20 / 03 / 2025

We are interested in proving the convergence of

$$x^{k+1} = x^k - \tau \nabla f(x^k) \quad \text{where } \tau > 0 \quad \text{(Gradient Step)}$$

the gradient descent algorithm is the following

1. fix $x_0 \in X \leftarrow \begin{cases} \mathbb{R}^n \\ \text{Hilbert space } \ell^2, L^2, \dots \end{cases}$
2. Compute a gradient step

$$x^{k+1} = x^k - \tau \nabla f(x^k) \quad \text{for } k \geq 0$$

In this way you compute $x_1, x_2, \dots$ that defines a sequence of points in X

Theorem Let $f \in C^1(X; \mathbb{R})$ with ∇f L-Lipschitz and $\inf f > -\infty$. (the function f is bounded from below.)

If you choose $0 < \tau < 2/L$ then the method converges in the sense that

$$* \quad \nabla f(x^k) \to 0 \quad \text{up to subsequences.}$$

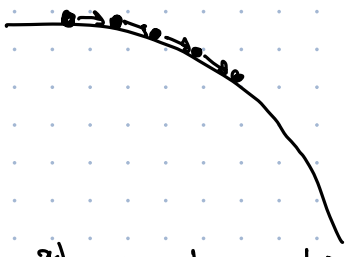
Rem 1. We are working under milder assumptions (no convexity, no coercivity etc.) so we cannot expect to say much about the convergence of the sequence.

Rem 2. In ML we don't need to actually say too much about the convergence of the weights.

Why $\inf f > -\infty$?

e^x is bounded from below $e^x \geq 0$ $\inf e^x = 0$

$$\inf -e^x = -\infty$$



$x^k \to +\infty$ ∇f along this sequence $\to -\infty$

$$(-e^x)' = -e^x$$

But $\nabla(-e^x)$ is not even Lipschitz

Another example is $f(x) = -x$

is a C^1 function

The $\nabla f = f' = -1$

$$|-1-(-1)| \leq L|x-y| \quad \forall x, y \in \mathbb{R}$$

$$0 \leq L|x-y| \quad \forall L$$

A function $g: X \to Y$ is said to be L-Lipschitz if

$$|g(x)-g(y)|_Y \leq L\,|x-y|_X$$

Simple examples of Lipschitz functions are C^1 function with $|g'|$ bounded

$$L = \sup_x |g'(x)|$$

But $f(x) = -x$ has the same problem of $-e^x$

$\nabla f \to -1$

Proof $f(x^{k+1}) = f\left(x^k - \tau \nabla f(x^k)\right)$ by definition of x^{k+1}

$\phi(t) = f(x + ty) \quad \forall x, y \in X \quad t \in \mathbb{R}$

Since $f \in C^1$ then ϕ is differentiable and

$$\phi'(t) = \nabla f(x+ty) \cdot y$$

Now I integrate between 0 and τ both sides

$$\int_0^\tau \phi'(s)\,ds = \int_0^\tau \nabla f(x+sy)\cdot y\,ds$$

$$\phi(\tau) - \phi(0) = \int_0^\tau \nabla f(x+sy)\cdot y\,ds.$$

$$f(x+\tau y) = f(x) + \int_0^\tau \nabla f(x+sy)\cdot y\,ds.$$

Lemma 1 If $f: X \to \mathbb{R}$ is C^1 then $\forall\, \tau$

$$f(x+\tau y) = f(x) + \int_0^\tau \nabla f(x+sy)\cdot y\,ds \qquad \forall\, x, y \in X.$$

$$f(x^{k+1}) = f\left(x^k - \tau \nabla f(x^k)\right)$$

Let us now use Lemma 1 with $x = x^k$ and $y = -\nabla f(x^k)$

$$f(x^k - \tau \nabla f(x^k)) = f(x^k) + \int_0^\tau \nabla f\left(x^k - s\nabla f(x^k)\right)\cdot\left(-\nabla f(x^k)\right) ds.$$

$$= f(x^k) - \int_0^\tau \nabla f\left(x^k - s\nabla f(x^k)\right)\cdot \nabla f(x^k)\,ds.$$

The idea now is to use Lipshitz condition on ∇f

$$\|\nabla f(x) - \nabla f(y)\|$$

$$f(x^k) - \int_0^\tau \left(\nabla f(x^k - s\nabla f(x^k)) - \nabla f(x^k)\right)\cdot \nabla f(x^k)\,ds - \tau\,|\nabla f(x^k)|^2$$

$$\int_0^\tau \left(\nabla f(x^k - s\nabla f(x^k)) - \nabla f(x^k) + \nabla f(x^k)\right)\cdot \nabla f(x^k)\,ds$$

$$\int_0^\tau \nabla f(x^k)\cdot\nabla f(x^k)\,ds = |\nabla f|^2$$

$$f(x^{k+1}) = f(x^k) - \tau |\nabla f(x^k)|^2 - \int_0^\tau \underbrace{(\nabla f(x^k - s\nabla f(x^k)) - \nabla f(x^k))}_{\text{Lipshitz condition.}} \cdot \nabla f(x^k)\, ds$$

We use the fact that $\forall x \in \mathbb{R} \quad x \leq |x|$

$$\leq f(x^k) - \tau |\nabla f(x^k)|^2 + \int_0^\tau |(\nabla f(x^k - s\nabla f(x^k)) - \nabla f(x^k)) \cdot \nabla f(x^k)|\, ds.$$

Now by Cauchy-Schwartz

$$\left(|x \cdot y| \leq \|x\| \|y\| \right)$$

$$\leq f(x^k) - \tau |\nabla f(x^k)|^2 + \int_0^\tau |\nabla f(x^k - s\nabla f(x^k)) - \nabla f(x^k)|\, |\nabla f(x^k)|\, ds$$

Finally I can use Lipshitz condition on ∇f

$$\leq f(x^k) - \tau |\nabla f(x^k)|^2 + \int_0^\tau L |x^k - s\nabla f(x^k) - x^k|\, |\nabla f(x^k)|\, ds.$$

$$= f(x^k) - \tau |\nabla f(x^k)|^2 + L |\nabla f(x^k)|^2 \int_0^\tau s\, ds$$

$$= f(x^k) - \tau |\nabla f(x^k)|^2 + L \frac{\tau^2}{2} |\nabla f(x^k)|^2$$

$$= f(x^k) + \tau \left(\frac{L}{2}\tau - 1 \right) |\nabla f(x^k)|^2.$$

$$= f(x^k) - \tau \left(1 - \frac{L\tau}{2} \right) |\nabla f(x^k)|^2$$

So in the end we had

$$f(x^{k+1}) \leq f(x^k) - \tau \left(1 - \frac{L\tau}{2} \right) |\nabla f(x^k)|^2. \quad (*)$$

Now I can start from x^0 and use $(*)$ recursively

$$f(x^1) \leq f(x^0) - \tau \left(1 - \frac{L\tau}{2} \right) |\nabla f(x^0)|^2$$

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

$$\leq f(x^0) - \tau \left(1 - \frac{L\tau}{2} \right) \left(|\nabla f(x^0)|^2 + |\nabla f(x^1)|^2 \right)$$

$$\vdots$$

$$f(x^n) \leq f(x^0) - \tau \underbrace{\left(1 - \frac{L\tau}{2} \right)}_{>0} \sum_{i=0}^{n-1} |\nabla f(x^i)|^2.$$

Remember that by hypothesis $\tau < \frac{2}{L}$, so $1 - \frac{L\tau}{2} > 0$ since

$$-\frac{L\tau}{2} > -1 \qquad \frac{L\tau}{2} < 1 \qquad \tau < \frac{2}{L}$$

$$-\infty < \inf f \leq f(x^n) \leq f(x^0) - \tau\left(1 - \frac{L\tau}{2}\right) \sum_{i=0}^{n-1} |\nabla f(x^i)|^2$$

$$\tau\left(1 - \frac{L\tau}{2}\right) \sum_{i=0}^{n-1} |\nabla f(x^i)|^2 \leq f(x^0) - \inf f$$

≥ 0 $\qquad$ ≥ 0

Notice that $f(x^0) - \inf f$ is a fixed positive quantity that does not depend on N.

This means that if we define

$$S_n = \sum_{i=0}^{n-1} |\nabla f(x^i)|^2$$

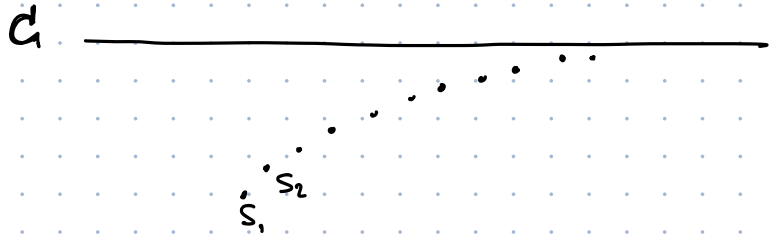
This quantity as $n \to +\infty$ is an increasing sequence meaning that

$$S_{n+1} \geq S_n.$$

Moreover (S_n) is bounded from above by

$$C = \frac{f(x_0) - \inf f}{\tau\left(1 - \frac{L\tau}{2}\right)} \geq 0$$

The situation is this



By Bolzano - Weirstrass Theorem, an increasing sequence bounded from above converges.

what is $\lim_{n \to +\infty} S_n$? $\qquad S_n = \sum_{i=0}^{n-1} |\nabla f(x^i)|^2$

By definition

$$\lim_{n \to +\infty} S_n = \sum_{i=0}^{+\infty} |\nabla f(x^i)|^2 < +\infty$$

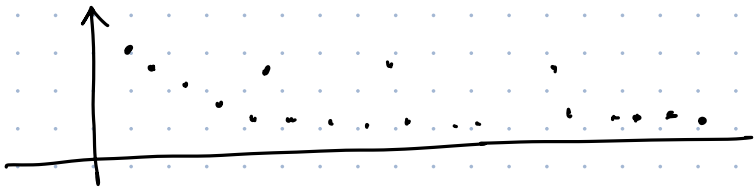
And the series converges.

$\sum_{i=0}^{+\infty} a_k < +\infty$ Then it is not properly true to say $a_k \to 0$.

This means that $\exists$ a subsequence i_k $k = 1, 2, \ldots$ such that

$$|\nabla f(x^{i_k})|^2 \to 0 \quad \text{as} \quad k \to +\infty$$

Rem From our perspective convergence up to subsequences is perfectly fine.



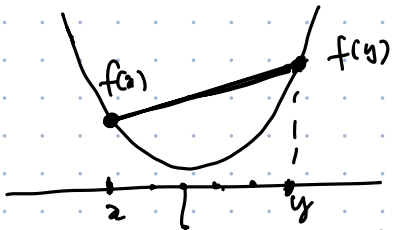
From a numerical perspective you basically fix a threshold $\varepsilon = 10^{-7}$ (this is an example) and you stop your algorithm when

$$|\nabla f|^2 < \varepsilon.$$

This theorem says that you will reach that point.
But in general we don't (and we can't) have an estimate on the time (= number of steps) we will need to reach that condition.

Convex Case

Disclaimer: now we work in $\mathbb{R}^n$, so $X = \mathbb{R}^n$ (i.e. in a finite dimensional Hilbert space)



$f((1-\lambda)x + \lambda y) \leq (1-\lambda) f(x) + \lambda f(y) \quad \forall \lambda \in [0,1]$.

↓ Linear function in λ that connects $f(x)$ with $f(y)$.

$(1-\lambda)x + \lambda y \quad \lambda \in [0,1]$

$\lambda = 0 \to x$

$\lambda = 1 \to y$

f is convex if and only if $\forall x, y \in X$

- $f((1-\lambda)x + \lambda y) \leq (1-\lambda) f(x) + \lambda f(y) \quad \forall \lambda \in [0,1]. \ \forall x, y \in X$

You should remember that

$g \in C^2((a,b); \mathbb{R})$ then g is convex iff $g'' \geq 0$ on (a,b)

↙ now we are in finite dim

Proposition: If $f \in C^2(\mathbb{R}^n, \mathbb{R})$ the f is convex if and only if $D^2 f \geq 0$

Hessian: $(D^2 f)_{ij} = \frac{\partial}{\partial x_i} \frac{\partial}{\partial x_j} f \quad \forall i, j = 1, \dots, n.$

Proof: Fix $x, y \in \mathbb{R}^n$ and define. $\psi(s) = f(x + sy)$.

Since f is convex then ψ is convex too $(1-\lambda)(x + ty)$

$$\psi((1-\lambda)t + \lambda s) = f(x + [(1-\lambda)t + \lambda s] y) = f(x - \lambda x + \lambda x + [(1-\lambda)t + \lambda s] y)$$

$$= f((1-\lambda)x + (1-\lambda) ty + \lambda(x + sy)) = f((1-\lambda)(x+ty) + \lambda(x+sy)) \leq (1-\lambda)\psi(t) + \lambda \psi(s)$$

by convexity of f ↙ $f(x+ty)$ $f(x+sy)$

this proves that if f is convex then ψ is convex. But if $\forall x, y \in \mathbb{R}^n$ ψ is convex, by definition of convexity also f is convex

But ψ is convex

if and only if $\psi''(s) \geq 0$

$$\psi'(s) = \nabla f(x+sy) \cdot y, \qquad \psi''(s) = \boxed{y \cdot D^2 f\, y \geq 0} \quad \forall y \in \mathbb{R}^n$$

On the other hand saying that

$$y \cdot D^2 f\, y \geq 0 \quad \forall y \in \mathbb{R}^n \quad \text{means} \quad D^2 f \geq 0 \qquad \square$$

Theorem (Baillon – Haddad) If f is convex and $C^2(\mathbb{R}^n, \mathbb{R})$ and ∇f is L-lipshitz then

$$\boxed{(\nabla f(x) - \nabla f(y)) \cdot (x-y) \geq \frac{1}{L} |\nabla f(x) - \nabla f(y)|^2.}$$

Lemma If $f : X \to \mathbb{R}$ is C^1 then $\forall \tau$

$$f(x+\tau y) = f(x) + \int_0^\tau \nabla f(x+sy) \cdot y \, ds \qquad \forall x, y \in X.$$

$$\nabla f(x) - \nabla f(y) = \int_0^1 D^2 f(\text{[illegible]}) \cdot (x-y) \, ds.$$

$$X = x + y \qquad y = X - Y \Rightarrow x + sy = Y + s(X-Y)$$

$$Y = x$$

Then

$$\nabla f(x) - \nabla f(y) = \int_0^1 D^2 f(y + s(x-y))\,(x-y)\, ds.$$

Then we know that $0 \leq D^2 f \leq L\,Id$

Define $A = \int_0^1 D^2 f(y + s(x-y))\, ds$ which is symmetric since $A = \int D^2 f$ and $0 \leq A \leq L\,Id$.

(*) $\nabla f(x) - \nabla f(y) = A(x-y)$ by def of $|z|^2$.

$$|\nabla f(x) - \nabla f(y)|^2 = |A(x-y)|^2 = A(x-y) \cdot A(x-y)$$

Then $A^{1/2}$ exists $(A^{1/2} \cdot A^{1/2} = A)$ and commute with A

$[A, A^{1/2}] = 0 \qquad A A^{1/2} = A^{1/2} A$

$$|\nabla f(x) - \nabla f(y)|^2 = A^{1/2} A (x-y) \cdot A^{1/2}(x-y) = A A^{1/2}(x-y) \cdot A^{1/2}(x-y)$$

Since $A \in L\,Id$

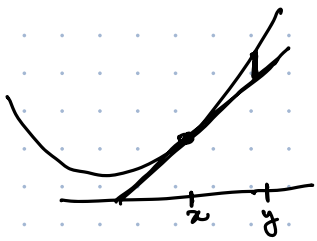
$$\leq L\,[A^{1/2}(x-y)] \cdot [A^{1/2}(x-y)] = L\,[A(x-y)] \cdot (x-y)$$

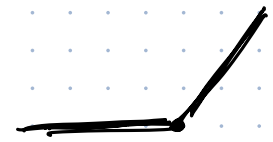
$$\overset{(*)}{=} L\,(\nabla f(x) - \nabla f(y)) \cdot (x-y)$$

$$|\nabla f(x) - \nabla f(y)|^2 \leq L\,(\nabla f(x) - \nabla f(y)) \cdot (x-y)$$

$f \in C^2$ is convex $\iff$ $D^2 f \geq 0$

$f \in C^1$ is convex $\iff$ $f(y) \geq f(x) + \nabla f(x) \cdot (x-y)$ (P)





$\partial f = [0, 1]$

$$\psi(t) = f(x + t(y-x))$$

$\psi(0) = f(x)$

$\psi(1) = f(y)$

$$\psi'(t) = \nabla f(x + t(y-x)) \cdot (y-x)$$

$$\psi'(0) = \nabla f(x) \cdot (y-x)$$

$\psi(1) - \psi(0) \geq \psi'(0)$ since ψ is convex.

$f(y) - f(x) \geq \nabla f(x) \cdot (y-x)$.

We need the following fact:

Define $T_\tau = id - \tau \nabla f$ $\qquad T_\tau(x) = x - \tau \nabla f(x)$

$$x^{k+1} = x^k - \tau \nabla f(x^k)$$

$$x^{k+1} = T_\tau(x^k)$$

T_τ is the operator that generate the sequence of points of GD.

Lemma T_τ is 1- Lipshitz.

$$|T_\tau(x) - T_\tau(y)| \leq 1 \cdot |x-y|$$

Proof. $|T_\tau(x) - T_\tau(y)|^2 = |x - \tau\nabla f(x) - y + \tau\nabla f(y)|^2$

$= |x - y - \tau(\nabla f(x) - \nabla f(y))|^2 = |x-y|^2 - 2\tau(x-y)\cdot(\nabla f(x) - \nabla f(y)) + \tau^2 |\nabla f(x) - \nabla f(y)|^2$

$\leq |x-y|^2 + \tau^2 |\nabla f(x) - \nabla f(y)|^2 - \frac{2}{L}\tau |\nabla f(x) - \nabla f(y)|^2$

$= |x-y|^2 - \frac{2\tau}{L}\left(1 - \frac{\tau L}{2}\right)|\nabla f(x) - \nabla f(y)|^2.$

($\frac{2\tau}{L} > 0$; $\left(1 - \frac{\tau L}{2}\right) > 0$; $|\nabla f(x) - \nabla f(y)|^2 \geq 0$)

Since $\tau < \frac{2}{L}$

$\leq |x-y|^2$ ∎

Now we have everything to give a bound on the convergence of GD

Assume that f has a minimum x^*.

$f(x^*) \geq f(x^k) + \nabla f(x^k)\cdot(x^k - x^*)$ by the usual characterization of convex functions

$f(x^*) - f(x^k) \geq \nabla f(x^k)\cdot(x^k - x^*)$

$f(x^k) - f(x^*) \leq \nabla f(x^k)\cdot(x^* - x^k) \leq |\nabla f(x^k)|\,|x^k - x^*|$ (*)

Now, since x^* is a minimum $T_\tau(x^*) = x^*$

$|x^{k+1} - x^*| = |T_\tau(x^k) - T_\tau(x^*)| \leq |x^k - x^*|$

$|x^k - x^*| \leq |x^0 - x^*|$

From (*) $\Delta_k = f(x^k) - f(x^*)$

$$\Delta_{k+1} \leq \Delta_k - C\frac{\Delta_k^2}{|x_0 - x^*|^2} \quad (**)$$

If Δ_k satisfies (**) then

$$\boxed{\Delta_k \leq \frac{|x_0 - x^*|^2}{C(k+1)}}$$

Chambolle Antonin, Leon Bottou

Robbins-Monroe '60 (very very hard to read)

Remember that f in ML is the empirical risk so

$$f(x) = \frac{1}{m}\sum_{i=1}^{m} f_i(x)$$

$$\hat{L}_N(w) = \frac{1}{m}\sum_{i=1}^{m} \ell\left(f(x_i, w), y_i\right)$$

f here is the model.

$\ell(f(x_i, \cdot), y_i) \to f_i$

The functions f_i are basically the loss function computed on the i-th example of the training set.

The dimension of the training set m is very very large

For modern ML problems it is not feasible to compute ∇f

The reason why we can still optimize f with a gradient based method is SGD

SGD:

1. Choose $x_0 \in \mathbb{R}^n$ randomly and you set $x^0 = x_0$
2. For every $k \geq 0$ select $i_k \in \{1, \dots, m\}$ sampled with probability $1/m$ and you compute

$$x^{k+1} = x^k - \tau_k \nabla f_{i_k}(x^k)$$

$$x^{k+1} = x^k - \tau \nabla f(x^k)$$

$$f(x^k) = \frac{1}{m}\sum_{i=1}^{m} f_i(x^k)$$

$$\nabla f(x^k) = \frac{1}{m}\sum_{i=1}^{m} \nabla f_i(x^k)$$

- $\inf f > -\infty$
- $f_i \in C^1$
- ∇f_i should be L_i-lipshitz

$$X^{k+1} = X^k - \tau \nabla f_{I_k}(X^k)$$

Now choose $j \in \{1, \dots, m\}$

$$f_j(X^{k+1}) = f_j\left(X^k - \tau \nabla f_{I_k}(X^k)\right)$$

Lemma 1 If $f : X \to \mathbb{R}$ is C^1 then $\forall\, \tau$

$$f(x + \tau y) = f(x) + \int_0^\tau \nabla f(x + sy) \cdot y \, ds \qquad \forall\, x, y \in X.$$

We use this lemma with $x = X^k$ $y = -\nabla f_{I_k}(X^k)$ $f = f_j$

$$f_j\left(X^k - \tau \nabla f_{j_k}(X^k)\right) = f_j(X^k) + \int_0^\tau \nabla f_j\left(X^k - s\nabla f_{I_k}(X^k)\right) \cdot \left(-\nabla f_{I_k}(X^k)\right) ds$$

$$= f_j(x^k) - \int_0^\tau \nabla f_j(x^k - s\nabla f_{I_k}(x^k)) \cdot \nabla f_{I_k}(x^k)\, ds.$$

$$= f_j(x^k) - \int_0^\tau \left(\nabla f_j(x^k - s\nabla f_{I_k}(x^k)) - \nabla f_j(x^k) + \nabla f_j(x^k)\right) \cdot \nabla f_{I_k}(x^k)\, ds.$$

$$= f_j(x^k) - \tau\, \nabla f_j(x^k)\cdot \nabla f_{I_k}(x^k) - \int_0^\tau \left(\nabla f_j(x^k - s\nabla f_{I_k}(x^k)) - \nabla f_j(x^k)\right) \cdot \nabla f_{I_k}(x^k)\, ds$$

↙ Cauchy-Schwartz + Lipshitz of ∇f_j

$$\le f_j(x^k) - \tau\, \nabla f_j(x^k)\cdot \nabla f_{I_k}(x^k) + \int_0^\tau L_j \left| -s\nabla f_{I_k}(x^k)\right| \left|\nabla f_{I_k}(x^k)\right| ds.$$

$$= f_j(x^k) - \tau\, \nabla f_j(x^k)\cdot \nabla f_{I_k}(x^k) + \frac{\tau^2 L_j}{2} \left|\nabla f_{I_k}(x^k)\right|^2$$

$$f_j(x^{k+1}) \le f_j(x^k) - \tau\, \nabla f_j(x^k)\cdot \nabla f_{I_k}(x^k) + \frac{\tau^2 L_j}{2} \left|\nabla f_{I_k}(x^k)\right|^2.$$

Now we sum both sides for $j = 1 \dots m$ and divide by m $\quad \frac{1}{m}\sum_{j=1}^m$

$$\frac{1}{m}\sum_{j=1}^m f_j(x^{k+1}) = f(x^{k+1})$$

$$f(x^{k+1}) \le f(x^k) - \tau\, \nabla f(x^k)\cdot \nabla f_{I_k}(x^k) + \frac{\tau^2 \bar{L}}{2} \left|\nabla f_{I_k}(x^k)\right|^2$$

where $\bar{L} = \frac{1}{m}\sum_{i=1}^m L_i$

$$E\left(f(x^{k+1}) \mid x^k\right) \le f(x^k) - \tau\, \nabla f(x^k)\cdot \nabla f(x^k) + \frac{\tau^2 \bar{L}}{2} \frac{1}{m}\sum_{i=1}^m \left|\nabla f_i(x^k)\right|^2$$

$\left|\nabla f(x^k)\right|^2$

$\left|\nabla f_i(x^k) - \nabla f(x^k)\right|^2$ $\qquad E(X - EX)^2 = EX^2 - (EX)^2$

$$\frac{1}{m}\sum_{i=1}^m \left|\nabla f_i(x^k) - \nabla f(x^k)\right|^2 = \frac{1}{m}\sum_{i=1}^m \left|\nabla f_i(x^k)\right|^2 + \left|\nabla f(x^k)\right|^2 - 2\nabla f_i(x^k)\cdot \nabla f(x^k)$$

$$= \frac{1}{m}\sum_{i=1}^m \left|\nabla f_i(x^k)\right|^2 + \left|\nabla f(x^k)\right|^2 - 2\left|\nabla f(x^k)\right|^2$$

$$= \frac{1}{m}\sum_{i=1}^m \left|\nabla f_i(x^k)\right|^2 - \left|\nabla f(x^k)\right|^2$$

$$\frac{1}{m}\sum_{i=1}^m \left|\nabla f_i(x^k)\right|^2 = \frac{1}{m}\sum_{i=1}^m \left|\nabla f_i(x^k) - \nabla f(x^k)\right|^2 + \left|\nabla f(x^k)\right|^2$$

$$E\left(f(x^{k+1}) \mid x^k\right) \le f(x^k) - \tau\left(1 - \frac{\tau \bar{L}}{2}\right)\left|\nabla f(x^k)\right|^2 + \frac{\tau^2 \bar{L}}{2}\frac{1}{m}\sum_{i=1}^m \left|\nabla f_i(x^k) - \nabla f(x^k)\right|^2$$

$$E[f(x^{k+1})|x^k] \leq f(x^k) - \tau\left(1 - \frac{\tau L}{2}\right)|\nabla f(x^k)|^2 + \frac{\tau^2 L}{2}\frac{1}{m}\sum_{i=1}^{m}|\nabla f_i(x^k) - \nabla f(x^k)|^2 \quad (*)$$

$$E(E(X|Y)) = EX$$

$$E(E(f(x^{k+1})|x^k)) = E\,f(x^{k+1})$$

$$E\,f(x^{k+1}) \leq E\,f(x^k) - \tau\left(1 - \frac{\tau L}{2}\right)E|\nabla f(x^k)|^2 + \frac{\tau^2 L}{2}E\left(\frac{1}{m}\sum_{i=1}^{m}|\nabla f_i(x^k) - \nabla f(x^k)|^2\right)$$

In order to prove the convergence of SGD we need to add an hypothesis on the variance of the gradients.

$$\frac{1}{m}\sum_{i=1}^{m}|\nabla f_i(x) - \nabla f(x)|^2 \leq \sigma^2 \quad \underline{\forall\, x \in \mathbb{R}^n}$$ This is a reasonable assumption

$$E\,f(x^{k+1}) \leq E\,f(x^k) - \tau\left(1 - \frac{\tau L}{2}\right)E|\nabla f(x^k)|^2 + \frac{\tau^2 L}{2}\sigma^2$$

Idea is that $\tau \to \tau_k$

$$\sum_{k=0}^{+\infty}\tau_k^2 < +\infty$$

$$\sum_{k=0}^{+\infty}\tau_k = +\infty$$