Convergence Theorems for Gradient Descent

Robert M. Gower.

September 16, 2019

Abstract

Here you will find a growing collection of proofs of the convergence of gradient and stochastic gradient descent type method on convex, strongly convex and smooth functions. Important disclaimer: Theses notes do not compare to a good book or well prepared lecture notes. You should only read these notes if you have sat through my lecture on the subject and would like to see detailed notes based on my lecture as a reminder. Under any other circumstances, I highly recommend reading instead the first few chapters of the books [4] and [1].

Contents

1 Assumptions and Lemmas

1.1 Convexity

Let $f : \mathbb{R}^d \mapsto \mathbb{R}$ be a twice continuously differentiable 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 *Let f be twice continuously differentiable. Then f is convex is 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) $\Rightarrow$ (1) Let $x_t = tx + (1-t)y$. From (2) we have that

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

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) $\Rightarrow$ (3) Since f is twice differentiable, taking a directional derivative in the v direction in (2) gives

$$0 \geq \langle \nabla f(x), v \rangle + \left\langle \nabla^2 f(x) v, y - x \right\rangle - \langle \nabla f(x), v \rangle = \left\langle \nabla^2 f(x) v, y - x \right\rangle, \quad \forall x, y, v \in \mathbb{R}^d. \tag{4}$$

Setting $y = x - v$ then gives

$$0 \leq \left\langle \nabla^2 f(x) v, v \right\rangle, \quad \forall x, v \in \mathbb{R}^d. \tag{5}$$

The above is equivalent to saying the $\nabla^2 f(x) \succeq 0$ is positive semi-definite for every $x \in \mathbb{R}^d$.

(3) $\Rightarrow$ (2) Using Taylor expansion we have that

$$\begin{aligned} f(x) &= f(y) + \langle \nabla f(y), x-y \rangle + \int_{\tau=0}^{1} \int_{s=0}^{\tau} \left\langle \nabla^2 f(y+\tau(x-y))(x-y), (x-y) \right\rangle d\tau ds. \\ &\overset{(5)}{\geq} f(y) + \langle \nabla f(y), x-y \rangle . \quad \blacksquare \end{aligned} \tag{6}$$

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

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

Lemma 1.2 *Let $f : \mathbb{R}^d \mapsto \mathbb{R}$ be a twice differentiable function. If f is L–smooth then the following holds*

$$\left\langle \nabla^2 f(x) v, v \right\rangle \ \leq \ L\|v\|_2^2, \quad \forall x, v \in \mathbb{R}^d, \tag{10}$$

$$f(y) \ \leq \ f(x) + \langle \nabla f(x), y-x \rangle + \frac{L}{2}\|y-x\|_2^2. \tag{11}$$

Proof:

(9) $\Rightarrow$ (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

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

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) $\Rightarrow$ (11) Using the Taylor expansion of $f(x)$ we have that

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

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

Lemma 1.3 *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

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

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 1.4 *If $f(x)$ is convex and L–smooth then*

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

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

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

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

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

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

1.4 Strong convexity

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

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

.

Lemma 1.5 *Let f be twice continuously differentiable. The following is equivalent to f being μ–strongly convex*

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

Proof:

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

Lemma 1.6 *If f is μ–strongly convex then it also satisfies the* Polyak–Lojasiewicz *condition, that is*

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

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 2.2 we will prove sublinear convergence under the assumption that f is convex. In Theorem 2.3 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

2.1.1 Average iterates

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*

$$\frac{1}{T}\sum_{t=1}^{T}[f(x^t) - f(x^*)] \leq \frac{\alpha}{4L}\frac{f(x^1) - f(x^*)}{T} + \frac{1}{\alpha}\frac{\|x^1 - x^*\|_2^2}{T}. \tag{26}$$

2.1.2 Last iterates

Theorem 2.2 *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*

$$f(x^n) - f(x^*) \leq \frac{2L\|x^1 - x^*\|^2}{n - 1}. \tag{27}$$

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

$$\begin{aligned}\|x^{t+1}-x^*\|_2^2 &= \|x^t-x^*-\tfrac{1}{L}\nabla f(x^t)\|_2^2 \\ &= \|x^t-x^*\|_2^2 - 2\tfrac{1}{L}\left\langle x^t-x^*,\nabla f(x^t)\right\rangle + \tfrac{1}{L^2}\|\nabla f(x^t)\|_2^2 \\ &\overset{(17)}{\leq} \|x^t-x^*\|_2^2 - \tfrac{1}{L^2}\|\nabla f(x^t)\|_2^2. \end{aligned} \tag{28}$$

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

$$\|x^t-x^*\|_2 \leq \|x^1-x^*\|_2. \tag{29}$$

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

$$f(x^{t+1})-f(x^*) \leq f(x^t)-f(x^*)-\frac{1}{2L}\|\nabla f(x^t)\|_2^2. \tag{30}$$

Applying convexity we have that

$$\begin{aligned} f(x^t)-f(x^*) &\leq \left\langle \nabla f(x^t), x^t-x^*\right\rangle \\ &\leq \|\nabla f(x^t)\|_2\|x^t-x^*\| \overset{(29)}{\leq} \|\nabla f(x^t)\|_2\|x^1-x^*\|. \end{aligned} \tag{31}$$

Isolating $\|\nabla f(x^t)\|_2$ in the above and inserting in (30) gives

$$f(x^{t+1})-f(x^*) \overset{(30)+(31)}{\leq} f(x^t)-f(x^*)-\underbrace{\frac{1}{2L}\frac{1}{\|x^1-x^*\|^2}}_{\beta}(f(x^t)-f(x^*))^2 \tag{32}$$

Let $\delta_t = f(x^t)-f(x^*)$. Since $\delta_{t+1}\leq\delta_t$, and by manipulating (32) 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

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

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

Re-arranging (28) gives

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

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

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(x^t)\|_2^2 &\leq \frac{L^2}{T}\left(\|x^1-x^*\|_2^2-\|x^{T+1}-x^*\|_2^2\right) \\ &\leq \frac{L^2}{T}\|x^1-x^*\|_2^2. \end{aligned} \tag{33}$$

2.3 Convergence for strongly convex and smooth functions

Now we prove some bounds that hold for strongly convex and smooth functions. In fact, if you observe, we will only use PL inequality (23) to establish the convergence result. Assuming a function satisfies the PL condition is a strictly weaker assumption then assuming strong convexity [2]. This proof is taken from [2].

Theorem 2.3 *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

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

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 Stochastic Gradient Descent

Consider the problem

$$w^* = \arg\min_{w \in \mathbb{R}^d} \mathbf{E}_{\mathcal{D}} \left[f(w, x) \right] \overset{\text{def}}{=} f(w), \tag{37}$$

where $x \sim \mathcal{D}$. Here we will consider the stochastic subgradient method. Let $f(w, x)$ be convex in w, and consider the iterates

$$w^{t+1} = w^t - \alpha_t g(w^t, x^t), \tag{38}$$

where $g(w^t, x^t) \in \partial_w f(w^t, x^t)$ and

$$\mathbf{E}\left[g(w^t, x^t) \mid w^t \right] = g(w^t) \in \partial f(w), \tag{39}$$

and where $x^t \sim \mathcal{D}$ is sampled i.i.d at each iteration. We will also consider (40) with a projection step.

3.1 Convex

Theorem 3.1 (With projection step) *Let $r, B > 0$. Let $f(w, x)$ be convex in w. Assume that and that $\mathbf{E}_{\mathcal{D}}\left[\|g(w^t, x)\|_2^2\right] \leq B^2$ where $B > 0$ for all t, and that the inputs and optimal point are in the ball $||w||_2 \leq r$. From a given $w^0 \in \mathbb{R}^d$ and $\alpha_t = \frac{D}{B\sqrt{2t}}$, consider the iterates of the projected stochastic subgradient method*

$$w^{t+1} = proj_D\left(w^t - \alpha_t g(w^t, x^t)\right), \tag{40}$$

The iterates satisfy

$$\mathbf{E}\left[f(\bar{x}_T)\right] - f(x^*) \leq \frac{3B}{D\sqrt{T}}. \tag{41}$$

The proof technique for convex function will always follow the following steps.

Proof:

$$\begin{aligned} \|w^{t+1} - w^*\|_2^2 &\overset{(40)+\text{proj}}{\leq} \|w^t - w^* - \alpha g(w^t, x^t)\|_2^2 \\ &= \|w^t - w^*\|_2^2 - 2\alpha\left\langle g(w^t, x^t), w^t - w^*\right\rangle + \alpha^2\|g(w^t, x^t)\|_2^2. \end{aligned} \tag{42}$$

Taking expectation conditioned on w^t we have that

$$\begin{aligned} \mathbf{E}\left[\|w^{t+1} - w^*\|_2^2 \,|\, w^t\right] &\overset{(39)}{=} \|w^t - w^*\|_2^2 - 2\alpha_t\left\langle g(w^t), w^t - w^*\right\rangle + \alpha_t^2\mathbf{E}_{\mathcal{D}}\left[\|g(w^t, x^t)\|_2^2\right] \\ &\leq \|w^t - w^*\|_2^2 - 2\alpha_t\left\langle g(w^t), w^t - w^*\right\rangle + \alpha_t^2 B^2 \\ &\overset{(7)}{\leq} \|w^t - w^*\|_2^2 - 2\alpha_t(f(w^t) - f(w*)) + \alpha_t^2 B^2. \end{aligned} \tag{43}$$

Taking expectation and re-arranging we have that

$$\mathbf{E}\left[f(w^t)\right] - f(w*)) \leq \frac{1}{2\alpha_t}\mathbf{E}\left[\|w^t - w^*\|_2^2\right] - \frac{1}{2\alpha_t}\mathbf{E}\left[\|w^{t+1} - w^*\|_2^2\right] + \frac{\alpha_t B^2}{2}. \tag{44}$$

Summing up from $t = 1, \ldots, T$ and using that α_t is a non-increasing sequence, we have that

$$\begin{aligned} \sum_{t=1}^{T}\mathbf{E}\left[f(w^t)\right] - f(w*)) &\leq \frac{1}{2\alpha_1}\|w^1 - w^*\|_2^2 + \frac{1}{2}\sum_{t=1}^{T-1}\left(\frac{1}{\alpha_{t+1}} - \frac{1}{\alpha_t}\right)\mathbf{E}\left[\|w^{t+1} - w^*\|_2^2\right] \\ &\quad - \frac{1}{2\alpha_{T+1}}\mathbf{E}\left[\|w^{T+1} - w^*\|_2^2\right] + \frac{B^2}{2}\sum_{t=1}^{T}\alpha_t \end{aligned} \tag{45}$$

$$\begin{aligned} &\overset{\|w\|_2 \leq r}{\leq} \frac{2}{\alpha_1}r^2 + \sum_{t=1}^{T-1}\left(\frac{1}{\alpha_{t+1}} - \frac{1}{\alpha_t}\right)2r^2 + \frac{B^2}{2}\sum_{t=1}^{T}\alpha_t \\ &\overset{\text{telescopic}}{=} \frac{2r^2}{\alpha_T} + \frac{B^2}{2}\sum_{t=1}^{T}\alpha_t. \end{aligned} \tag{46}$$

Finally let $\bar{x}_T = \frac{1}{T}\sum_{t=0}^{T-1} x_t$ and dividing by T and using Jensen's inequality we have

$$\begin{aligned} \mathbf{E}\left[f(\bar{x}_T)\right] - f(x^*) &\overset{\text{Jensen's}}{\leq} \frac{1}{T}\sum_{t=0}^{T-1}\mathbf{E}\left[f(w^t)\right] - f(w*)) \\ &\overset{(46)}{\leq} \frac{2r^2}{T\alpha_T} + \frac{B^2}{2T}\sum_{t=1}^{T}\alpha_t. \end{aligned} \tag{47}$$

Now plugging in $\alpha_t = \frac{\alpha_0}{\sqrt{t}}$ we have that

$$\mathbf{E}\left[f(\bar{x}_T)\right] - f(x^*) \leq \frac{2r^2}{\sqrt{T}\alpha_0} + \frac{\alpha_0 B^2}{2T}\sum_{t=1}^{T}\frac{1}{\sqrt{t}} \leq \frac{1}{\sqrt{T}}\left(\frac{2r^2}{\alpha_0} + \alpha_0 B^2\right). \tag{48}$$

Minimizing the right-hand side in α_0 gives $\alpha_0 = \frac{\sqrt{2}r}{B}$ and the result since

$$\frac{2r^2}{\alpha_0} + \alpha_0 B^2 = 2\sqrt{2}rB \leq 3rB. \quad \blacksquare$$

3.2 Strongly convex

3.2.1 Constant stepsize

Theorem 3.2 *Let $f(w,x)$ be convex in w. From a given $w^0 \in \mathbb{R}^d$ and $\frac{1}{\mu} > \alpha \equiv \alpha_t > 0$, consider the iterates of the stochastic subgradient method* (40)*. Assume that and that $\mathbf{E}_{\mathcal{D}}\left[\|g(w^t,x)\|_2^2\right] \leq B^2$ where $B > 0$ for all t.* [1] *The iterates satisfy*

$$\mathbf{E}\left[\|w^{t+1} - w^*\|_2^2\right] \quad \leq \quad (1-\alpha\mu)^{t+1}\|w^0 - w^*\|_2^2 + \frac{\alpha}{\mu}B^2. \tag{49}$$

Proof: From (40) we have that

$$\begin{aligned}\|w^{t+1} - w^*\|_2^2 &\overset{(39)}{=} \|w^t - w^* - \alpha g(w^t,x^t)\|_2^2 \\ &= \|w^t - w^*\|_2^2 - 2\alpha\left\langle g(w^t,x^t), w^t - w^*\right\rangle + \alpha^2\|g(w^t,x^t)\|_2^2.\end{aligned}$$

Taking expectation condition on w^t in the above gives

$$\begin{aligned}\mathbf{E}\left[\|w^{t+1} - w^*\|_2^2 \,|\, w^t\right] &= \|w^t - w^*\|_2^2 - 2\alpha\left\langle g(w^t), w^t - w^*\right\rangle + \alpha^2\mathbf{E}_{\mathcal{D}}\left[\|g(w^t,x^t)\|_2^2\right] \\ &\leq \|w^t - w^*\|_2^2 - 2\alpha\left\langle g(w^t), w^t - w^*\right\rangle + \alpha^2 B^2 \qquad (50) \\ &\overset{(21)}{\leq} (1-\alpha\mu)\|w^t - w^*\|_2^2 + \alpha^2 B^2 \qquad (51) \\ &\overset{\text{recurrence}}{\leq} (1-\alpha\mu)^{t+1}\|w^0 - w^*\|_2^2 + \sum_{i=0}^{t}(1-\alpha\mu)^i\alpha^2 B^2. \qquad (52)\end{aligned}$$

Since

$$\sum_{i=0}^{t}(1-\alpha\mu)^i\alpha^2 B^2 = \alpha^2 B^2\frac{1-(1-\alpha\mu)^{t+1}}{\alpha\mu} \leq \frac{\alpha^2 B^2}{\alpha\mu} = \frac{\alpha B^2}{\mu}, \tag{53}$$

we have that

$$\mathbf{E}\left[\|w^{t+1} - w^*\|_2^2 \,|\, w^t\right] \quad \overset{(53)+(52)}{\leq} \quad (1-\alpha\mu)^{t+1}\|w^0 - w^*\|_2^2 + \frac{\alpha B^2}{\mu}. \tag{54}$$

It now remains to taking expectation over the above.

[1] This is a very awkward assumption, and should really be proven instead.

3.3 Strongly quasi-convex

Let us assume instead that the average of functions if μ–strongly quasi convex, that is

$$f(x^*) \geq f(x) + \langle \nabla f(x), x^* - x \rangle + \frac{\mu}{2}\|x^* - x\|_2^2, \quad \forall x \in \mathbb{R}^d. \tag{55}$$

Even non-convex functions satisfy this assumption [3].

Further let each f_i be L_i–smooth and convex, thus Lemma 1.4 holds and

$$f_i(y) - f_i(x) \quad \leq \quad \langle \nabla f_i(y), y - x \rangle - \frac{1}{2L_i}\|\nabla f_i(y) - \nabla f_i(x)\|_2^2. \tag{56}$$

References

[1] Boyd. *Convex optimization theory*. Ed. by C. U. Press. Vol. 25. 3. Cambridge University Press, 2010. Chap. 1,10,11, pp. 487–487. arXiv: `1111.6189v1`.

[2] H. Karimi, J. Nutini, and M. W. Schmidt. "Linear Convergence of Gradient and Proximal-Gradient Methods Under the Polyak-Łojasiewicz Condition". In: *CoRR* abs/1608.04636 (2016).

[3] I. Necoara, Y. Nesterov, and F. Glineur. "Linear convergence of first order methods for non-strongly convex optimization". In: *Mathematical Programming* (2018), pp. 1–39.

[4] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. 1st ed. Springer Publishing Company, Incorporated, 2014.