Optimization for Data Science

Stéphane Gaïffas

November 8, 2016

Abstract

In this short note we give basic convergence proofs for stochastic gradient descent, using classical arguments from literature.

1 Stochastic Gradient Descent (SGD)

We want to minimize

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

keeping in mind the main example

$$f_i(x) = \ell(b_i, \langle a_i, x \rangle) + \frac{\lambda}{2} \|x\|_2^2.$$

Introduce the ball $B = \{x \in \mathbb{R}^d : \|x\| \leq r\}$. We'll restrict the iterateds of SGD inside this ball.

We study in this section the Stochastic Gradient Algorithm (SGD) that procedes uses the following iteration

$$x_t = \mathrm{proj}_B(x_{t-1} - \eta_t \nabla f_{i_t}(x_{t-1})) \tag{1}$$

where at each iteration t, we sample i_t uniformly in $\{1, \ldots, n\}$, so that the sequence $(i_t)_t$ is i.i.d.

Let us stress that in this note, $\nabla f_i(x)$ will stand for any subgradient of the subdifferential $\partial f_i(x)$ of f_i. It is indeed not required that the f_i are differentiable, but only that the subgradients of all f_i are bounded by some constant.

We denote by $\mathcal{F}_t$ the minimal $\sigma-$field that makes $i_1, \ldots, i_t$ measurable. We need the following properties on the conditional expectation:

$$\mathbb{E}[\nabla f_{i_t}(x_{t-1}) | \mathcal{F}_{t-1}] = \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(x_{t-1}) = \nabla f(x_{t-1}) \tag{2}$$

and the *chain rule* of conditional expectation, that says that

$$\mathbb{E}[\mathbb{E}[\cdot | \mathcal{F}_{t-1}]] = \mathbb{E}[\cdot]. \tag{3}$$

Theorem 1.1. *Consider* (x_t) *a sequence given by* (1) *with* $\eta_t = \frac{2r}{b\sqrt{t}}$. *Assume that* f *is convex, that* $\|\nabla f_i(x)\| \leq b$ *for any* $i = 1, \ldots, n$, *any* $x \in B$ *and any* $\nabla f_i(x) \in \partial f_i(x)$. *Furthermore, assume that any* $x_* \in \operatorname{argmin}_{x \in \mathbb{R}^d} f(x)$ *belongs to* B. *Then, the following inequality holds*

$$\mathbb{E} f\Big(\frac{1}{t}\sum_{s=0}^{t-1} x_s\Big) - f(x_*) \leq \frac{3rb}{\sqrt{t}}.$$

This proves that the convergence rate of averaged SGD is $O(1/\sqrt{t})$ under under convexity, when subgradients are bounded (once again, no differentiability or L-smoothness is required here).

Proof. Write

$$\begin{aligned}
\|x_t - x_*\|^2 &= \|\operatorname{proj}_B(x_{t-1} - \eta_t \nabla f_{i_t}(x_{t-1})) - x_*\|^2 \\
&= \|\operatorname{proj}_B(x_{t-1} - \eta_t \nabla f_{i_t}(x_{t-1})) - \operatorname{proj}_B(x_*)\|^2 \\
&\leq \|x_{t-1} - \eta_t \nabla f_{i_t}(x_{t-1}) - x_*\|^2 \\
&= \|x_{t-1} - x_*\|^2 + \eta_t^2 \|\nabla f_{i_t}(x_{t-1})\|^2 - 2\eta_t \langle \nabla f_{i_t}(x_{t-1}), x_{t-1} - x_* \rangle.
\end{aligned}$$

In the first line, we used Equation (1), in the second we used that by assumption $x_* \in B$, for the third we used the fact that $\|\operatorname{proj}_B(x) - \operatorname{proj}_B(y)\| \leq \|x - y\|$ and the last line is simple algebra. We assumed that $\|\nabla f_{i_t}(x_{t-1})\|^2 \leq b^2$, so we arrive at

$$\|x_t - x_*\|^2 \leq \|x_{t-1} - x_*\|^2 + \eta_t^2 b^2 - 2\eta_t \langle \nabla f_{i_t}(x_{t-1}), x_{t-1} - x_* \rangle.$$

Now, taking the conditional expectation $\mathbb{E}[\cdot|\mathcal{F}_{t-1}]$ on both sides leads to

$$\begin{aligned}
\mathbb{E}[\|x_t - x_*\|^2 | \mathcal{F}_{t-1}] &\leq \|x_{t-1} - x_*\|^2 + \eta_t^2 b^2 - 2\eta_t \mathbb{E}[\langle \nabla f_{i_t}(x_{t-1}), x_{t-1} - x_* \rangle | \mathcal{F}_{t-1}] \\
&= \|x_{t-1} - x_*\|^2 + \eta_t^2 b^2 - 2\eta_t \langle \mathbb{E}[\nabla f_{i_t}(x_{t-1}) | \mathcal{F}_{t-1}], x_{t-1} - x_* \rangle \\
&= \|x_{t-1} - x_*\|^2 + \eta_t^2 b^2 - 2\eta_t \langle \nabla f(x_{t-1}), x_{t-1} - x_* \rangle.
\end{aligned}$$

In the first line, we used the fact that x_{t-1} is $\mathcal{F}_{t-1}$-measurable, on the second line we used linearity of the conditional expectation and we used Equation (2) in the third line.

By convexity of f and by definition of the subdifferential, we have that

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

for any $x, y \in \mathbb{R}^d$ and any $\nabla f(x) \in \partial f(x)$. This entails

$$\mathbb{E}[\|x_t - x_*\|^2 | \mathcal{F}_{t-1}] \leq \|x_{t-1} - x_*\|^2 + \eta_t^2 b^2 - 2\eta_t (f(x_{t-1}) - f(x_*)).$$

Now, taking the expectation $\mathbb{E}[\cdot]$ on both sides, we obtain using Equation (3) that

$$\mathbb{E}\|x_t - x_*\|^2 \leq \mathbb{E}\|x_{t-1} - x_*\|^2 + \eta_t^2 b^2 - 2\eta_t (\mathbb{E} f(x_{t-1}) - f(x_*)).$$

Now, simple algebra leads to

$$\mathbb{E} f(x_{t-1}) - f(x_*) \leq \frac{1}{2\eta_t}(\mathbb{E}\|x_{t-1} - x_*\|^2 - \mathbb{E}\|x_t - x_*\|^2) + \frac{b^2 \eta_t}{2}.$$

Let us now consider the following sum

$$\begin{aligned}
\sum_{s=1}^{t}(\mathbb{E}f(x_{s-1})-f(x_*)) &\le \frac{1}{2}\sum_{s=1}^{t}\frac{1}{\eta_s}(\mathbb{E}\|x_{s-1}-x_*\|^2-\mathbb{E}\|x_s-x_*\|^2)+\frac{b^2}{2}\sum_{s=1}^{t}\eta_s \\
&= \frac{1}{2}\Big(\frac{1}{\eta_1}\mathbb{E}\|x_0-x_*\|^2-\frac{1}{\eta_t}\mathbb{E}\|x_t-x_*\|^2\Big) \\
&\quad +\frac{1}{2}\sum_{s=1}^{t-1}\Big(\frac{1}{\eta_{s+1}}-\frac{1}{\eta_s}\Big)\mathbb{E}\|x_s-x_*\|^2+\frac{b^2}{2}\sum_{s=1}^{t}\eta_s \\
&\le \frac{2r^2}{\eta_1}+2r^2\sum_{s=1}^{t-1}\Big(\frac{1}{\eta_{s+1}}-\frac{1}{\eta_s}\Big)+\frac{b^2}{2}\sum_{s=1}^{t}\eta_s \\
&\le \frac{2r^2}{\eta_1}+\frac{2r^2}{\eta_t}+\frac{b^2}{2}\sum_{s=1}^{t}\eta_s.
\end{aligned}$$

We used simple algebra in the first and second lines, and the fact that $\|x_t-x_*\|\le 2r$ for any t in the third and last lines (since x_t and x_* belong to B).

Now, we can conclude the proof by noticing that $\sum_{s=1}^{t}\frac{1}{\sqrt{s}}\le 2\sqrt{t}-1$ (this comes by induction, using the following trick $\frac{1}{\sqrt{s}}\le\frac{2}{\sqrt{s}+\sqrt{s-1}}=2(\sqrt{s}-\sqrt{s-1})$), so that

$$\sum_{s=1}^{t}(\mathbb{E}f(x_{s-1})-f(x_*))\le 3rb\sqrt{t}$$

and using the convexity of f yields

$$\mathbb{E}f\Big(\frac{1}{t}\sum_{s=0}^{t-1}x_s\Big)-f(x_*)\le\frac{1}{t}\sum_{s=0}^{t-1}(\mathbb{E}f(x_s)-f(x_*))\le\frac{3rb}{\sqrt{t}},$$

which concludes the proof of Theorem 1.1. ∎

Under strong convexity, the rate is better, as described in the next Theorem.

Theorem 1.2. *Assume the same as in Theorem 1.1, and assume that f is μ-strongly convex. If (x_t) is a sequence given by* (1) *with $\eta_t=\frac{2}{\mu(t+1)}$, we have*

$$\mathbb{E}f\Big(\frac{2}{t(t+1)}\sum_{s=1}^{t}sx_{s-1}\Big)-f(x_*)\le\frac{2b^2}{\mu(t+1)}.$$

Proof. We start similarly as in the proof of Theorem 1.1 and get

$$\mathbb{E}[\|x_t-x_*\|^2|\mathcal{F}_{t-1}]\le\|x_{t-1}-x_*\|^2+\eta_t^2b^2-2\eta_t\langle\nabla f(x_{t-1}),x_{t-1}-x_*\rangle.$$

But now, we use the fact that since f is μ-strongly convex, we have

$$f(y)-f(x)\ge\langle\nabla f(x),y-x\rangle+\frac{\mu}{2}\|y-x\|^2$$

for any $x, y \in \mathbb{R}^d$ and any $\nabla f(x) \in \partial f(x)$. This entails using some simple algebra, and taking the expectation $\mathbb{E}[\cdot]$ on both sides, that

$$\begin{aligned}\mathbb{E}f(x_{t-1}) - f(x_*) &\leq \frac{1}{2}\Big(\frac{1}{\eta_t} - \mu\Big)\mathbb{E}\|x_{t-1} - x_*\|^2 - \frac{1}{2\eta_t}\mathbb{E}\|x_t - x_*\|^2 + \frac{b^2\eta_t}{2} \\ &= \frac{\mu(t-1)}{4}\mathbb{E}\|x_{t-1} - x_*\|^2 - \frac{\mu(t+1)}{4}\mathbb{E}\|x_t - x_*\|^2 + \frac{b^2}{\mu(t+1)}.\end{aligned}$$

Now, write

$$\begin{aligned}\sum_{s=1}^{t} s\mathbb{E}(f(x_{s-1}) - f(x_*)) &= \frac{\mu}{4}\sum_{s=1}^{t}\Big((s-1)s\mathbb{E}\|x_{s-1} - x_*\|^2 - s(s+1)\mathbb{E}\|x_s - x_*\|^2\Big) \\ &\quad + \frac{b^2}{\mu}t \\ &\leq \frac{b^2}{\mu}t,\end{aligned}$$

and by convexity of f, we obtain

$$\mathbb{E}f\Big(\frac{2}{t(t+1)}\sum_{s=1}^{t} s x_{s-1}\Big) - f(x_*) \leq \frac{2}{t(t+1)}\sum_{s=1}^{t} s(\mathbb{E}f(x_{s-1}) - f(x_*)) \leq \frac{2b^2}{\mu(t+1)}$$

which concludes the proof of the theorem. ■