Convergence Theorems for Gradient Descent

Robert M. Gower

May 2, 2022

Abstract

Here you will find a growing collection of simple proofs of the convergence of gradient and stochastic gradient descent type method on convex, strongly convex and smooth functions. Our focus is on "good proofs" that are also simple. Each section can also be consulted separately. Some noteworthy proofs here include a "good" convergence proof of SGD with momentum.

Disclaimer: Theses notes are not proper review of the literature, and we welcome earlier pointers to missing bibliography, of which we are sure is missing. Further disclaimer: These notes do not compare to a good book or well prepared lecture notes. If you are new to convex optimization I highly recommend reading instead the first few chapters of the books [4] and [1].

Contents

0.1 What this is not

We only focus on the simplest gradient methods. We do not touch upon second order methods, or optimization at large. Our formalism in convex analysis is kept at a minimal (and sometime informal). When discussing stochastic gradient descent, we use the finite-sum notation for simplicity. Yet, much (if not all) our results for stochastic gradient hold for a continuous measure and minimizing an expectation.

1 Assumptions and Lemmas

1.1 Convexity

Let $f : \mathbb{R}^d \mapsto \mathbb{R}$ be a function. We say that f is convex if

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y), \quad \forall x, y \in \mathbb{R}^d,\ t \in [0,\, 1]. \tag{1}$$

Lemma 1.1. A function $f : \mathbb{R}^d \mapsto \mathbb{R}$ is convex if and only if the univariate function $\phi : \mathbb{R} \mapsto \mathbb{R}$ given by $\phi : t \mapsto f(x_0 + tv)$ is convex for all $x_0, v \in \mathbb{R}^d$. Let $x_0, v \in \mathbb{R}^d$, $\operatorname{dom} \phi = \{t \in \mathbb{R} : x_0 + tv \in \operatorname{dom} f\}$.

Proof: Homework! Try, and if you fail, ask me.

Lemma 1.2. Let f be twice continuously differentiable. Then f is convex if either of the following hold

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle\,, \quad \forall x, y \in \mathbb{R}^d, \tag{2}$$

or

$$\left\langle \nabla^2 f(x) v, v \right\rangle \geq 0, \quad \forall x, v \in \mathbb{R}^d. \tag{3}$$

Proof: We will prove this by showing that (1) $\Leftrightarrow$ (2) $\Leftrightarrow$ (3).

(1) $\Rightarrow$ (2) We can deduce (2) from (1) by dividing by t and re-arranging

$$\frac{f(y + t(x-y)) - f(y)}{t} \leq f(x) - f(y).$$

Now taking the limit $t \to 0$ gives

$$\langle \nabla f(y), x - y \rangle \leq f(x) - f(y).$$

(2) ⇒ (1) Let $x_t = tx + (1-t)y$. From (2) we have that

$$\begin{array}{rclcl} f(x) & \geq & f(x_t) + \langle \nabla f(x_t), x - x_t \rangle & = & f(x_t) - (1-t) \langle \nabla f(x_t), y - x \rangle \\ f(y) & \geq & f(x_t) + \langle \nabla f(x_t), y - x_t \rangle & = & f(y) + t \langle \nabla f(x_t), y - x \rangle \end{array}$$

Multiplying the first inequality by t and the second inequality by $(1-t)$ and adding the result together gives

$$tf(x) + (1-t)f(y) \geq f(x_t)$$

which is equivalent to (1).

(2) ⇒ (3) Let us prove this in dimension 1 and then generalize it using Lemma 1.1.

Let $x, y \in \operatorname{dom} f \subseteq \mathbb{R}$ $(d = 1)$, with $x < y$. We have

$$f(y) \geq f(x) + f'(x)(y - x) \tag{4}$$

$$f(x) \geq f(y) + f'(y)(x - y) \tag{5}$$

The inequalities (4) and (5) imply that

$$f'(x)(y - x) \leq f(y) - f(x) \leq f'(y)(y - x)$$

Diving by $(y-x)^2 > 0$ and noticing that the same inequality holds for $y < x$ gives

$$\frac{f'(y) - f'(x)}{y - x} \geq 0 \quad \forall x, y \in \mathbb{R} : x \neq y$$

As we let $y \to x$, we get $f''(x) \geq 0$ for all $x \in \operatorname{dom} f$.

Now let us establish (2) ⇒ (3) in general dimension d. We recall that convexity of f is equivalent convexity along all lines (see Lemma 1.1). So, if (2) is true, then we have that f is convex (1). Then, by Lemma 1.1 we have that $\phi : t \mapsto f(x_0 + tv)$ is convex for all ("proper") $x_0, v \in \mathbb{R}^d$. According to the proof above in dimension 1, the latter implies that

$$\phi''(t) = \left\langle \nabla^2 f(x_0 + tv)v, v \right\rangle \geq 0 \quad \forall x_0, v \in \mathbb{R}^d, \forall t \in \mathbb{R} : x_0 + tv \in \operatorname{dom} f$$

Hence, if f is convex, then (2) is true, which implies finally $\nabla^2 f(x) \succeq 0$ for all $x \in \operatorname{dom} f$.

(3) ⇒ (2) Using Taylor expansion we have that

$$\begin{array}{rcl} f(x) & = & f(y) + \langle \nabla f(y), x - y \rangle + \displaystyle\int_{b=0}^{1} \int_{s=0}^{b} \left\langle \nabla^2 f(y + b(x-y))(x-y), (x-y) \right\rangle dbds. \\ & \overset{(3)}{\geq} & f(y) + \langle \nabla f(y), x - y \rangle . \quad \square \end{array} \tag{6}$$

□

An analogous property to (2) holds even when the function is not differentiable. Indeed for every convex function, we say that $g \in \mathbb{R}^d$ subgradient is a subdifferential at x if

$$f(y) \geq f(x) + \langle g, y - x \rangle\,, \quad \forall y. \tag{7}$$

We refer to the set of subgradients as the subdifferential $\partial f(x)$, that is

$$\partial f(x) \overset{\text{def}}{=} \{g \;:\; f(y) \geq f(x) + \langle g, y - x \rangle\,, \quad \forall y\}. \tag{8}$$

1.2 Smoothness

A differential function f is said to be L–smooth if its gradients are Lipschitz continuous, that is

$$\|\nabla f(x) - \nabla f(y)\| \quad \leq \quad L\|x - y\|. \tag{9}$$

Lemma 1.3. Let $f : \mathbb{R}^d \mapsto \mathbb{R}$ be a twice differentiable function. If f is L–smooth then the following holds

$$\begin{aligned} \left\langle \nabla^2 f(x) v, v \right\rangle \quad &\leq \quad L\|v\|_2^2, \quad \forall x, v \in \mathbb{R}^d, &(10)\\ f(y) \quad &\leq \quad f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|_2^2. &(11) \end{aligned}$$

Proof:

(9) ⇒ (10) If f is twice differentiable then we have, by using first order expansion

$$\nabla f(x) - \nabla f(x + \alpha v) = \int_{t=0}^{\alpha} \nabla^2 f(x + tv) v dt. \tag{12}$$

Taking the inner product with v gives

$$\begin{aligned} \int_{t=0}^{\alpha} \left\langle \nabla^2 f(x + tv) v, v \right\rangle dt \quad &= \quad \langle \nabla f(x) - \nabla f(x + \alpha v), v \rangle \\ &\leq \quad \|\nabla f(x) - \nabla f(x + \alpha v)\|\|v\| \\ &\overset{(9)}{\leq} \quad L\alpha\|v\|^2. \end{aligned}$$

Dividing by α and taking the limit of $\alpha \to 0$ gives

$$\lim_{\alpha \to 0} \frac{1}{\alpha} \int_{t=0}^{\alpha} \left\langle \nabla^2 f(x + tv) v, v \right\rangle dt \;= \left\langle \nabla^2 f(x) v, v \right\rangle \quad \leq \quad L\|v\|^2.$$

(9) ⇒ (11) Using the Taylor expansion of $f(x)$ we have that

$$\begin{aligned} f(x) \quad &= \quad f(y) + \int_{b=0}^{1} \langle \nabla f(y + b(x - y)) \rangle \, db. \\ &= \quad f(y) + \langle \nabla f(y), x - y \rangle + \int_{b=0}^{1} \langle \nabla f(y + b(x - y)) - \nabla f(y), (x - y) \rangle \, db. \\ &\leq \quad f(y) + \langle \nabla f(y), x - y \rangle + \int_{b=0}^{1} \|\nabla f(y + b(x - y)) - \nabla f(y)\|\|x - y\| db \\ &\overset{(9)}{\leq} \quad f(y) + \langle \nabla f(y), x - y \rangle + L \int_{b=0}^{1} b\|x - y\|^2 db \;=\; (11). \end{aligned}$$

Some direct consequences of the smoothness are given in the following lemma.

Lemma 1.4. If f is L–smooth then

$$f(x - \tfrac{1}{L}\nabla f(x)) - f(x) \leq -\frac{1}{2L}\|\nabla f(x)\|_2^2, \tag{13}$$

and

$$f(x^*) - f(x) \leq -\frac{1}{2L}\|\nabla f(x)\|_2^2, \tag{14}$$

hold for all $x \in \mathbb{R}^d$.

Proof: The first inequality (13) follows by inserting $y = x - \frac{1}{L}\nabla f(x)$ in the definition of smoothness (9) since

$$\begin{aligned} f(x - \tfrac{1}{L}\nabla f(x)) &\leq f(x) - \tfrac{1}{L}\langle \nabla f(x), \nabla f(x)\rangle + \frac{L}{2}\|\tfrac{1}{L}\nabla f(x)\|_2^2 \\ &= f(x) - \frac{1}{2L}\|\nabla f(x)\|_2^2. \end{aligned}$$

Furthermore, by using (13) combined with $f(x^*) \leq f(y) \quad \forall y$, we get (14). Indeed since

$$f(x^*) - f(x) \leq f(x - \tfrac{1}{L}\nabla f(x)) - f(x) \leq -\frac{1}{2L}\|\nabla f(x)\|_2^2. \quad \square \tag{15}$$

1.3 Smooth and Convex

There are many problems in optimization where the function is both smooth and convex. Furthermore, such a combination results in some interesting consequences and Lemmas. Lemmas that we will then use to prove convergence of the Gradient method.

Lemma 1.5. If $f(x)$ is convex and L–smooth then

$$f(y) - f(x) \leq \langle \nabla f(y), y - x\rangle - \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|_2^2. \tag{16}$$

$$\langle \nabla f(y) - \nabla f(x), y - x\rangle \geq \frac{1}{L}\|\nabla f(x) - \nabla f(y)\| \quad \text{(Co-coercivity).} \tag{17}$$

Proof: To prove (16), it follows that

$$\begin{aligned} f(y) - f(x) &= f(y) - f(z) + f(z) - f(x) \\ &\overset{(2)+(11)}{\leq} \langle \nabla f(y), y - z\rangle + \langle \nabla f(x), z - x\rangle + \frac{L}{2}\|z - x\|_2^2. \end{aligned}$$

To get the tightest upper bound on the right hand side, we can minimize the right hand side in z, which gives

$$z = x - \frac{1}{L}(\nabla f(x) - \nabla f(y)). \tag{18}$$

Substituting this in gives

$$
\begin{aligned}
f(y)-f(x) &= \left\langle \nabla f(y), y-x+\frac{1}{L}(\nabla f(x)-\nabla f(y)) \right\rangle - \frac{1}{L}\left\langle \nabla f(x), \nabla f(x)-\nabla f(y)\right\rangle + \frac{1}{2L}\|\nabla f(x)-\nabla f(y)\|_2^2 \\
&= \langle \nabla f(y), y-x\rangle - \frac{1}{L}\|\nabla f(x)-\nabla f(y)\|_2^2 + \frac{1}{2L}\|\nabla f(x)-\nabla f(y)\|_2^2 \qquad (19) \\
&= \langle \nabla f(y), y-x\rangle - \frac{1}{2L}\|\nabla f(x)-\nabla f(y)\|_2^2 \qquad (20)
\end{aligned}
$$

Finally (17) follows from applying (16) once

$$f(y)-f(x) \leq \langle \nabla f(y), y-x\rangle - \frac{1}{2L}\|\nabla f(y)-\nabla f(x)\|_2^2,$$

then interchanging the roles of x and y to get

$$f(x)-f(y) \leq \langle \nabla f(x), x-y\rangle - \frac{1}{2L}\|\nabla f(y)-\nabla f(x)\|_2^2.$$

Finally adding together the two above inequalities gives

$$0 \leq \langle \nabla f(y)-\nabla f(x), y-x\rangle - \frac{1}{L}\|\nabla f(y)-\nabla f(x)\|_2^2. \quad \square$$

1.4 Strong convexity

We can "strengthen" the notion of convexity by defining μ–strong convexity, that is

$$f(y) \geq f(x) + \langle \nabla f(x), y-x\rangle + \frac{\mu}{2}\|y-x\|_2^2, \quad \forall x,y \in \mathbb{R}^d. \qquad (21)$$

.

Lemma 1.6. Let f be twice continuously differentiable. The following is equivalent to f being μ–strongly convex

$$\left\langle \nabla^2 f(x)v, v\right\rangle \geq \mu\|v\|_2^2. \qquad (22)$$

Proof:

The following inequality (23) is of such importance in optimization that is merits its own name.

Lemma 1.7. If f is μ–strongly convex then it also satisfies the *Polyak–Lojasiewicz* condition, that is

$$\|\nabla f(x)\|_2^2 \geq 2\mu(f(x)-f(x^*)). \qquad (23)$$

Proof: Multiplying (21) by minus one and substituting $y = x^*$ we have that

$$
\begin{aligned}
f(x)-f(x^*) &\leq \langle \nabla f(x), x-x^*\rangle - \frac{\mu}{2}\|x^*-x\|_2^2 \\
&= -\frac{1}{2}\|\sqrt{\mu}(x-x^*) - \frac{1}{\sqrt{\mu}}\nabla f(x)\|_2^2 + \frac{1}{2\mu}\|\nabla f(x)\|_2^2 \\
&\leq \frac{1}{2\mu}\|\nabla f(x)\|_2^2.
\end{aligned}
$$

2 Gradient Descent

Consider the problem

$$x^* = \arg\min_{x\in\mathbb{R}^d} f(x), \tag{24}$$

and the following gradient method

$$x^{t+1} = x^t - \alpha \nabla f(x^t), \tag{25}$$

where f is L–smooth. We will now prove that the iterates (25) converge. In Theorem 2.1 we will prove sublinear convergence under the assumption that f is convex. In Theorem 2.2 we will prove linear convergence (a stronger form of convergence) under the assumption that f is μ–strongly convex.

2.1 Convergence for convex and smooth functions

Theorem 2.1. Let f be convex and L–smooth and let x^t for $t = 1, \ldots, n$ be the sequence of iterates generated by the gradient method (25). It follows that

$$f(x^n) - f(x^*) \leq \frac{2L\|x^1 - x^*\|^2}{n-1}. \tag{26}$$

Proof: Let f be convex and L–smooth. It follows that

$$\begin{aligned}
\|x^{t+1} - x^*\|_2^2 &= \|x^t - x^* - \tfrac{1}{L}\nabla f(x^t)\|_2^2 \\
&= \|x^t - x^*\|_2^2 - 2\tfrac{1}{L}\left\langle x^t - x^*, \nabla f(x^t)\right\rangle + \tfrac{1}{L^2}\|\nabla f(x^t)\|_2^2 \\
&\overset{(17)}{\leq} \|x^t - x^*\|_2^2 - \tfrac{1}{L^2}\|\nabla f(x^t)\|_2^2.
\end{aligned} \tag{27}$$

Thus $\|x^t - x^*\|_2^2$ is a decreasing sequence in t, and thus consequently

$$\|x^t - x^*\|_2 \leq \|x^1 - x^*\|_2. \tag{28}$$

Calling upon (13) and subtracting $f(x^*)$ from both sides gives

$$f(x^{t+1}) - f(x^*) \;\leq\; f(x^t) - f(x^*) - \frac{1}{2L}\|\nabla f(x^t)\|_2^2. \tag{29}$$

Applying convexity we have that

$$\begin{aligned}
f(x^t) - f(x^*) &\leq \left\langle \nabla f(x^t), x^t - x^*\right\rangle \\
&\leq \|\nabla f(x^t)\|_2\|x^t - x^*\| \overset{(28)}{\leq} \|\nabla f(x^t)\|_2\|x^1 - x^*\|.
\end{aligned} \tag{30}$$

Isolating $\|\nabla f(x^t)\|_2$ in the above and inserting in (29) gives

$$f(x^{t+1}) - f(x^*) \overset{(29)+(30)}{\leq} f(x^t) - f(x^*) - \underbrace{\frac{1}{2L}\frac{1}{\|x^1 - x^*\|^2}}_{\beta}(f(x^t) - f(x^*))^2 \tag{31}$$

Let $\delta_t = f(x^t) - f(x^*)$. Since $\delta_{t+1} \leq \delta_t$, and by manipulating (31) we have that

$$\delta_{t+1} \leq \delta_t - \beta\delta_t^2 \overset{\times\frac{1}{\delta_t\delta_{t+1}}}{\Leftrightarrow} \beta\frac{\delta_t}{\delta_{t+1}} \leq \frac{1}{\delta_{t+1}} - \frac{1}{\delta_t} \overset{\delta_{t+1}\leq\delta_t}{\Leftrightarrow} \beta \leq \frac{1}{\delta_{t+1}} - \frac{1}{\delta_t}.$$

Summing up both sides over $t = 1, \ldots, n-1$ and using telescopic cancellation we have that

$$(n-1)\beta \leq \frac{1}{\delta_n} - \frac{1}{\delta_1} \leq \frac{1}{\delta_n}. \quad \square$$

2.2 Convergence of the gradient norm for non-convex and smooth

As a side note, we can use the previous proof to show that the iterates converge to a stationary point even when f is not convex. Indeed, re-arranging (27) gives

$$\|\nabla f(x^t)\|_2^2 \quad \leq \quad L^2\|x^t - x^*\|_2^2 - L^2\|x^{t+1} - x^*\|_2^2.$$

Summing up from $t = 1, \ldots, T$ and dividing by T gives

$$\begin{aligned} \frac{1}{T}\sum_{t=1}^{T}\|\nabla f(x^t)\|_2^2 \quad &\leq \quad \frac{L^2}{T}\left(\|x^1 - x^*\|_2^2 - \|x^{T+1} - x^*\|_2^2\right) \\ &\leq \quad \frac{L^2}{T}\|x^1 - x^*\|_2^2. \end{aligned} \tag{32}$$

Consequently

$$\min_{t=1,\ldots,T}\|\nabla f(x^t)\|_2^2 \;\leq\; \frac{1}{T}\sum_{t=1}^{T}\|\nabla f(x^t)\|_2^2 \;\leq\; \frac{L^2}{T}\|x^1 - x^*\|_2^2. \tag{33}$$

Thus we know that, by recording the iterate x^t with the least gradient norm, this gradient norm will converge sublinearly to zero, and thus x^t converges to a stationary point.

2.3 Convergence for strongly convex and smooth functions

Now we prove some bounds that hold for strongly convex and smooth functions.

Theorem 2.2. Let f be L–smooth and μ–strongly convex. From a given $x_0 \in \mathbb{R}^d$ and $\frac{1}{L} \geq \alpha > 0$, the iterates

$$x^{t+1} = x^t - \alpha\nabla f(x^t), \tag{34}$$

converge according to

$$\|x^{t+1} - x^*\|_2^2 \leq (1 - \alpha\mu)^{t+1}\|x^0 - x^*\|_2^2. \tag{35}$$

In particular, or $\alpha = \frac{1}{L}$ the iterates (25) enjoy a linear convergence with a rate of μ/L.

Proof: From (25) we have that

$$\begin{aligned}
\|x^{t+1} - x^*\|_2^2 &= \|x^t - x^* - \alpha \nabla f(x^t)\|_2^2 \\
&= \|x^t - x^*\|_2^2 - 2\alpha \left\langle \nabla f(x^t), x^t - x^* \right\rangle + \alpha^2 \|\nabla f(x^t)\|_2^2 \\
&\overset{(21)}{\leq} (1 - \alpha\mu)\|x^t - x^*\|_2^2 - 2\alpha(f(x^t) - f(x^*)) + \alpha^2 \|\nabla f(x^t)\|_2^2 \\
&\overset{(14)}{\leq} (1 - \alpha\mu)\|x^t - x^*\|_2^2 - 2\alpha(f(x^t) - f(x^*)) + 2\alpha^2 L(f(x^t) - f(x^*) \\
&= (1 - \alpha\mu)\|x^t - x^*\|_2^2 - 2\alpha(1 - \alpha L)(f(x^t) - f(x^*)). \qquad (36)
\end{aligned}$$

Since $\frac{1}{L} \geq \alpha$ we have that $-2\alpha(1 - \alpha L)$ is negative, and thus can be safely dropped to give

$$\|x^{t+1} - x^*\|_2^2 \leq (1 - \alpha\mu)\|x^t - x^*\|_2^2.$$

It now remains to unroll the recurrence. □

3 Proximal gradient descent

We now add a bit more structure and consider the minimization of the sum of two objective functions

$$x^* = \arg\min_{w \in \mathbb{R}^d} F(w) \overset{\text{def}}{=} f(w) + R(w), \qquad (37)$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is a loss function and $R : \mathbb{R}^d \to \mathbb{R}$ a regularizor.

3.1 Subgradients

First we need the notion of subgradients to extend the definition of convex to non-smooth functions.

Definition 3.1. Let $F : \mathbb{R}^d \to \mathbb{R}$. We say that $g \in \mathbb{R}^d$ is a subgradient of F at $x \in \mathbb{R}^d$ if

$$F(w) \geq F(x) + \langle g, w - x \rangle, \quad \forall w \in \mathbb{R}^d. \qquad (38)$$

We use $\partial F(x) \subset \mathbb{R}^d$ to denote the set of subgradients

Convex functions are guaranteed to have subgradients. What is more, Subgradient can be used to characterize convex functions.

Lemma 3.2. Let $F : \mathbb{R}^d \to R$. If for every $x \in \mathbb{R}^d$ we have that $\partial f(x)$ is non-empty, then F is convex. Furthermore, if F is convex then $\partial F(x)$ is non-empty for every $x \in \mathbb{R}^d$.

3.2 Properties of the Proximal Operator

To exploit this structure we use the proximal operator.

Definition 3.3. For every convex function $g : \mathbb{R}^d \to \mathbb{R}$ and vector $v \in \mathbb{R}^d$ we define the proximal operator of f applied to v as

$$\mathrm{prox}_g(v) := \arg\min_w \frac{1}{2}||w - v||_2^2 + g(w) \tag{39}$$

The proximal operator is well defined because, since g is convex the sum $\frac{1}{2}||w - v||_2^2 + g(w)$ is strongly convex in w. Thus there exists only one solution.

One way to characterize the proximal operator is using the subgradient

Lemma 3.4. Let $p_v = \mathrm{prox}_g(v)$. It follows that

$$v - p_v \in \partial g(p_v). \tag{40}$$

Furthermore using the above together with the definition of subgradient (38) we have that

$$h(w) \geq h(p_v) + \langle v - p_v, w - p_v \rangle . \tag{41}$$

Proof: This follows by taking the derivative of (39) in w and setting to zero to find the solution, which gives

$$0 \in w - v + \partial g(w).$$

If w solves the above, it is necessarily the argmin, thus

$$0 \in p_v - v + \partial g(p_v).$$

Rearranging gives (40). □

The proximal operator also behaves a bit like a projection, in that it is also non-expansive.

Lemma 3.5 (Non-expansiveness)**.** For every convex function $g : \mathbb{R}^d \to \mathbb{R}$ and vectors $v, u \in \mathbb{R}^d$ we have that

$$\|p_u - p_v\|^2 \leq \langle u - v, p_u - p_v \rangle . \tag{42}$$

Furthermore, using Cauchy-Schwarz we have that

$$\|\mathrm{prox}_g(v) - \mathrm{prox}_g(u)\|_2 \leq \|v - u\|_2 \tag{43}$$

Proof: Let $p_v \overset{\text{def}}{=} \mathrm{prox}_g(v)$ and $p_u \overset{\text{def}}{=} \mathrm{prox}_g(u)$. Using the proximal gradient characterization (40) (with $w = p_u$ and then $w = p_v$) gives

$$\begin{aligned} h(p_u) &\geq h(p_v) + \langle v - p_v, p_u - p_v \rangle && (44) \\ h(p_v) &\geq h(p_u) + \langle u - p_u, p_v - p_u \rangle . && (45) \end{aligned}$$

Adding together the above two inequalities gives

$$\langle v - u - p_v + p_u, p_u - p_v \rangle \leq 0.$$

Expanding the left argument of the inner product gives

$$\|p_u - p_v\|^2 \leq \langle u - v, p_u - p_v \rangle \,,$$

which proves (42). Now using the Cauchy-Schwartz inequality gives

$$\|p_u - p_v\|^2 \leq \langle u - v, p_u - p_v \rangle \leq \|u - v\| \|p_u - p_v\|.$$

Dividing through by $\|p_u - p_v\|$ (assuming this is non-zero otherwise (43) holds trivially) we have (43).

3.3 Proximal gradient descent

Consider the method

$$x^{t+1} = \mathrm{prox}_{\frac{\lambda}{L}R}(x^t - \frac{1}{L}\nabla f(x^t))). \tag{46}$$

First note that

Lemma 3.6 (Fixed point viewpoint)**.**

$$x^* = \mathrm{prox}_{\lambda\gamma R}\left(x^* - \gamma \nabla f(x^*)\right) \tag{47}$$

Proof: Homework! Also in the slides.

Now we give an example of how to prove convergence of the proximal gradient descent method using the fixed point viewpoint (47) followed by the non-expansiveness property (43).

Theorem 3.7. Let $R : \mathbb{R}^d \to \mathbb{R}$ be a convex function and let $L : \mathbb{R}^d \to \mathbb{R}$ be μ–strongly convex, $L_{\max}$–smooth and twice continuously differentiable. It follows that the iterates (46) converges according to

$$\|x^{t+1} - x^*\|_2 \quad \leq \quad \left(1 - \frac{\mu}{L_{\max}}\right) \|x^t - x^*\|_2. \tag{48}$$

Proof: Note that

$$\begin{aligned}
\|x^{t+1} - x^*\|_2 &\overset{(46)}{=} \|\mathrm{prox}_{\frac{\lambda}{L}R}(x^t - \frac{1}{L}\nabla f(x^t))) - x^*\|_2 \\
&\overset{(47)}{=} \|\mathrm{prox}_{\frac{\lambda}{L}R}(x^t - \frac{1}{L}\nabla f(x^t))) - \mathrm{prox}_{\frac{\lambda}{L_{\max}}R}\left(x^* - \frac{1}{L_{\max}}\nabla f(x^*)\right)\|_2 \\
&\overset{(43)}{\leq} \|(x^t - \frac{1}{L}\nabla f(x^t))) - \left(x^* - \frac{1}{L_{\max}}\nabla f(x^*)\right)\|_2 \\
&= \|x^t - x^* - \frac{1}{L}\left(\nabla f(x^t)) - \nabla f(x^*)\right)\|_2 \qquad (49)
\end{aligned}$$

Now using Taylor's expansion we have that

$$\begin{aligned}
x^t - x^* - \frac{1}{L}\left(\nabla f(x^t)) - \nabla f(x^*)\right) &= x^t - x^* - \frac{1}{L}\int_{s=0}^{1} \nabla^2 f(x^t + s(x^* - x^t)))(x^t - x^*)ds \\
&= \left(I - \frac{1}{L}\int_{s=0}^{1} \nabla^2 f(x^t + s(x^* - x^t)))\right)(x^t - x^*)ds.
\end{aligned}$$

For shorthand let $w_s^t \overset{\text{def}}{=} x^t + s(x^* - x^t)$. Taking norms in the above gives

$$\|x^t - x^* - \frac{1}{L}\left(\nabla f(x^t)) - \nabla f(x^*)\right)\|_2 \quad \leq \quad \|I - \frac{1}{L}\int_{s=0}^1 \nabla^2 f(w_s^t)ds\|\|x^t - x^*\|_2.$$

Let λ_i be the eigenvalues of $\int_{s=0}^1 \nabla^2 f(w_s^t)ds$. Thus we have that

$$\|I - \frac{1}{L}\int_{s=0}^1 \nabla^2 f(w_s^t)ds\| = \max_{i=1,\ldots,d} |1 - \lambda_i/L| \leq (1 - \mu/L),$$

which follows since L is μ–strongly convex and L–smooth . Thus finally combining this with (49) gives

$$\|x^{t+1} - x^*\|_2 \quad \leq \quad \left(1 - \frac{\mu}{L}\right)\|x^t - x^*\|_2.$$

4 Stochastic Gradient Descent

Now we assume that our objective function has a sum of term structure given by

$$\min f(x) \quad \overset{\text{def}}{=} \quad \frac{1}{n}\sum_{i=1}^n f_i(x). \tag{50}$$

For this section let $\mathcal{X}^*$ be the set of minimizers of (50) which we assume in non-empty.

Here we will consider the convergence of SGD (Stochastic Gradient Descent)

$$\begin{aligned} & i \in \{1, \ldots n\} & & \text{Sampled with probability } \frac{1}{n} & (51)\\ & x^{t+1} = x^t - \alpha_t \nabla f_i(x^t), & & & (52) \end{aligned}$$

where $\alpha_t > 0$ are a sequence of step sizes (also known as learning rates). Because we uniformly sample the index i over all indices $\{1, \ldots, n\}$ we have that

$$\mathbb{E}\left[\nabla f_i(x^t)\right] = \sum_{i=1}^n \frac{1}{n}\nabla f_i(x^t) = \nabla f(x^t). \tag{53}$$

We will assume throughout this section that each $f_i(x)$ is smooth and convex.

Assumption 4.1. Each f_i is convex, that is

$$f_i(x) \geq f_i(y) + \langle \nabla f_i(y), x - y\rangle, \quad \text{for all } x, y \in \mathbb{R}^d. \tag{54}$$

Furthermore there exists $L_{\max} \geq 0$ such that f_i is $L_{\max}$–smooth.

All of our proofs also rely on the follow definition of gradient noise.

Definition 4.2. We refer to σ^2 as

$$\sigma^2 \overset{\text{def}}{=} \max_{x^* \in \mathcal{X}} \frac{1}{n} \sum_{i=1}^{2} \|\nabla f_i(x^*)\|^2. \tag{55}$$

It is this gradient noise σ^2 that prevents SGD from converging when using constant stepsizes, as we will see later on.

4.1 Properties of Smooth and Convex

As a direct consequence of Assumption 4.1 we have the following bound

Lemma 4.3. Let Assumption 4.1 hold and let $x^* \in \mathbb{R}^d$ be a minimizer of (50). It follows that

$$\frac{1}{n} \sum_{i=1}^{n} \|\nabla f_i(x) - \nabla f_i(x^*)\|^2 \;\leq\; L_{\max}(f(x) - f(x^*)). \tag{56}$$

Consequently for the iterates x^t of SGD we have that

$$\mathbb{E}_t\left[\|\nabla f_i(x^t) - \nabla f_i(x^*)\|^2\right] \;\leq\; L_{\max}(f(x^t) - f(x^*)), \tag{57}$$

where $\mathbb{E}_t\left[x\right] = \mathbb{E}\left[x \mid x^t\right]$ is expectation conditioned on x^t.

Proof: Using (16) in Lemma 1.5 applied to f_i we have, after re-arranging, that

$$\|\nabla f_i(x^*) - \nabla f_i(x)\|_2^2 \;\;\leq\;\; 2L_{\max}(f_i(x) - f_i(x^*) + \langle \nabla f_i(x^*), x^* - x\rangle). \tag{58}$$

Using (53) we have that $\frac{1}{n}\sum_{i=1}^n \nabla f_i(x^*) = \nabla f(x^*) = 0$, which follows since x^* is a minima and thus it is also a stationary point. Consequently summing up (58) over $i = 1, \ldots, n$ gives

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^{n} \|\nabla f_i(x^*) - \nabla f_i(x)\|_2^2 \;&\leq\; 2L_{\max}(f(x) - f(x^*) + \langle \nabla f(x^*), x^* - x\rangle) \\ &=\; 2L_{\max}(f(x) - f(x^*)), \end{aligned} \tag{59}$$

where we also used $\frac{1}{n}\sum_{i=1}^n f_i(x) = f(x)$.

Because of the above lemma, we can now derive a convenient bound on the norm of the stochastic gradient.

Using the definition of gradient noise, we have the following.

Lemma 4.4. Let Assumption 4.1 hold. It follows that

$$\mathbb{E}_t\left[\|\nabla f_i(x^t)\|^2\right] \leq 4L_{\max}(f(x) - f^*) + 2\sigma^2, \tag{60}$$

Proof: Using

$$\|\nabla f_i(x^t)\|^2 \leq 2\|\nabla f_i(x^t) - g(x^*)\|^2 + 2\|g(x^*)\|^2,$$

and taking the supremum over $x^* \in \mathcal{X}^*$ and expectation together with (57) gives the result.

4.2 Convergence for convex and smooth functions

This next proof is a simplified version of Theorem D.6 in [2]

Theorem 4.5. Assume each $f_i(x)$ is convex and $L_{\max}$–smooth. Let $0 < \alpha_k < \frac{1}{2L_{\max}}$ for all $k \in \mathbb{N}$. It follows that

1. If $\alpha_k = \alpha \leq \frac{1}{2L_{\max}}$, then for every k we have that
$$\mathbb{E}\left[f(\bar{x}^k) - f(x^*)\right] \leq \frac{1}{k}\frac{\|x^0 - x^*\|^2}{2\alpha(1-2\alpha L_{\max})} + \frac{\alpha\sigma^2}{1-2\alpha L_{\max}}, \tag{61}$$

2. If $\alpha_k = \frac{\alpha}{\sqrt{k+1}}$ with $\alpha \leq \frac{1}{2L_{\max}}$, then for every k we have that
$$\mathbb{E}\left[f(\bar{x}^k) - f(x^*)\right] \leq \frac{\|x^0 - x^*\|^2 + 2\alpha^2\sigma^2(\log(k)+1)}{4\alpha\big(\sqrt{k} - 1 - \alpha L_{\max}(\log(k)+1)\big)} \sim O\left(\frac{\log(k)}{\sqrt{k}}\right). \tag{62}$$

Proof: To prove both of the above two statements, for $\alpha_t \leq \frac{1}{2L_{\max}}$ we will first prove the following statement

$$\mathbb{E}\left[f(\bar{x}^k) - f(x^*)\right] \leq \frac{\|x^0 - x^*\|^2}{2\sum_{i=0}^{k-1}\alpha_i(1-2\alpha_i L_{\max})} + \sigma^2 \frac{\sum_{t=0}^{k-1}\alpha_t^2}{\sum_{i=0}^{k-1}\alpha_i(1-2\alpha_i L_{\max})}, \tag{63}$$

where

$$\bar{x}^k \overset{\text{def}}{=} \sum_{i=0}^{k-1} p_i x^i, \quad \text{and} \quad p_k \overset{\text{def}}{=} \frac{\alpha_k(1-2\alpha_k L_{\max})}{\sum_{i=0}^{t-1}\alpha_i(1-2\alpha_i L_{\max})} \tag{64}$$

and where

$$p_k \overset{\text{def}}{=} \frac{\alpha_k(1-2\alpha_k L_{\max})}{\sum_{i=0}^{t-1}\alpha_i(1-2\alpha_i L_{\max})}$$

$$\|x^{k+1} - x^*\|^2 = \|x^k - x^*\|^2 - 2\alpha_k\langle \nabla f_i(x^k), x^k - x^*\rangle + \alpha_k^2\|\nabla f_i(x^k)\|^2$$

Hence, taking expectation conditioned on x_k, we have:

$$\begin{aligned}\mathbb{E}_k\left[\|x^{k+1} - x^*\|^2\right] &= \|x^k - x^*\|^2 - 2\alpha_k\langle \nabla f(x^k), x^k - x^*\rangle + \alpha_k^2\mathbb{E}_k\left[\|\nabla f_i(x_k)\|^2\right] \\ &\overset{(54)+(60)}{\leq} \|x^k - x^*\| - 2\alpha_k(1-2\alpha_k L_{\max})(f(x^k) - f^*)) + 2\alpha_k^2\sigma^2.\end{aligned}$$

Rearranging and taking expectation, we have

$$2\alpha_k(1-2\alpha_k L_{\max})\mathbb{E}\left[f(x^k) - f^*\right] \leq \mathbb{E}\left[\|x^k - x^*\|^2\right] - \mathbb{E}\left[\|x^{k+1} - x^*\|^2\right] + 2\alpha_k^2\sigma^2.$$

Summing over $k = 0, \ldots, t-1$ and using telescopic cancellation gives

$$2\sum_{k=0}^{t-1}\alpha_k(1-2\alpha_k L_{\max})\mathbb{E}\left[f(x_k) - f^*\right] \leq \|x^0 - x^*\|^2 - \mathbb{E}\left[\|x^k - x^*\|^2\right] + 2\sigma^2\sum_{k=0}^{t-1}\alpha_k^2.$$

Since $\mathbb{E}\left[\|x^k - x^*\|^2\right] \geq 0$, dividing both sides by $2\sum_{i=1}^{t} \alpha_i(1-2\alpha_k L_{\max})$ gives:

$$\sum_{k=0}^{t-1} \mathbb{E}\left[\frac{\alpha_k(1-2\alpha_k L_{\max})}{\sum_{i=0}^{t-1}\alpha_i(1-2\alpha_i L_{\max})}(f(x^k)-f^*)\right] \leq \frac{\|x^0-x^*\|^2}{2\sum_{i=0}^{t-1}\alpha_i(1-2\alpha_i L_{\max})} + \frac{\sigma^2\sum_{k=0}^{t-1}\alpha_k^2}{\sum_{i=0}^{t-1}\alpha_i(1-2\alpha_i L_{\max})}.$$

Finally, let

$$p_k \overset{\text{def}}{=} \frac{\alpha_k(1-2\alpha_k L_{\max})}{\sum_{i=0}^{t-1}\alpha_i(1-2\alpha_i L_{\max})}$$

and note that $p_k \geq 0$ and $\sum_{i=0}^{t-1} p_i = 1$. This allows us to treat the p_i's as if they were probabilities. Indeed, using that $f(x)$ is convex together with Jensen's inequality gives

$$\begin{aligned}\mathbb{E}\left[f(\bar{x}^k) - f(x^*)\right] &\leq \sum_{k=0}^{t-1}\mathbb{E}\left[\frac{\alpha_k(1-2\alpha_k L_{\max})}{\sum_{i=0}^{t-1}\alpha_i(1-2\alpha_i L_{\max})}(f(x^k)-f^*)\right] \\ &\leq \frac{\|x^0-x^*\|^2}{2\sum_{i=0}^{t-1}\alpha_i(1-2\alpha_i L_{\max})} + \frac{\sigma^2\sum_{k=0}^{t-1}\alpha_k^2}{\sum_{i=0}^{t-1}\alpha_i(1-2\alpha_i L_{\max})}.\end{aligned}$$

For the different choices of step sizes:

1. If $\forall k \in \mathbb{N},\ \alpha_k = \alpha \leq \frac{1}{2L_{\max}}$, then it suffices to replace $\alpha_k = \alpha$ in (63).
2. For $\alpha_t = \frac{\alpha}{\sqrt{t+1}}$ and using the integral bound, we have that

$$\sum_{t=0}^{k-1}\alpha_t^2 = \alpha^2\sum_{t=0}^{k-1}\frac{1}{t+1} \ \leq\ \alpha^2\left(\log(k)+1\right). \tag{65}$$

Furthermore using the integral bound again we have that

$$\sum_{t=0}^{k-1}\alpha_t \ \geq\ 2\alpha\left(\sqrt{k}-1\right). \tag{66}$$

Now using (93) and (66) we have that

$$\begin{aligned}\sum_{i=0}^{k-1}\alpha_i(1-2\alpha_i L_{\max}) &= \sum_{i=0}^{k-1}\alpha_i - 2L_{\max}\sum_{i=0}^{k-1}\alpha_i^2 \\ &\geq 2\alpha\left(\sqrt{k}-1-\alpha L_{\max}\left(\log(k)+1\right)\right).\end{aligned}$$

It remains to replace bound the sums in (63) by the values we have computed.

4.3 Convergence for mini-batching and strongly convex

In practice we compute the stochastic gradient using a small batch of data, instead of a single data point. That is, we update following

$$x^{t+1} = x^t - \alpha_t \nabla f_B(x^t) \tag{67}$$

where $B \subset \{1, \ldots, n\}$ is a subset of the data and

$$\nabla f_B(x^t) \overset{\text{def}}{=} \frac{1}{|B|} \sum_{i \in B} \nabla f_i(x^t).$$

We refer to this a mini-batching. Mini-batching makes better use of parallel computational resources and it also speeds-up the convergence of SGD, as we show next. All the proofs in this section are taken from [3]. We assume that the batches B are sampled uniformly from all batches of size $b \in \mathbb{N}$, that is

$$\mathbb{P}[B] = \frac{1}{\binom{n}{b}} = \frac{(n-b)!b!}{n!} \tag{68}$$

To prove convergence, we will assume that f_i is convex and smooth. We now prove a bound on the expected norm of the gradient that is similar to Lemma 4.4. The difference is that our bound now improves as the batch size b increases.

Lemma 4.6. Let f_i be convex and $L_{\max}$–smooth for $i = 1, \ldots, n$, as stated in Assumption 4.1. Consider the definition of the gradient noise σ given in (55). It follows that

$$\mathbb{E}_t \left[\|\nabla f_B(x^t)\|^2 \right] \leq 4\mathcal{L}_b(f(x) - f^*) + 2\sigma_b^2, \tag{69}$$

where

$$\begin{aligned} \mathcal{L}_b &= \frac{n(b-1)}{b(n-1)} L + \frac{n-b}{b(n-1)} L_{\max} \\ \sigma_b &= \frac{1}{nb} \cdot \frac{n-b}{n-1} \sum_{i=1}^{n} \|\nabla f_i(x^*)\|^2 = \frac{1}{b} \cdot \frac{n-b}{n-1} \sigma. \end{aligned} \tag{70}$$

Proof: See Proposition 3.8 and 3.10 in [3].
Note that for $b = 1$ we have that $\mathcal{L}_b = L_{\max}$ and $\sigma_b = \sigma$, thus Lemma 4.4 recovers the results in Lemma 4.4 as a special case. On the other extreme, when $b = n$ we have that $\mathcal{L}_b = L$ and $\sigma_b = 0$, thus Lemma 4.4 recovers the bound (14) as a special case. Consequently on either extreme, $b = 1$ or $b = n$, we get a type bound on $\mathbb{E}_t \left[\|\nabla f_B(x^t)\|^2 \right]$. This bound (69) can now be used to establish a simple proof of convergence.

Theorem 4.7. Assume f is μ-strongly convex. In particular, we only need that f be strongly convex *around* x^* ,that is

$$f(x^*) \geq f(x) + \langle \nabla f(x), x^* - x \rangle + \frac{\mu}{2} \|x^* - x\|^2. \tag{71}$$

Assume that f_i is convex and $L_{\max}$–smooth for $i = 1, \ldots, n$. Choose $\alpha \in (0, \frac{1}{2\mathcal{L}_b}]$. The iterates of SGD given by (67) satisfy:

$$\mathbb{E}\|x^k - x^*\|^2 \leq (1 - \alpha\mu)^k \|x^0 - x^*\|^2 + \frac{2\alpha\sigma_b^2}{\mu}. \tag{72}$$

In particular, for $\alpha = \frac{1}{2\mathcal{L}_b}$ and plugging in the definition of $\mathcal{L}_b$ and σ_b in (70) in the above gives

$$\mathbb{E}\|x^k - x^*\|^2 \leq \left(1 - \frac{\mu}{2}\frac{b(n-1)}{n(b-1)L + (n-b)L_{\max}}\right)^k \|x^0 - x^*\|^2 + \sigma\frac{(n-b)}{n(b-1)L + (n-b)L_{\max}}. \tag{73}$$

Proof: Let $r^k = x^k - x^*$. From (67), we have

$$\begin{aligned} \|r^{k+1}\|^2 &\overset{(67)}{=} \|x^k - x^* - \alpha\nabla f_B(x^k)\|^2 \\ &= \|r^k\|^2 - 2\alpha\langle r^k, \nabla f_B(x^k)\rangle + \alpha^2\|\nabla f_B(x^k)\|^2. \end{aligned}$$

Taking expectation conditioned on x^k we obtain:

$$\begin{aligned} \mathbb{E}_k\left[\|r^{k+1}\|^2\right] &\overset{(53)}{=} \|r^k\|^2 - 2\alpha\langle r^k, \nabla f(x^k)\rangle + \alpha^2\mathbb{E}_k\left[\|\nabla f_B(x^k)\|^2\right] \\ &\overset{(71)}{\leq} (1-\alpha\mu)\|r^k\|^2 - 2\alpha[f(x^k) - f(x^*)] + \alpha^2\mathbb{E}_k\left[\|\nabla f_B(x^k)\|^2\right]. \end{aligned}$$

Taking expectations again and using Lemma 4.6:

$$\begin{aligned} \mathbb{E}\left[\|r^{k+1}\|^2\right] &\overset{(69)}{\leq} (1-\alpha\mu)\mathbb{E}\|r^k\|^2 + 2\alpha^2\sigma_b^2 + 2\alpha(2\alpha\mathcal{L}_b - 1)\mathbb{E}[f(x^k) - f(x^*)] \\ &\leq (1-\alpha\mu)\mathbb{E}\left[\|r^k\|^2\right] + 2\alpha^2\sigma_b^2, \end{aligned}$$

where we used in the last inequality that $2\alpha\mathcal{L}_b \leq 1$ since $\alpha \leq \frac{1}{2\mathcal{L}_b}$. Recursively applying the above and summing up the resulting geometric series gives

$$\begin{aligned} \mathbb{E}\|r^k\|^2 &\leq (1-\alpha\mu)^k\|r^0\|^2 + 2\sum_{j=0}^{k-1}(1-\alpha\mu)^j\alpha^2\sigma_b^2 \\ &\leq (1-\alpha\mu)^k\|r^0\|^2 + \frac{2\alpha\sigma_b^2}{\mu}. \quad \square \end{aligned} \tag{74}$$

5 Stochastic Proximal Gradient Descent

Robert: Coming soon!

6 Stochastic Momentum

This section is based on [5]. For most, if not all, machine learning applications SGD is used with *momentum*. In the machine learning community, the *momentum method* is often written as

$$\begin{aligned} m^t &= \hat{\beta}_t m^{t-1} + \nabla f_i(x^t) \\ x^{t+1} &= x^t - \alpha_t m^t, \end{aligned} \tag{75}$$

where $\hat{\beta}_t \in [0,\ 1]$ is the tth momentum parameter. In the optimization community it is more often written in the *heavy ball* format which is

$$x^{t+1} = x^t - \alpha_t\,\nabla f_i(x^t) + \beta_t(x^t - x^{t-1}), \tag{76}$$

where $\beta_t \in [0,\ 1]$ is another momentum parameter, $i \in \{1,\ldots,n\}$ is sampled uniformly and i.i.d at each iteration.

These two ways of writing down momentum in (75) and (76) are equivalent, as we show next.

Lemma 6.1. If

$$\beta_t = \frac{\alpha_t \hat{\beta}_t}{\alpha_{t-1}}, \tag{77}$$

then x^t iterates given by (75) and (76) are equal.

Proof: Starting (75) we have that

$$\begin{aligned} x^{t+1} &= x^t - \alpha_t m^t \\ &\overset{(75)}{=} x^t - \alpha_t \hat{\beta}_t m^{t-1} - \alpha_t \nabla f_i(x^t). \end{aligned}$$

Now using (75) at time $t-1$ we have that $m^{t-1} = \frac{x^{t-1}-x^t}{\alpha_{t-1}}$ in the above gives

$$\begin{aligned} x^{t+1} &= x^t - \frac{\alpha_t \hat{\beta}_t}{\alpha_{t-1}}(x^{t-1} - x^t) - \alpha_t \nabla f_i(x^t) \\ &\overset{(77)}{=} x^t - \alpha_t \nabla f_i(x^t) + \beta_t (x^t - x^{t-1}) \end{aligned}$$

There is yet a third equivalent way of writing down the momentum method that will be useful in establishing convergence.

Theorem 6.2. Let $\eta_k, \lambda_k \in \mathbb{R}$. Consider the iterate-moving-average (IMA) method:

$$\begin{aligned} z_k &= z_{k-1} - \eta_k \nabla f_i(x_k), \\ x_{k+1} &= \frac{\lambda_{k+1}}{\lambda_{k+1}+1} x_k + \frac{1}{\lambda_{k+1}+1} z_k, \end{aligned} \tag{78}$$

when we set $z_0 = x_0$. If

$$\alpha_k = \frac{\eta_k}{1+\lambda_{k+1}} \quad \text{and} \quad \beta_k = \frac{\lambda_k}{1+\lambda_{k+1}}, \tag{79}$$

then the x_k iterates in (78) are equal to the x_k iterates of (76) .

Proof: Consider the iterate-averaging method

$$z_k = z_{k-1} - \eta_k \nabla f_i(x_k), \tag{80}$$

$$x_{k+1} = \frac{\lambda_{k+1}}{\lambda_{k+1}+1} x_k + \frac{1}{\lambda_{k+1}+1} z_k, \tag{81}$$

and let

$$\alpha_k = \frac{\eta_k}{\lambda_{k+1}+1} \quad \text{and} \quad \beta_k = \frac{\lambda_k}{\lambda_{k+1}+1}. \tag{82}$$

Substituting (80) into (81) gives

$$x_{k+1} = \frac{\lambda_{k+1}}{\lambda_{k+1}+1} x_k + \frac{1}{\lambda_{k+1}+1} \left(z_{k-1} - \eta_k \nabla f_i(x_k)\right). \tag{83}$$

Now using (81) at the previous iteration we have that that

$$z_{k-1} = (\lambda_k + 1)\left(x_k - \frac{\lambda_k}{\lambda_k + 1} x_{k-1}\right) = (\lambda_k + 1)x_k - \lambda_k x_{k-1}.$$

Substituting the above into (83) gives

$$\begin{aligned} x_{k+1} &= \frac{\lambda_{k+1}}{\lambda_{k+1}+1} x_k + \frac{1}{\lambda_{k+1}+1}\left((\lambda_k + 1)x_k - \lambda_k x_{k-1} - \eta_k \nabla f_i(x_k)\right) && (84)\\ &= x_k - \frac{\eta_k}{\lambda_{k+1}+1}\nabla f_i(x_k) + \frac{\lambda_k}{\lambda_{k+1}+1}\left(x_k - x_{k-1}\right). && (85)\end{aligned}$$

Consequently by using (82) gives the result.

6.1 Convergence for Convex

First we provide a convergence theorem for any sequence of step sizes. Later we develop special cases of this theorem through different choices of the step sizes.

Theorem 6.3. Let each f_i be convex and $L_{\max}$–smooth. Let $x_{-1} = x_0$ and let

$$\sigma^2 \overset{\text{def}}{=} \max_{x^* \in \mathcal{X}} \frac{1}{n}\sum_{i=1}^{2} \|\nabla f_i(x^*)\|^2.$$

Let $(\eta_k)_k$ be such that $0 < \eta_k \leq \frac{1}{4L_{\max}}$ for all $k \in \mathbb{N}$. Let

$$\lambda_0 \overset{\text{def}}{=} 0 \quad \text{and} \quad \lambda_k = \frac{\sum_{t=0}^{k-1}\eta_t}{2\eta_k} \text{ for all } k \geq 1. \tag{86}$$

Consider the iterates of (78). It follows that

$$\mathbb{E}\left[f(x_k) - f_*\right] \leq \frac{\|x_0 - x^*\|^2}{\sum_{t=0}^{k}\eta_t} + 2\sigma^2 \frac{\sum_{t=0}^{k}\eta_t^2}{\sum_{t=0}^{k}\eta_t}. \tag{87}$$

Proof:

The proof uses the following Lyaponuv function

$$L_k = \mathbb{E}\left[\|z_k - x_*\|^2\right] + 2\eta_k\lambda_k \mathbb{E}\left[f(x_{k-1}) - f_*\right]$$

We have

$$\begin{aligned} \|z_{k+1} - x_*\|^2 &= \|z_k - x_* - \eta_k \nabla f_{i_k}(x_k)\|^2 \\ &\overset{(78)}{=} \|z_k - x_*\|^2 - 2\eta_k\langle \nabla f_{i_k}(x_k), z_k - x_*\rangle + \eta_k^2\|\nabla f_{i_k}(x_k)\|^2 \\ &\overset{(78)}{=} \|z_k - x_*\|^2 - 2\eta_k\langle \nabla f_{i_k}(x_k), x_k - x_*\rangle - 2\eta_k\lambda_k\langle \nabla f_{i_k}(x_k), x_k - x_{k-1}\rangle + \eta_k^2\|\nabla f_{i_k}(x_k)\|^2 \end{aligned}$$

Then taking conditional expectation $\mathbb{E}_k[\cdot] \overset{\text{def}}{=} \mathbb{E}[\cdot \mid x_k]$ we have

$$\begin{aligned}
\mathbb{E}_k\left[\|z_{k+1}-x_*\|^2\right] &= \|z_k - x_*\|^2 - 2\eta_k\langle\nabla f(x_k), x_k - x_*\rangle \\
&- 2\eta_k\lambda_k\langle\nabla f(x_k), x_k - x_{k-1}\rangle + \eta_k^2\mathbb{E}_k\left[\|\nabla f_{i_k}(x_k)\|^2\right], \\
&\overset{(56)+(54)}{\leq} \|z_k - x_*\|^2 + 4\eta_k^2 L_{\max}\left(f(x_k) - f_*\right) + 2\eta_k^2\sigma^2 \\
&- 2\eta_k\left(f(x_k) - f_*)\right) - 2\eta_k\lambda_k\left(f(x_k) - f(x_{k-1})\right) \\
&= \|z_k - x_*\|^2 - 2\eta_k\left(1 + \lambda_k - 2\eta_k L_{\max}\right)\left(f(x_k) - f_*\right) \\
&+ 2\eta_k\lambda_k\left(f(x_{k-1}) - f_*\right) + 2\eta_k^2\sigma^2. && (88) \\
&\leq \|z_k - x_*\|^2 - 2\eta_k\left(\frac{1}{2} + \lambda_k\right)\left(f(x_k) - f_*\right) \\
&+ 2\eta_k\lambda_k\left(f(x_{k-1}) - f_*\right) + 2\eta_k^2\sigma^2, && (89)
\end{aligned}$$

where we used the fact that $\eta_k \leq \frac{1}{4L_{\max}}$ in the last inequality. Since $\lambda_{k+1} = \frac{\sum_{t=0}^k \eta_t}{2\eta_{k+1}}$ we have that

$$\eta_{k+1}\lambda_{k+1} = \eta_k\left(\frac{1}{2} + \lambda_k\right).$$

Using this in (88) then taking expectation and rearranging gives

$$\mathbb{E}\left[\|z_{k+1}-x_*\|^2\right] + 2\eta_{k+1}\lambda_{k+1}\mathbb{E}\left[f(x_k) - f_*\right] \leq \mathbb{E}\left[\|z_k - x_*\|^2\right] + 2\eta_k\lambda_k\mathbb{E}\left[f(x_{k-1}) - f_*\right] + 2\eta_k^2\sigma^2.$$

Summing over $t = 0$ to k and using a telescopic sum, we have

$$\mathbb{E}\left[\|z_{k+1}-x_*\|^2\right] + \left(\sum_{t=0}^k \eta_t\right)\mathbb{E}\left[f(x_k) - f_*\right] \leq \|x_0 - x^*\|^2 + 2\sigma^2\sum_{t=0}^k \eta_t^2,$$

where we used that $\lambda_0 = 0$. Thus, writing λ_k explicitly, gives

$$\mathbb{E}\left[f(x_k) - f_*\right] \leq \frac{\|x_0 - x^*\|^2}{\sum_{t=0}^k \eta_t} + \frac{2\sigma^2\sum_{t=0}^k \eta_t^2}{\sum_{t=0}^k \eta_t}.$$

In Theorem 6.3 the only free parameters are the η_k's which in the iterate-moving-average viewpoint (78) play the role of a learning rate. The scaled step sizes α_k and the momentum parameters β_k of(76) are given by (79) once we have chosen η_k. We now explore three different settings of the η_k's in the following corollaries.

Corollary 6.4. Consider the setting of Theorem 6.3. Let $\eta \leq 1/4L_{\max}$.

1. Let $\eta_k = \eta$. Then,
$$\mathbb{E}\left[f(x_k) - f_*\right] \leq \frac{\|x_0 - x_*\|^2}{\eta\,(k+1)} + 2\eta\sigma^2. \tag{90}$$

2. Let $\eta_k = \frac{\eta}{\sqrt{k+1}}$. Then,
$$\mathbb{E}\left[f(x_k) - f_*\right] \leq \frac{\|x^0 - x^*\|_2^2 + 4\sigma^2\eta^2\left(\log(k+1) + 1\right)}{2\eta\left(\sqrt{k+1} - 1\right)} \sim\ O\left(\frac{\log(k)}{\sqrt{k}}\right). \tag{91}$$

3. Suppose Algorithm (76) is run for T iterations. Let $\eta_k = \frac{\eta}{\sqrt{T+1}}$ for all $k \in \{0, \ldots, T\}$. Then,

$$\mathbb{E}\left[f(x_T) - f_*\right] \leq \frac{\|x^0 - x^*\|_2^2 + 2\sigma^2\eta^2}{\eta\sqrt{T+1}}. \tag{92}$$

Proof: The bounds (90) and (92) can be easily derived from Theorem 6.3. As for (91), using the integral bound and plugging in our choice of η_k gives

$$\sum_{t=0}^{k-1} \eta_t^2 = \eta^2 \sum_{t=0}^{k-1} \frac{1}{t+1} \ \leq\ \eta^2 \left(\log(k) + 1\right) \quad \text{and} \quad \sum_{t=0}^{k-1} \eta_t \geq 2\eta\left(\sqrt{k} - 1\right), \tag{93}$$

which we use to obtain (91).

References

[1] S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge University Press, 2004.

[2] R. M. Gower, O. Sebbouh, and N. Loizou. “SGD for Structured Nonconvex Functions: Learning Rates, Minibatching and Interpolation”. In: *arXiv:2006.10311* (2020).

[3] R. M. Gower et al. “SGD: General Analysis and Improved Rates”. In: *International Conference on Machine Learning.* 2019, pp. 5200–5209.

[4] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course.* Vol. 87. Springer Science & Business Media, 2013.

[5] O. Sebbouh, R. M. Gower, and A. Defazio. “On the convergence of the Stochastic Heavy Ball Method”. In: *arXiv:2006.07867* (2020).