Convergence de SGD pour une fonction localement μ-Polyak Lojasiewicz

Lucas Ketels

April 2023

1 Hypothèses et notations

Soit L la fonction définie par

$$\begin{aligned} \mathcal{L} : \mathbb{R}^n &\longrightarrow \mathbb{R} \\ x &\longmapsto \frac{1}{n}\sum_{i=1}^{n} l_i(x) \end{aligned}$$

où les fonctions $l_i : \mathbb{R}^n \longrightarrow \mathbb{R}$ sont des fonctions positives. On suppose que $\mathcal{L}$ est $\mu - PL^*$ sur une boule $\mathcal{B}(x_0, R)$ (i.e. $||\nabla\mathcal{L}(x)||^2 \geq \mu\mathcal{L}(x)\ \forall x \in \mathcal{B}(x_0, R)$).
On suppose de plus que les l_i sont $\beta - smooth\ \forall i = 1, .., n$ (ce qui implique que $\mathcal{L}$ est $\lambda - smooth$ avec $\lambda \leq \beta$).
On considère l'algorithme SGD avec taille de batch $m = 1$ et pas constant $\eta > 0$:

$$\begin{aligned} x_0 &= x \in \mathbb{R}^n \\ x_{t+1} &= x_t - \eta\nabla l_{i_t}(x_t) \end{aligned}$$

où les $i_t \sim Unif(\{1, ..., n\})\ \forall t \in \mathbb{N}$ sont des variables aléatoires indépendantes.

2 Lemmes

On présente ici un lemme permettant "d'appliquer" l'inégalité $\mu - PL^*$ en espérance:

Lemma 1. *Soit* $t \in \mathbb{N}$.
Si $\mathbb{E}[||x_t - x_0|| \mid x_{t-1}] \leq \frac{R}{n}$*, alors* $\mathbb{E}[||\nabla\mathcal{L}(x_t)||^2 \mid x_{t-1}] \geq \mu\mathbb{E}[\mathcal{L}(x_t) \mid x_{t-1}]$

Proof. Posons $x_t^{(k)} = x_{t-1} - \eta \nabla l_k(x_{t-1})$.
Premièrement on a:

$$\begin{aligned}
\mathbb{E}[||x_t - x_0|| \mid x_{t-1}] &= \sum_{k=1}^{n} \mathbb{P}(i_t = k)||x_t^{(k)} - x_0|| \\
&= \frac{1}{n}\sum_{k=1}^{n} ||x_t^{(k)} - x_0|| \\
&\leq \frac{R}{n}
\end{aligned}$$

Ce qui implique que $||x_t^{(k)} - x_0|| \leq R \ \forall k = 1, ..., n$, donc $x_t^{(k)} \in \mathcal{B}(x_0, R) \ \forall k = 1, ..., n$.
Deuxièmement,

$$\begin{aligned}
\mathbb{E}[||\nabla \mathcal{L}(x_t)||^2 \mid x_{t-1}] &= \sum_{k=1}^{n} \mathbb{P}(i_t = k)||\nabla \mathcal{L}(x_t^{(k)})||^2 \\
&= \frac{1}{n}\sum_{k=1}^{n} ||\nabla \mathcal{L}(x_t^{(k)})||^2
\end{aligned}$$

Or $x_t^{(i)} \in \mathcal{B}(x_0, R) \ \forall i = 1, ..., n$ et $\mathcal{L}$ est $\mu - PL$ sur $\mathcal{B}(x_0, R)$. Donc

$$\begin{aligned}
\frac{1}{n}\sum_{k=1}^{n} ||\nabla \mathcal{L}(x_t^{(i)})||^2 &\geq \frac{\mu}{n}\sum_{k=1}^{n} \mathcal{L}(x_t^{(k)}) \\
&= \mu \mathbb{E}[\mathcal{L}(x_t) \mid x_{t-1}]
\end{aligned}$$

On obtient bien $\mathbb{E}[||\nabla \mathcal{L}(x_t)||^2 \mid x_{t-1}] \geq \mu \mathbb{E}[\mathcal{L}(x_t) \mid x_{t-1}]$. ☐

Lemma 2. *Si $\mathcal{L}$ est $\mu - PL$ sur $\mathcal{B}(x_0, \bar{R})$ où $\bar{R} := \frac{4n\sqrt{2\beta}\sqrt{\mathcal{L}(x_0)}}{\mu}$, alors $\forall t \in \mathbb{N}$ on a:*

$$\mathbb{E}[\mathcal{L}(x_t)] \leq (1 - \frac{\mu \eta^*}{2})^t \mathcal{L}(x_0) \text{ où } \eta^* = \frac{\mu}{2\lambda\beta} \tag{1}$$

$$\mathbb{E}[||x_t - x_0|| \mid x_{t-1}] \leq \frac{4\sqrt{2\beta}\sqrt{\mathcal{L}(x_0)}}{\mu} \tag{2}$$

Proof. On fait la preuve par récurrence.

Initialisation: Pour $t = 1$, on a:

$$\begin{aligned}
\mathbb{E}[||x_1 - x_0|| \mid x_0] &= \eta\mathbb{E}[||\nabla l_{i_0}(x_0)|| \mid x_0] \\
&= \eta\mathbb{E}[||\nabla l_{i_0}(x_0)||] \\
&\overset{l_i\ \beta\text{-smooth}}{\leq} \eta\mathbb{E}[\sqrt{2\beta l_{i_0}(x_0)}] \\
&\overset{\text{Jensen}}{\leq} \eta\sqrt{2\beta\mathbb{E}[l_{i_0}(x_0)]} \\
&= \eta\sqrt{2\beta\sum_{k=1}^{n}\underbrace{\mathbb{P}(i_0 = k)}_{=\frac{1}{n}} l_k(x_0)} \\
&= \eta\sqrt{2\beta\mathcal{L}(x_0)} \\
&\leq \frac{4\sqrt{2\beta\mathcal{L}(x_0)}}{\mu} \qquad (3)
\end{aligned}$$

on a (3) si $\eta \leq \frac{4}{\mu}$, et dans ce cas (2) est vraie au rang $t = 1$.
De plus, par λ-smoothness de $\mathcal{L}$, on a:

$$\begin{aligned}
\mathcal{L}(x_1) &\leq \mathcal{L}(x_0) + \langle\nabla\mathcal{L}(x_0), x_1 - x_0\rangle + \frac{\lambda}{2}||x_1 - x_0||^2 \\
\Leftrightarrow \mathcal{L}(x_0) - \mathcal{L}(x_1) &\geq \eta\langle\nabla\mathcal{L}(x_0), \nabla l_{i_0}(x_0)\rangle - \eta^2\frac{\lambda}{2}||\nabla l_{i_0}(x_0)||^2 \\
\Rightarrow \mathbb{E}[\mathcal{L}(x_0) - \mathcal{L}(x_1) \mid x_0] &\geq \eta\underbrace{||\nabla\mathcal{L}(x_0)||^2}_{\geq\mu\mathcal{L}(x_0)} - \eta^2\frac{\lambda}{2}\mathbb{E}[\underbrace{||\nabla l_{i_0}(x_0)||^2}_{\leq 2\beta l_{i_0}(x_0)} \mid x_0] \\
&\geq \eta\mu\mathcal{L}(x_0) - \eta^2\frac{\lambda}{2}\mathbb{E}[2\beta l_{i_1}(x_0) \mid x_0]
\end{aligned}$$

Or $\mathbb{E}[2\beta l_{i_1}(x_0) \mid x_0] = \mathbb{E}[2\beta l_{i_1}(x_0)] = 2\beta\frac{1}{n}\sum_{k=1}^{n} l_k x_0 = 2\beta\mathcal{L}(x_0)$. On obtient:

$$\begin{aligned}
\mathbb{E}[\mathcal{L}(x_0) - \mathcal{L}(x_1) \mid x_0] &\geq \eta\mu\mathcal{L}(x_0) - \eta^2\frac{\lambda}{2}2\beta\mathcal{L}(x_0) \\
&= (\eta\mu - \eta^2\lambda\beta)\mathcal{L}(x_0) \\
\Leftrightarrow \mathbb{E}[\mathcal{L}(x_1) \mid x_0] &\leq (-\eta\mu + \eta^2\lambda\beta)\mathcal{L}(x_0) + \underbrace{\mathbb{E}[\mathcal{L}(x_0) \mid x_0]}_{=\mathcal{L}(x_0)} \\
&= (1 - \eta\mu + \eta^2\lambda\beta)\mathcal{L}(x_0)
\end{aligned}$$

En optimisant $(1 - \eta\mu + \eta^2\lambda\beta)$, on obtient $\mathbb{E}[\mathcal{L}(x_1) \mid x_0] \leq (1 - \frac{\mu\eta^*}{2})\mathcal{L}(x_0)$. Comme $\mathbb{E}[\mathcal{L}(x_1) \mid x_0] = \mathbb{E}[\mathcal{L}(x_1)]$, on a (1) au rang $t = 1$.

Hérédité: On pose l'hypthèse de récurrence (HR) suivante: soit $t \in \mathbb{N}$ tel que $\forall\ t' \leq t$, $\mathbb{E}[||x_{t'} - x_0|| \mid x_{t'-1}] \leq \frac{\bar{R}}{n}$ et $\mathbb{E}[\mathcal{L}(x_{t'})] \leq (1 - \frac{\mu\eta^*}{2})\mathcal{L}(x_0)$.

Par λ-smoothness de $\mathcal{L}$:

$$\mathcal{L}(x_{t+1}) \leq \mathcal{L}(x_t) + \langle \nabla \mathcal{L}(x_t), x_{t+1} - x_t \rangle + \frac{\lambda}{2} ||x_{t+1} - x_t||^2$$
$$\Leftrightarrow \mathcal{L}(x_t) - \mathcal{L}(x_{t+1}) \geq \eta \langle \nabla \mathcal{L}(x_t), \nabla l_{i_t}(x_t) \rangle - \eta^2 \frac{\lambda}{2} ||\nabla l_{i_t}(x_t)||^2$$
$$\Rightarrow \mathbb{E}[\mathcal{L}(x_t) - \mathcal{L}(x_{t+1}) \mid x_t] \geq \eta ||\nabla \mathcal{L}(x_t)||^2 \underbrace{-\eta^2 \frac{\lambda}{2} \mathbb{E}[\underbrace{||\nabla l_{i_t}(x_t)||^2}_{\leq 2\beta l_{i_t}(x_t)} \mid x_t]}_{\geq -\eta^2 \lambda \beta \mathcal{L}(x_t)}$$
$$\geq \eta ||\nabla \mathcal{L}(x_t)||^2 - \eta^2 \lambda \beta \mathcal{L}(x_t)$$

Comme $\sigma(x_{t-1}) \subset \sigma(x_t)$, on a:

$$\mathbb{E}[\mathcal{L}(x_t) - \mathcal{L}(x_{t+1}) \mid x_{t-1}] = \mathbb{E}[\mathbb{E}[\mathcal{L}(x_t) - \mathcal{L}(x_{t+1}) \mid x_t] \mid x_{t-1}]$$
$$\geq \mathbb{E}[||\nabla \mathcal{L}(x_t)||^2 \mid x_{t-1}] - \eta^2 \lambda \beta \mathbb{E}[\mathcal{L}(x_t) \mid x_{t-1}]$$

Par le lemme 1 et (HR), $\mathbb{E}[||\nabla \mathcal{L}(x_t)||^2 \mid x_{t-1}] \geq \mu \mathbb{E}[\mathcal{L}(x_t) \mid x_{t-1}]$, donc:

$$\mathbb{E}[\mathcal{L}(x_t) - \mathcal{L}(x_{t+1}) \mid x_{t-1}] \geq \mu \mathbb{E}[\mathcal{L}(x_t) \mid x_{t-1}] - \eta^2 \lambda \beta \mathbb{E}[\mathcal{L}(x_t) \mid x_{t-1}]$$
$$\Rightarrow \mathbb{E}[\mathcal{L}(x_{t+1})] \leq -\eta \mu \mathbb{E}[\mathcal{L}(x_t)] + \eta^2 \lambda \beta \mathbb{E}[\mathcal{L}(x_t)] + \mathbb{E}[\mathcal{L}(x_t)]$$
$$= (1 - \eta\mu + \eta^2 \lambda \beta) \mathbb{E}[\mathcal{L}(x_t)]$$

Par (HR), et en optimisant $(1 - \eta\mu + \eta^2 \lambda\beta)$, on obtient:

$$\mathbb{E}[\mathcal{L}(x_{t+1})] \leq (1 - \frac{\mu \eta^*}{2})^{t+1} \mathcal{L}(x_0)$$

On a donc (1) au rang $t + 1$.
Il nous reste maintenant à prouver qu'on a (2). La preuve bloque à cet endroit.
$\square$