) & = \frac{1}{n}\sum_{i=1}^{n} f_i(w) \end{cases} . \tag{2.1}$$

No, we will specify a generic version of the stochastic gradient descent algorithm and take a closer look on the components of the algorithm in Remark 2.1.2.
Except for the stochastic part included in the gradient calculation, this generic algorithm is close to the original generic algorithm for general descent methods, which we analysed in the lecture *Optimization 1.*

Remark 2.1.2 (Interpretation of Algorithm 1)**.** By analysing Algorithm 1, we can remark a few important characteristics.

Algorithm 1 Stochastic gradient descent method (SGD method)

Require: initial guess $w_1 \in \mathbb{R}^d$
1: **for** $k = 1, 2, \ldots$ **do**
2: generate realization of the random variable ξ_k
3: compute the stochastic vector $g(w_k; \xi_k) \in \mathbb{R}^d$
4: choose a step size α_k
5: $w_{k+1} \leftarrow w_k - \alpha_k \cdot g(w_k; \xi_k)$
6: **end for**

(i) The generation of the realization of the random variable ξ_k directly corresponds to the risk measure we use. If we uniformly choose ξ_k over the finite training set, this corresponds to the empirical risk R_n. Alternatively selecting samples according to the distribution $\mathbb{P}$, we obtain the expected risk R. Furthermore, we assume that the random variable sequence $\{\xi_k\}_{k \in \mathbb{N}}$, generated by the algorithm, is independent.

(ii) The computation of the stochastic vector in line 3 is also very generic. This computation could represent SGD or a (mixed) batch approach

$$g(w_k, \xi_k) = \begin{cases} \nabla f(w_k; \xi_k) \\ \frac{1}{n_k} \sum_{i=1}^{n_k} \nabla f(w_k; \xi_{k,i}) \\ H_k \cdot \frac{1}{n_k} \sum_{i=1}^{n_k} \nabla f(w_k; \xi_{k,i}) \end{cases}, \tag{2.2}$$

with $H_k \in \mathbb{R}^{d \times d}$ being a symmetric positive definite matrix.

(iii) The choice of α_k in line 4 allows us to use a fixed step size as well as a sequence of diminishing step sizes. In the following analysis, we will take a closer look on both options for the step size and its effect on convergence.

2.2. Two fundamental lemmas

The following two fundamental lemmas are build on the smoothness of the objective function. Hence, we need some assumptions regarding the first and second moments.

Assumption 2.2.1 (Lipschitz continuity of the objective functions gradient)**.** *In the following report, we assume that the objective function $F : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable and the gradient function $\nabla F : \mathbb{R}^d \to \mathbb{R}^d$ is Lipschitz continuous with*

Lipschitz constant $L > 0$. In particular, for all $w, \bar{w} \in \mathbb{R}^d$ it holds

$$\|\nabla F(w) - \nabla F(\bar{w})\|_2 \leq L \cdot \|w - \bar{w}\|_2 .$$

Recalling the convergence proof of the gradient descent method in *Optimization 1*, we can see that Assumption 2.2.1 is typical for a gradient based descent method. Otherwise, the gradient would not be a good indicator for how far to move to achieve a decrease in F. We obtain the following proposition.

Proposition 2.2.2. *In the setting of Assumption 2.2.1, we obtain for all $w, \bar{w} \in \mathbb{R}^d$ the inequality*

$$F(w) \leq F(\bar{w}) + \nabla F(\bar{w})^{\mathsf{T}}(w - \bar{w}) + \frac{L}{2} \|w - \bar{w}\|_2^2 . \tag{2.3}$$

Proof. See [BCN18, Appendix B]. □

Before stating the first of two fundamental lemmas, we need some more notations. We use $\mathbb{E}_{\xi_k}[\cdot]$ to denote the expected value with respect to the distribution of the random variable ξ_k given by w_k. As a consequence, the expression $\mathbb{E}_{\xi_k}[F(w_{k+1})]$ makes sense because w_{k+1} depends on ξ_k (cf. Algorithm 1).

Lemma 2.2.3. *Under Assumption 2.2.1, the iterates generated by Algorithm 1 satisfy for all $k \in \mathbb{N}$ the inequality*

$$\mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k) \leq -\alpha_k \nabla F(w_k)^{\mathsf{T}} \mathbb{E}_{\xi_k}[g(w_k, \xi_k)] + \frac{\alpha_k^2 L}{2} \mathbb{E}_{\xi_k}[\|g(w_k, \xi_k)\|_2^2]. \tag{2.4}$$

Proof. See [BCN18, Lemma 4.2]. □

Lemma 2.2.3 shows that Algorithm 1 behaves similarly to a *Markov chain*. Regardless of how the iterates arrived at the step w_k, the expected descent just depends on two properties:

(i) the expected directional derivative of F at w_k along the direction $g(w_k, \xi_k)$, and

(ii) the second moment[1] of $g(w_k, \xi_k)$.

[1] If we recall (1.8) and assume that $g(w_k, \xi_k)$ is an unbiased estimate of $\nabla F(w_k)$, Lemma 2.2.3 yields

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

If we could ensure that the right-hand side of the inequality in (2.4) is negative and bounded from above by a *deterministic* quantity, this would lead to an asymptotically sufficient descent in F. Therefore, we need more assumptions regarding the first and second moments of the stochastic directions. But first of all, we define the variance with respect to the distribution of ξ_k as

$$\begin{aligned}\mathbb{V}\mathrm{ar}_{\xi_k}[g(w,\xi_k)] &:= \mathbb{E}_{\xi_k}\left[(g(w,\xi_k) - \mathbb{E}_{\xi_k}[g(w,\xi_k)])^2\right] \\ &\overset{(*)}{=} \mathbb{E}_{\xi_k}\left[\|g(w,\xi_k)\|_2^2\right] - \|\mathbb{E}_{\xi_k}[g(w,\xi_k)]\|_2^2, \end{aligned} \tag{2.5}$$

for $w \in \mathbb{R}^d$. The equality $(*)$ was proven in the course *Probability Theory* [Kle13, p. 103] and is a basic property of the variance.

Assumption 2.2.4 (Limits for first and second moments)**.** *We assume that the objective function $F : \mathbb{R}^d \to \mathbb{R}$ satisfies the following quantities:*

(i) *$\{w_k\}_k \subseteq \mathcal{O}$, for an open set $\mathcal{O} \subseteq \mathbb{R}^d$, in which the objective function is bounded from below. In particular: for all $w \in \mathcal{O}$ it holds $F(w) \geq F_{\inf}$.*

(ii) *There exist constants $\mu_G \geq \mu > 0$, such that*

$$\nabla F(w_k)^\mathsf{T} \mathbb{E}_{\xi_k}[g(w_k,\xi_k)] \geq \mu \|\nabla F(w_k)\|_2^2 \tag{2.6a}$$

$$\|\mathbb{E}_{\xi_k}[g(w_k,\xi_k)]\|_2 \leq \mu_G \|\nabla F(w_k)\|_2 \tag{2.6b}$$

hold for all $k \in \mathbb{N}$.

(iii) *There are constants $M, M_V \geq 0$, such that the inequality*

$$\mathbb{V}\mathrm{ar}_{\xi_k}[g(w_k,\xi_k)] \leq M + M_V \|\nabla F(w_k)\|_2^2 \tag{2.7}$$

is satisfied for all $k \in \mathbb{N}$.

We want to briefly interpret Assumption 2.2.4:

(i) The first assumption states, that the objective function F is bounded from below in the area explored by the algorithm. Basically, we had the same assumption in the analysis of a generic descent algorithm in *Optimization 1.*

(ii) This point is also similar to a condition of the full gradients analysis. We want to make sure that, in expectation, the algorithm's directions are of sufficient descent. We will see this in Lemma 2.2.5.

(iii) The last assumption restricts the variance in a minor way. If, for example F is a convex quadratic function, the inequality (2.7) allows the variance to be non-zero in every stationary point of F and to grow quadratically in each direction.

Combining Assumption 2.2.4 and Eq. (2.5) lead to

$$\begin{aligned}\mathbb{E}_{\xi_k}\left[\|g(w_k,\xi_k)\|_2^2\right] &= \mathbb{V}\mathrm{ar}_{\xi_k}[g(w_k,\xi_k)] + \|\mathbb{E}_{\xi_k}[g(w_k,\xi_k)]\|_2^2 \\ &\leq M + M_G\|\nabla F(w_k)\|_2^2, \end{aligned} \tag{2.8}$$

for a constant

$$M_G := M_V + \mu_G^2 \geq \mu^2 > 0. \tag{2.9}$$

Lemma 2.2.5. *Under Assumptions 2.2.1 and 2.2.4, the iterates $\{w_k\}_{k\in\mathbb{N}}$ of SGD produced by Algorithm 1 satisfy*

$$\mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k) \leq -\mu\alpha_k\|\nabla F(w_k)\|_2^2 + \frac{1}{2}\alpha_k^2 L\mathbb{E}_{\xi_k}\left[\|g(w_k,\xi_k)\|_2^2\right] \tag{2.10a}$$

$$\leq -\left(\mu - \frac{1}{2}\alpha_k L M_G\right)\alpha_k\|\nabla F(w_k)\|_2^2 + \frac{1}{2}\alpha_k^2 LM \tag{2.10b}$$

for all $k \in \mathbb{N}$.

Proof. See [BCN18, Lemma 4.4]. □

If we take a closer look at the inequalities above and reconsider the interpretation of Lemma 2.2.3, we can see again how SGD behaves in a *Markovian manner*. No matter how SGD arrived at w_k, the random variable w_{k+1} only depends on w_k and ξ_k and *not* on past iterates.

Notice that the difference on the left-hand side is bounded from above by a deterministic quantity.

2.3. SGD for strongly convex objective functions

In this section, we will proceed our analysis of Lemma 2.2.5, by assuming that our objective function is strongly convex. In a first observation, we will see that in expectation the difference of the left-hand side will converge to a deterministic (in general non-zero) quantity, while using SGD with a fixed step size.

Naturally, we will try to refine this result by using a diminishing step size sequence to achieve convergence in expectation[2].

Assumption 2.3.1 (Strongly convex objective function)**.** *We assume that for our generic objective function $F : \mathbb{R}^d \to \mathbb{R}$ there exists a constant $c > 0$ (the so-called modulus), such that for all $(\bar{w}, w) \in \mathbb{R}^d \times \mathbb{R}^d$ the inequality*

$$F(\bar{w}) \geq F(w) + \nabla F(w)^\mathsf{T}(\bar{w} - w) + \frac{1}{2}c \left\|\bar{w} - w\right\|_2^2 \tag{2.11}$$

is satisfied. In particular, in that case F has a unique minimizer $w_* \in \mathbb{R}$; compare Optimization 1. Furthermore, we will define $F_* := F(w_*)$. Taking a look at Eq. (2.3)*, we can also see, that $c \leq L$ has to hold.*

The following proposition provides a useful result, which relates the minimal value F_* with any other value $F(w)$, using the Euclidean-Norm of the gradient.

Proposition 2.3.2. *Using Assumption 2.3.1, the inequality*

$$2c\,(F(w) - F_*) \leq \left\|\nabla F(w)\right\|_2^2 \tag{2.12}$$

follows for all $w \in \mathbb{R}^d$.

Proof. See [BCN18, Appendix B]. □

As mentioned above, in the following first convergence theorem we will not get convergence towards a solution. Instead, we will only get convergence to a neighbourhood of the optimal value.

Remark 2.3.3 (Total expectation)**.** In step $k \in \mathbb{N}$ of the algorithm, we will write $\mathbb{E}[\cdot]$ to denote the expected value with respect to the joint distribution of *all* random variables. Particularly, the total expectation of $F(w_k)$ can be written as

$$\mathbb{E}[F(w_k)] := \mathbb{E}_{\xi_1}\left[\mathbb{E}_{\xi_2}\left[\ldots\left[\mathbb{E}_{\xi_{k-1}}\left[F(w_k)\right]\right]\ldots\right]\right],$$

because w_k is completely determined by the independent random variables $\{\xi_1, \ldots \xi_{k-1}\}$. The goal of our analysis is to achieve convergence *in expectation* for the sequence $\{F(w_k)\}_{k\in\mathbb{N}}$ towards a local minimizer. As usual, the choice of the step size is crucial

[2] It is also possible to achieve almost sure convergence, but this requires way more work and a change of perspective, using martingale techniques; cf. [BCN18, Inset 4.1].

to obtain convergence. As we will see in the numerical analysis, due to the stochastic approach, SGD methods are more sensitive to the step size sequence than the (exact) gradient method.

Now we are able to prove our first convergence result.

Theorem 2.3.4 (Strongly convex objective function, fixed step size)**.** *Suppose that Assumption 2.2.1, 2.2.4 and 2.3.1 holds, and $F_{\inf} = F_*$. For the SGD algorithm with the fixed step size $\alpha_k = \bar{\alpha}$ for all $k \in \mathbb{N}$, satisfying*

$$0 < \bar{\alpha} \leq \frac{\mu}{LM_G}, \tag{2.13}$$

the following (contraction) inequality holds

$$\mathbb{E}\left[F(w_k) - F_*\right] \leq \frac{\bar{\alpha}LM}{2c\mu} + (1 - \bar{\alpha}c\mu)^{k-1}\left(F(w_1) - F_* - \frac{\bar{\alpha}LM}{2c\mu}\right) \overset{k\to\infty}{\longrightarrow} \frac{\bar{\alpha}LM}{2c\mu}. \tag{2.14}$$

Proof. This proof is strongly based on the proof of [BCN18, Theorem 4.6]. As this result one of the main theoretical statements, we include the proof with some additional steps.

Using Lemma 2.2.5, Proposition 2.3.2 and Eq. (2.13), we infer

$$\begin{aligned}
\mathbb{E}_{\xi_k}[F(w_{k+1}) - F(w_k)] &\overset{(2.10b)}{\leq} -\left(\mu - \frac{1}{2}\bar{\alpha}LM_G\right)\bar{\alpha}\left\|\nabla F(w_k)\right\|_2^2 + \frac{1}{2}\bar{\alpha}^2 LM \\
&\overset{(2.13)}{\leq} -\frac{1}{2}\mu\bar{\alpha}\left\|\nabla F(w_k)\right\|_2^2 + \frac{1}{2}\bar{\alpha}^2 LM \\
&\overset{(2.12)}{\leq} -\frac{1}{2}\mu\bar{\alpha}2c\left(F(w_k) - F_*\right) + \frac{1}{2}\bar{\alpha}^2 LM
\end{aligned}$$

for all $k \in \mathbb{N}$. We now subtract F_* from both sides, take the total expectation and shift $\mathbb{E}[F(w_k)]$ around, to get

$$\begin{aligned}
\mathbb{E}\left[F(w_{k+1}) - F_*\right] &\leq (1 - \mu\bar{\alpha}c)\mathbb{E}\left[F(w_k) - F_*\right] + \frac{1}{2}\bar{\alpha}^2 LM \\
\Rightarrow \mathbb{E}\left[F(w_{k+1}) - F_*\right] - \frac{\bar{\alpha}LM}{2c\mu} &\leq (1 - \mu\bar{\alpha}c)\mathbb{E}\left[F(w_k) - F_*\right] + \frac{1}{2}\bar{\alpha}^2 LM - \frac{\bar{\alpha}LM}{2c\mu} \\
&= (1 - \mu\bar{\alpha}c)\left(\mathbb{E}\left[F(w_k) - F_*\right] - \frac{\bar{\alpha}LM}{2c\mu}\right) \\
\Rightarrow \mathbb{E}\left[F(w_{k+1}) - F_*\right] - \frac{\bar{\alpha}LM}{2c\mu} &\leq (1 - \mu\bar{\alpha}c)\left(\mathbb{E}\left[F(w_k) - F_*\right] - \frac{\bar{\alpha}LM}{2c\mu}\right).
\end{aligned} \tag{2.15}$$

Eq. (2.15) is a contraction inequality, since we have

$$0 < \bar{\alpha} c \mu \overset{(2.13)}{\leq} \frac{c\mu^2}{LM_G} \overset{(2.9)}{\leq} \frac{c}{L}\frac{\mu^2}{\mu^2} \leq \frac{c}{L} \leq 1.$$

We obtain $c \leq L$ from Eq. (2.11). Now, we can apply the inequality (2.15) repeatedly, to obtain

$$\mathbb{E}\left[F(w_k) - F_*\right] \leq \frac{\bar{\alpha} L M}{2c\mu} + (1 - \bar{\alpha} c \mu)^{k-1} \left(F(w_1) - F_* - \frac{\bar{\alpha} L M}{2c\mu} \right).$$

The convergence result easily follows using the fact that the factor $1 - \bar{\alpha} c \mu$ is smaller than one and $F(w_1)$ is just a scalar, which can be factored out of the expected value, as it is independent of the joint distribution. □

Let us remark a few things, regarding the results of Theorem 2.3.4. First of all, assume $g(w_k, \xi_k)$ being an unbiased estimate of $\nabla F(w_k)$ and having no noise in $g(w_k, \xi_k)$, e.g. $\mu = 1$ and $M_G = 1$. In this case, the process of selecting a fixed step size reduces to $\alpha \in \left(0, \frac{1}{L}\right]$ – a classical step size for SGD. Furthermore, Theorem 2.3.4 indicates the important interplay between the step size and the bound of the variance. If we have no noise in the computation of the gradient, Theorem 2.3.4 states a linear convergence rate of SGD. But, if the computation is too noisy, the fixed step size only leads to a neighbourhood of the optimum.

Consequently, we will now consider SGD methods with descending step size sequences.

Remark 2.3.5 (SGD in practice)**.** Let us assume running SGD with an initial step size $\alpha_1 \in \left(0, \frac{\mu}{LM_G}\right]$ for the iterations $k_1 = 1$ to k_2, whereas w_{k_2} is the first iterate to satisfy

$$\mathbb{E}[F(w_{k_2}) - F_*] \leq 2 \cdot F_{\alpha_1},$$

with $F_\alpha := \frac{\alpha L M}{2c\mu}$. Recalling Theorem 2.3.4, we obtain that F_α is the limit of the expected optimality gap, if we run SGD with fixed step size $\bar{\alpha} = \alpha_1$. Now, we cut the step size in half, meaning $\alpha_2 := \frac{\alpha_1}{2}$. This results in a step size sequence $\{\alpha_{r+1}\} = \{\alpha_1 2^{-r}\}$ with an index sequence $\{k_r\}_{r \in \mathbb{N}}$. Now we can see that the sequence

$$\{F_{\alpha_r}\}_{r \in \mathbb{N}} = \left\{ \frac{\alpha_r L M}{2c\mu} \right\}_{r \in \mathbb{N}}$$

satisfies

$$\lim_{r\to\infty} F_{\alpha_r} = 0,$$

as the diminishing step size sequence converges to zero.

We will now analyse the behaviour of the optimality gap with respect to the new step size sequence $\{\alpha_r\}_{r\in\mathbb{N}}$ more closely. For all $r \in \mathbb{N}_{\geq 2}$, we have

$$\mathbb{E}[F(w_{k_{r+1}}) - F_*] \leq 2 \cdot F_{\alpha_r}. \tag{2.16}$$

Due to our choice of the step size sequence, we can assume that

$$\mathbb{E}[F(w_{k_r}) - F_*] \approx 2 \cdot F_{\alpha_{r-1}} = 4F_{\alpha_r}.$$

Next, we want to calculate the *effective rate* of step size decrease we need in order to obtain the inequality (2.16). Therefore, we can invoke Theorem 2.3.4 to calculate the difference $k_{r+1} - k_r$ as

$$\begin{aligned}\mathbb{E}[F(w_{k_{r+1}}) - F_*] &\overset{(2.14)}{\leq} \frac{\alpha_r LM}{2c\mu} + (1 - \alpha_r c\mu)^{k_{r+1}-k_r} \left(\mathbb{E}\left[F(w_{k_r}) - F_*\right] - \frac{\alpha_r LM}{2c\mu}\right) \overset{!}{\leq} 2F_{\alpha_r} \\ &\Rightarrow (1 - \alpha_r \mu c)^{k_{r+1}-k_r} \left(4F_{\alpha_r} - F_{\alpha_r}\right) \leq F_{\alpha_r}.\end{aligned}$$

As a consequence, we get

$$k_{r+1} - k_r \geq \frac{\log\left(\frac{1}{3}\right)}{\log\left(1 - \alpha_r \mu c\right)} \approx \frac{\log(3)}{\alpha_r \mu c} = \mathcal{O}(2^r),$$

where we used that the Taylor expansion of log at $x_0 = 1$ has the form

$$\log(x) = (x - 1) + \mathcal{O}(\|x - 1\|^2)$$

and

$$\log\left(\frac{1}{3}\right) = \log(1) - \log(3) = -\log(3)$$

holds. This result is essential because it states that cutting the step size in half, leads to a doubled number of iterations – a *sublinear rate of convergence*. An example of this behaviour is illustrated in Fig. 2.1.

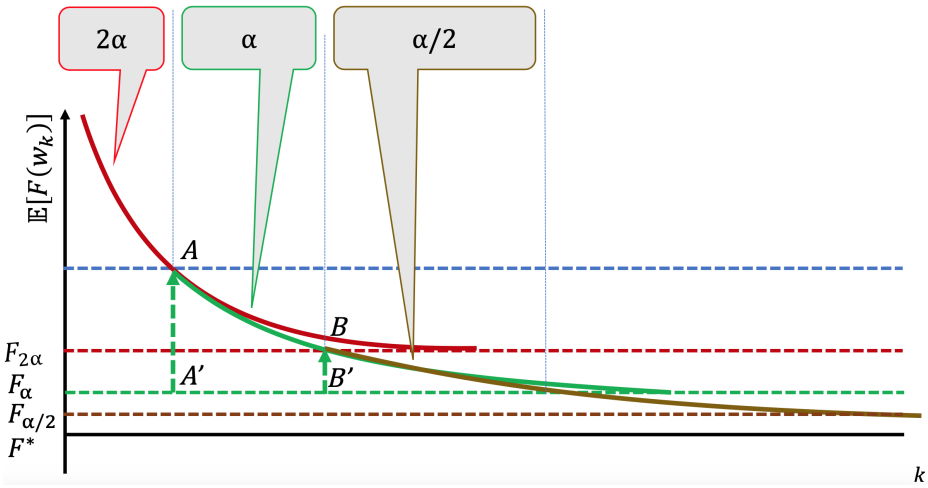

Figure 2.1.: Exemplary sublinear convergence behaviour, with a step size choice stated in Remark 2.3.5. One can see, the overall effective rate is at the order of $\mathcal{O}(\frac{1}{k})$. Source: [BCN18, Figure 4.1].

In the next theorem, we will analyse SGD methods with general diminishing step size sequences of the form

$$\sum_{k=1}^{\infty} \alpha_k = \infty \text{ and } \sum_{k=1}^{\infty} \alpha_k^2 < \infty. \tag{2.17}$$

The rate of convergence will stay the same, but one has more flexibility to choose the step size sequence.

Theorem 2.3.6 (Strongly convex objective function, diminishing step sizes)**.** *Under Assumptions 2.2.1, 2.2.4, 2.3.1 and $F_{\text{inf}} = F_*$, we consider the iterates $\{w_k\}_{k\in\mathbb{N}}$, generated by SGD with a step size sequence satisfying*

$$\{\alpha_k\}_{k\in\mathbb{N}} = \left\{\frac{\beta}{\gamma + k}\right\}_{k\in\mathbb{N}} \text{ with } \alpha_1 = \frac{\beta}{\gamma+1} \leq \frac{\mu}{LM_G}, \tag{2.18}$$

for some constants

$$\beta > \frac{1}{c\mu}, \gamma > 0.$$

Then, we get for all iterations $k \in \mathbb{N}$

$$\mathbb{E}[F(w_k) - F_*] \leq \frac{\nu}{\gamma + k}, \tag{2.19}$$

as an expected optimality gap, where

$$\nu := \max\left\{\frac{\beta^2 LM}{2(\beta c\mu - 1)}, (\gamma+1)\left(F(w_1) - F_*\right)\right\}. \tag{2.20}$$

Proof. This proof is strongly based on the proof of [BCN18, Theorem 4.7]. As this one of the main theoretical results, we include the proof with some additional steps. Using $\alpha_k = \frac{\beta}{\gamma+k}$, we get for all $k \in \mathbb{N}$

$$\alpha_k L M_G \leq \alpha_1 L M_G \leq \mu. \tag{$*$}$$

Now, we can use Lemma 2.2.5 and Proposition 2.3.2 and obtain

$$\begin{aligned}
\mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k)] &\overset{(2.10b)}{\leq} -\left(\mu - \frac{1}{2}\alpha_k L M_G\right)\alpha_k \left\|\nabla F(w_k)\right\|_2^2 + \frac{1}{2}\alpha_k^2 LM \\
&\overset{(*)}{\leq} -\frac{1}{2}\mu\alpha_k \left\|\nabla F(w_k)\right\|_2^2 + \frac{1}{2}\alpha_k^2 LM \\
&\overset{(2.12)}{\leq} -\alpha_k c\mu\left(F(w_k) - F(w_*)\right) + \frac{1}{2}\alpha_k^2 LM.
\end{aligned}$$

If we subtract F_* from both sides, take the expected value and rearrange the inequality, like in the proof of Theorem 2.3.4, it follows that

$$\mathbb{E}[F(w_{k+1})] - F_*] \leq (1 - \alpha_k c\mu)\mathbb{E}\left[F(w_k) - F(w_*)\right] + \frac{1}{2}\alpha_k^2 LM \tag{2.21}$$

holds. Now we can prove the statement of the theorem by induction.
$\underline{k = 1:}$ We have to show the inequality

$$\mathbb{E}[F(w_1)] - F_*] = F(w_1) - F_* \leq \frac{\nu}{\gamma+1},$$

which is a direct consequence of the definition of ν.
$\underline{k \to k+1:}$ We assume Eq. (2.19) holds for a $k \geq 1$, using (2.21), $\hat{k} := \gamma + k$ and $\alpha_k = \frac{\beta}{\gamma+k}$, we obtain

$$\begin{aligned}
\mathbb{E}[F(w_{k+1})] - F_*] &\overset{(2.21)}{\leq} (1 - \alpha_k c\mu)\mathbb{E}\left[F(w_k) - F(w_*)\right] + \frac{1}{2}\alpha_k^2 LM \\
&\leq \left(1 - \frac{\beta}{\hat{k}} c\mu\right)\frac{\nu}{\hat{k}} + \frac{\beta^2 LM}{2\hat{k}^2} = \left(\frac{\hat{k} - \beta c\mu}{\hat{k}^2}\right)\nu + \frac{\beta^2 LM}{2\hat{k}^2}
\end{aligned}$$

$$= \left(\frac{\hat{k}-1}{\hat{k}^2}\right)\nu - \left(\frac{\beta c\mu - 1}{\hat{k}^2}\right)\nu + \frac{\beta^2 LM}{2\hat{k}^2}$$
$$\leq \left(\frac{\hat{k}-1}{\hat{k}^2}\right)\nu. \tag{2.22}$$

The last inequality in (2.22) holds, because by the definition of ν we have

$$\nu \geq \frac{\beta^2 LM}{2(\beta c\mu - 1)},$$

which implies

$$\left(\frac{\beta c\mu - 1}{\hat{k}^2}\right)\nu \geq \frac{\beta^2 LM}{2\hat{k}^2}$$
$$\Rightarrow -\left(\frac{\beta c\mu - 1}{\hat{k}^2}\right)\nu + \frac{\beta^2 LM}{2\hat{k}^2} \leq 0.$$

Using $\hat{k}^2 \geq \hat{k}^2 - 1 = (\hat{k}-1)(\hat{k}+1)$, we get

$$\frac{\hat{k}-1}{\hat{k}^2} \leq \frac{1}{\hat{k}+1}. \tag{2.23}$$

Combing the inequalities (2.22) and (2.23) leads to

$$\mathbb{E}[F(w_{k+1})] - F_* \leq \frac{\nu}{\gamma + (k+1)},$$

which is our stated result. □

Finally, we want to interpret the results of Theorem 2.3.4 and Theorem 2.3.6.

Role of the strong convexity The parameter $c > 0$, gained by Assumption 2.3.1, occurs in both convergence theorems to achieve a contraction inequality. The impact of c on the step size differs between the strategies. While running SGD with a fixed step size, the interval to choose $\bar{\alpha}$ is independent of c, but for SG with a diminishing step size, the condition $\beta > \frac{1}{c\mu}$ is crucial to obtain a convergence rate of $\mathcal{O}\left(\frac{1}{k}\right)$.

Role of the initial guess In both theorems 2.3.4 and 2.3.6, the term $F(w_1) - F_*$ occurs. For a fixed step size, we get the initial gap with an exponentially decreasing factor.

For a diminishing step size sequence, one can spot the initial gap in the second term of the value ν. In [BCN18, p. 29] it is shown, that the first term of ν dominates the asymptotic behaviour. We will create a small example, which will illustrate this behaviour.

Assume running SGD with a fixed step size $\bar{\alpha}$ until step k_1, where

$$F(w_{k_1}) - F_* \leq \frac{\bar{\alpha} LM}{2c\mu},$$

as in Remark 2.3.5. Then choosing β, γ, as in Theorem 2.3.6 and $\bar{\alpha} = \alpha_1$, we get

$$(\gamma + 1)\mathbb{E}\left[F(w_{k_1}) - F_*\right] = \frac{\beta}{\alpha_1}\mathbb{E}\left[F(w_{k_1}) - F_*\right] \leq \frac{\beta}{\alpha_1}\frac{\alpha_1 LM}{2c\mu} = \frac{\beta LM}{2c\mu} < \frac{\beta^2 LM}{2(\beta c\mu - 1)}.$$

Therefore, the first term of ν in (2.20) asymptotically dominates.

We should always try to choose the initial step size as big as possible, e.g. $\alpha_1 = \frac{\mu}{LM_G}$. We obtain the best asymptotic behaviour, if the first value of ν is as small as possible, e.g. $\beta = \frac{2}{c\mu}$ (since only the first term is asymptotically relevant). In this case we get

$$\nu = \frac{\beta^2 LM}{2(\beta c\mu - 1)} = \frac{\left(\frac{2}{c\mu}\right)^2 LM}{2\left(\frac{2}{c\mu}c\mu - 1\right)} = \frac{2}{\mu^2}\frac{L}{c}\frac{M}{c}.$$

The last two ratios are essential, as they set the Lipschitz constant and the lower noise bound in proportion to the modulus of convexity.

2.4. Outlook: SGD for general objective functions

In this section we want to expand our analysis of SGD to non-convex objective functions. As we see in most practical application fields (e.g. the majority of machine learning models), the objective function is in general non-convex and therefore, further analysis of the SGD algorithm is needed. The analysis for non-convex objective functions is more challenging, as such functions can possess multiple local minima and other stationary points. The main result of this section will be the generalization of both of the previous theorems, but at the expense of the convergence factor in the case of a diminishing step size sequence. We want to make the same assumptions as before, except for the strong convexity of the objective function.

First, we want to take a look at the sequence of gradients of F when running SGD with fixed step sizes and hence try to generalize Lemma 2.2.3 for the non-convex case.

Theorem 2.4.1 (Non-convex objective function, fixed step size)**.** *Suppose that Assumptions 2.2.1 and 2.2.4 hold. Further, assume running SGD with a fixed step size, $\alpha_k = \bar{\alpha}$, for all $k \in \mathbb{N}$, satisfying the usual condition*

$$0 < \bar{\alpha} \leq \frac{\mu}{LM_G}. \tag{2.24}$$

Then, the expected sum-of-squares and average-squared gradients of the objective function F generated by the SGD iterates, satisfy the following inequalities for all $K \in \mathbb{N}$:

$$\mathbb{E}\left[\sum_{k=1}^{K} \|\nabla F(w_k)\|_2^2\right] \leq \frac{K\bar{\alpha}LM}{\mu} + \frac{2 \cdot (F(w_1) - F_{\text{inf}})}{\mu\bar{\alpha}} \tag{2.25a}$$

$$\text{and therefore } \mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K} \|\nabla F(w_k)\|_2^2\right] \leq \frac{\bar{\alpha}LM}{\mu} + \frac{2 \cdot (F(w_1) - F_{\text{inf}})}{K\mu\bar{\alpha}} \tag{2.25b}$$

$$\xrightarrow{K\to\infty} \frac{\bar{\alpha}LM}{\mu}.$$

Proof. See [BCN18, Theorem 4.8]. □

As in the result of Theorem 2.3.4, the right-hand side of the inequality contains the fixed step size $\bar{\alpha}$, a Lipschitz constant of the objective functions gradient, as well as the constants obtained by the assumptions regarding the first and second moments of the stochastic directions $g(w_k, \xi_k)$. The case of $M = 0$, meaning there is no noise or the noise reduces proportionally to $\|\nabla F(w_k)\|_2^2$, is also similar to Theorem 2.3.4.

Here, the sum of the squared gradients remains finite, implying $\{\|\nabla F(w_k)\|_2\} \to 0$ for $k \to \infty$.

In the situation of $M > 0$, Theorem 2.4.1 illustrates the interplay between the step size $\bar{\alpha}$ and the variance of the stochastic directions. While we lose the upper bound for the expected optimality gap due to the non-convexity of the objective function, inequality (2.25b) bounds the average of the objective functions gradient observed on $\{w_k\}_k$ during the first K iteration steps. The quantity on the right-hand side gets smaller when K increases, indicating that SGD spends increasingly more time in regions where the objective function has a (relatively) small gradient. This problem arises in most practical machine learning projects (a so-called *learning plateau*).

As we have already seen in the convex case, noise in the gradients delays further progress. One can decrease the upper bound for the average norm of the gradients to an arbitrarily small value by reducing the step size. However, Remark 2.3.5 illustrates that this tremendously reduces the speed of approaching this limit.

Now, we want to expand our result to the case with a diminishing step size sequence. We first state the next theorem and will later prove it as a corollary.

Theorem 2.4.2 (Non-convex objective function, diminishing step sizes)**.** *Under Assumptions 2.2.1 and 2.2.4, suppose running the SGD method with a step size sequence* $\{\alpha_k\}_{k\in\mathbb{N}}$*, satisfying*

$$\sum_{k=1}^{\infty} \alpha_k = \infty \text{ and } \sum_{k=1}^{\infty} \alpha_k^2 < \infty. \tag{2.26}$$

Then we obtain

$$\liminf_{k\to\infty} \mathbb{E}\left[\|\nabla F(w_k)\|_2^2\right] = 0 \tag{2.27}$$

for the sequence of gradients generated by SGD.

The proof of this theorem is a direct consequence of the next stated theorem. A "lim inf" result of this type is typical in the context of stochastic.

Theorem 2.4.3 (Non-convex objective function, diminishing step sizes)**.** *We again assume running the SGD method under Assumptions 2.2.1 and 2.2.4 with a diminishing step size sequence satisfying the conditions in* (2.26)*. Then, with* $A_K := \sum_{k=1}^{K} \alpha_k$*,*

for $K \in \mathbb{N}$, we obtain

$$\lim_{K\to\infty} \mathbb{E}\left[\sum_{k=1}^{K} \alpha_k \left\|\nabla F(w_k)\right\|_2^2\right] < \infty \tag{2.28a}$$

$$\textit{and therefore } \mathbb{E}\left[\frac{1}{A_k}\sum_{k=1}^{K} \alpha_k \left\|\nabla F(w_k)\right\|_2^2\right] \overset{K\to\infty}{\longrightarrow} 0. \tag{2.28b}$$

Proof. See [BCN18, Theorem 4.10]. □

In contrast to Theorem 2.4.1, the result of Theorem 2.4.3 states that the weighted average norm of the square gradients converges to zero – even under the presence of noise.

We can see now, why Theorem 2.4.2 is a direct consequence of Theorem 2.4.3. If Eq. (2.26) would not hold, it would contradict Theorem 2.4.3.

This also negates the fact that (2.28b) only specifies a property of a weighted average, as we can still conclude that the expected gradient norms cannot asymptotically stay far away from zero (cf. Theorem 2.4.2).

We can improve the "lim inf" convergence to a stronger convergence in probability, at the expense of only showing this property at randomly selected iterates of the objective functions gradient.

Corollary 2.4.4 (Convergence in probability for randomly selected iterates)**.** *We assume the conditions of Theorem 2.4.3 hold. For any $K \in \mathbb{N}$, let $k(K) \in \{1, \dots, K\}$ represent a randomly chosen index with probabilities proportional to $\{\alpha_k\}_{k=1}^{K}$. Then $\left\|\nabla F(w_{k(K)})\right\|_2^2 \to 0$ converges in probability as $K \to \infty$.*

Proof. The proof is a direct consequence of the Markov inequality (cf. Lemma A.2.2) and the result given in Eq. (2.28a). For arbitrary $\varepsilon > 0$, we can write

$$\begin{aligned}\mathbb{P}\left(\left\|\nabla F(w_{k(K)})\right\|_2 \geq \varepsilon\right) &= \mathbb{P}\left(\left\|\nabla F(w_{k(K)})\right\|_2^2 \geq \varepsilon^2\right) \\ &\overset{\text{Markov}}{\leq} \frac{\mathbb{E}\left[\mathbb{E}_k\left[\left\|\nabla F(w_{k(K)})\right\|_2^2\right]\right]}{\varepsilon^2} \overset{K\to\infty}{\longrightarrow} 0.\end{aligned}$$

This is exactly the definition of convergence in probability (cf. Definition A.2.1). □

Assuming additional regularity conditions, we can show a stronger convergence result.

Corollary 2.4.5. *We assume the conditions of Theorem 2.4.3 hold. Furthermore, let the objective function F be twice differentiable and we assume that the mapping $w \mapsto \|\nabla F(w)\|_2^2$ has a Lipschitz continuous derivative. Then,*

$$\lim_{k\to\infty} \mathbb{E}\left[\|\nabla F(w_k)\|_2^2\right] = 0.$$

Proof. See [BCN18, Appendix B]. □

In the next section we will analyse concrete implementations of various SGD variations and compare their behaviour using different objective functions.