21/03/2025

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

Remember

$f = \frac{1}{m}\sum_{i=1}^m f_i$ $\quad m \sim 10^5 - 10^{10}$ $\quad$ (Imagenet : dataset of images $\sim 10^6$)

$\uparrow$ loss function evaluated on a point of the Training set $(D = \{(x_1, y_1) \cdots (x_N, y_N)\})$.

All these functions are functions of the "weights" of a NN

We cannot directly use GD (= Gradient Descent) $\quad x^{k+1} = x^k - \tau \nabla f$

Instead we use SGD

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

$i_k$ is sampled with prob. $1/m$ (i.e. uniform probability) from $\{1, \ldots, m\}$

=

"I pick an example at random and on that example I compute the gradient"

- We already understood that we have to impose a bound on the variance of $\nabla f$
- $$\boxed{\frac{1}{m}\sum_{i=1}^m |\nabla f_i(x) - \nabla f(x)|^2 \leq \sigma^2}$$

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

$$\sum_{i=}^{+\infty} |\nabla f|^2 < +\infty$$

1. In order to control the variance term our only hope is $\tau^2$.
2. The variance term is proportional to $\tau^2$ while the gradient term $(\mathbb{E} |\nabla f(x^k)|^2)$ is proportional to $\tau$

First of all we understand that we need a variable learning rate $\tau \to \tau_k$

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

We recall the bound of the variance and we also choose $\underline{\tau_k \bar{L} \leq 1}$

$$-\tau_k\left(1 - \frac{\tau_k \bar{L}}{2}\right) \leq -\frac{\tau_k}{2}.$$

$$-\tau_k + \frac{\tau_k^2 \bar{L}}{2} \leq -\frac{\tau_k}{2}, \quad -\frac{\tau_k}{2} + \frac{\tau_k^2 \bar{L}}{2} \leq 0 \quad \tau_k\left(-1 + \tau_k \bar{L}\right) \leq 0$$

Now since $\tau_k > 0 \Rightarrow -1 + \tau_k \bar{L} \leq 0$

$$E f(x^{k+1}) \leq E f(x^k) - \frac{\tau_k}{2} E|\nabla f(x^k)|^2 + \frac{\tau_k^2 \bar{L}}{2} \sigma^2$$

We apply this inequality recursively $n$ times

$$E f(x^1) \leq E f(x^0) - \frac{\tau_0}{2} E|\nabla f(x^0)|^2 + \frac{\tau_0^2 \bar{L}}{2} \sigma^2$$

$$E f(x^2) \leq E f(x^1) - \frac{\tau_1}{2} E|\nabla f(x^1)|^2 + \frac{\tau_1^2 \bar{L}}{2} \sigma^2$$

$$\leq E f(x^0) + \quad -\frac{1}{2}\left(\tau_0 E|\nabla f(x^0)|^2 + \tau_1 E|\nabla f(x^1)|^2\right)$$

$$+ \frac{\bar{L}\sigma^2}{2} \quad \tau_0^2 + \tau_1^2$$

$$E f(x^n) \leq E f(x^0) - \frac{1}{2}\sum_{k=0}^{n-1} \tau_k E|\nabla f(x^k)|^2 + \frac{\bar{L}\sigma^2}{2}\sum_{k=0}^{n-1}\tau_k^2$$

As we did for GD.

$$-\infty < \underline{\inf f} \leq E f(x^k) \leq \underline{E f(x^0) - \frac{1}{2}\sum_{k=0}^{n-1} \tau_k E|\nabla f(x^k)|^2 + \frac{\bar{L}\sigma^2}{2}\sum_{k=0}^{n-1}\tau_k^2}$$

$$\frac{1}{2}\sum_{k=0}^{n-1} \tau_k E|\nabla f(x^k)|^2 \leq E f(x^0) - \inf f + \frac{\bar{L}\sigma^2}{2}\sum_{k=0}^{n-1}\tau_k^2$$

The problem now is that $\sum_{k=0}^{n-1} \tau_k^2 \to +\infty$ as $n \to +\infty$. We need to avoid this and therefore we assume that

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

$$\frac{1}{2}\sum_{k=0}^{n-1} \tau_k E|\nabla f(x^k)|^2 \leq \underbrace{E f(x^0) - \inf f + \frac{\bar{L}\sigma^2}{2} T}_{>0}$$

If we define $S_n = \frac{1}{2}\sum_{k=0}^{n-1}\tau_k E|\nabla f(x^k)|^2$ Then $S_n \nearrow$ and bounded from above $\Rightarrow$ $S_n$ converges which means that

$$\sum_{k=0}^{+\infty} \tau_k E|\nabla f(x^k)|^2 < +\infty$$

If $\boxed{\sum_{k=0}^{+\infty} \tau_k = +\infty}$ then we are safe in the sense that

$\exists\ i_k$ s.t. $\mathbb{E}\,|\nabla f(x^{i_k})|^2 \to 0$

$$\left( \begin{array}{c} \Downarrow \\ \nabla f \to 0 \quad a.s. \end{array} \right)$$

If you take for instance $\tau_k = \frac{1}{k}$

$$\sum_{k=0}^{+\infty} \frac{1}{k^2} = \left(\frac{\pi^2}{6}\ ?\right)$$

$$\sum_{k=0}^{+\infty} \frac{1}{k} = +\infty$$

What are the assumptions we did to prove this convergence result.

Assumptions on $f$

1. $f \in \mathcal{C}^1(\mathbb{R}^n; \mathbb{R})$ ( $f$ is continuosly differentiable )
2. $\nabla f_i$ is $L_i$-Lipschitz $\forall\, i = 1, \dots, m$
3. $\inf f > -\infty$ ( $f$ is bounded from below )
4. $\frac{1}{m}\sum_{i=1}^{m} |\nabla f_i(x) - \nabla f(x)|^2 < \sigma^2$ ( Variance of the gradients bounded )

Assumption on $\tau_k$

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

Remember tha the whole algorithm works because the indices $i_k$

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

are choosen <u>uniformly at random</u> from $\{1, \dots, m\}$

SGD is <u>batch mode</u> optimization method It is NOT an online method

Minimizing Movements. ( Actually This is the discrete version of MM )

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

$$x^{k+1} \in \underset{x\in X}{\arg\min}\; f(x) + \frac{1}{2\tau}|x - x^k|^2 \qquad (*)$$

"I have to choose the next point ($x^{k+1}$) in such a way that I minimize $f$ but <u>also</u> I don't <u>go too far away from $x^k$</u>"

1. Notice that there is no gradient here
2. Instead of $|x - x^k|^2$ I can put a generic distance or score

$$d(x, x^k)$$

Now if $f \in C^1(\mathbb{R}^n, \mathbb{R})$ then

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

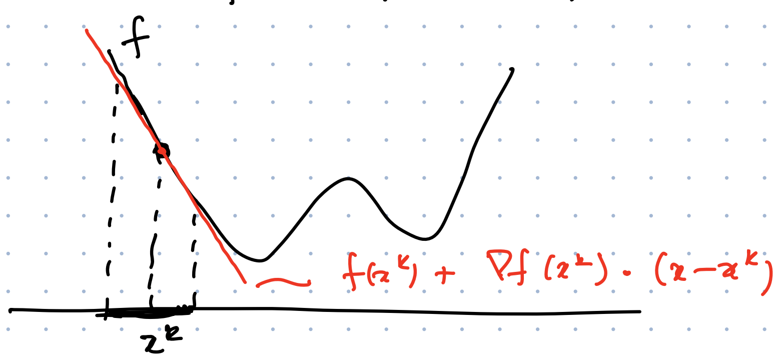
$$\nabla f(x^{k+1}) + \frac{1}{\tau}(x^{k+1} - x^k) = 0 \implies x^{k+1} = x^k - \tau \nabla f(\underline{\underline{x^{k+1}}})$$

implicit GD.

How do we recover explicit GD.

Since we are looking for solutions close to $x^k$ and since $f$ is diff. we can approximate $f$ around $x^k$ with

$$f(x) \approx f(x^k) + \nabla f(x^k) \cdot (x - x^k).$$




So I can say that instead of $(*)$ I use

$$x^{k+1} \in \underset{x\in X}{\arg\min}\; \underbrace{f(x^k) + \nabla f(x^k) \cdot (x - x^k)}_{f(x)} + \frac{1}{2\tau}|x - x^k|^2$$

Now the minimality cond. of is

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

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