Stochastic gradient descent

Giulio Galvan

26 marzo 2016

Consider the stochastic optimization problem

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} f(\boldsymbol{x}) = \mathbb{E}[\mathbb{F}(\boldsymbol{x}, \boldsymbol{\xi})], \tag{1}$$

where $\boldsymbol{\xi} \in \Omega \subset \mathbb{R}^d$ is a random vector. Suppose $f(\cdot)$ is continuous, strongly convex (with constant c) and there exists a compact level set of $f(\cdot)$, hence (1) has a unique optimal solution $\boldsymbol{x}_*$. We make the following two assumptions:

- It is possible to generate independent identically distributed samples of $\boldsymbol{\xi}$.
- There exists an oracle which, for a given point $(\boldsymbol{x}, \boldsymbol{\xi})$ returns a stochastic direction $D(\boldsymbol{x}, \boldsymbol{\xi})$ such that $d(\boldsymbol{x}) \triangleq \mathbb{E}[D(\boldsymbol{x}, \boldsymbol{\xi})]$ satisfies:

$$-(\boldsymbol{x} - \boldsymbol{x}_*)^T (f' - d(\boldsymbol{x})) \geq -\mu L \left\| \boldsymbol{x} - \boldsymbol{x}_* \right\|_2^2 \quad \text{for some } f' \in \partial f(\boldsymbol{x}), \tag{2}$$

for some $\mu \in (0, \frac{c}{L})$, L is some chosen positive constant. We assume further that there exists $M > 0$ such that

$$\| d(\boldsymbol{x}) \|_2^2] \leq M^2 \quad \forall \boldsymbol{x} \in \mathbb{R}^n. \tag{3}$$

Consider an algorithm defined by

$$\boldsymbol{x}_{j+1} = \boldsymbol{x}_j - \gamma_j D(\boldsymbol{x}_j, \boldsymbol{\xi}_j). \tag{4}$$

Each iterate $\boldsymbol{x}_j$ of such a random process is a function of the history $\boldsymbol{\xi}_{[j-1]} = (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_{j-1})$.

Let $A_j \triangleq \| \boldsymbol{x}_j - \boldsymbol{x}_* \|_2^2$ and $a_j \triangleq \mathbb{E}[A_j]$. From (4) we get

$$\begin{aligned} A_{j+1} &= \frac{1}{2} \left\| \boldsymbol{x}_j - \gamma_j D(\boldsymbol{x}_j, \boldsymbol{\xi}_j) - \boldsymbol{x}_* \right\|_2^2 \\ &= A_j + \frac{1}{2} \gamma_j^2 \left\| D(\boldsymbol{x}_j, \boldsymbol{\xi}_j) \right\|_2^2 - \gamma_j (\boldsymbol{x}_j - \boldsymbol{x}_*)^T D(\boldsymbol{x}_j, \boldsymbol{\xi}_j). \end{aligned} \tag{5}$$

We can write:

$$\mathbb{E}_{\boldsymbol{\xi}_{[j]}}[(\boldsymbol{x}_j - \boldsymbol{x}_*)^T D(\boldsymbol{x}_j, \boldsymbol{\xi}_j)] = \mathbb{E}_{\boldsymbol{\xi}_{[j-1]}}[\mathbb{E}_{\boldsymbol{\xi}_{[j]}}[(\boldsymbol{x}_j - \boldsymbol{x}_*)^T D(\boldsymbol{x}_j, \boldsymbol{\xi}_j)] | \boldsymbol{\xi}_{[j-1]}] \tag{6}$$

$$= \mathbb{E}_{\boldsymbol{\xi}_{[j-1]}}[(\boldsymbol{x}_j - \boldsymbol{x}_*)^T \mathbb{E}_{\boldsymbol{\xi}_{[j]}}[D(\boldsymbol{x}_j, \boldsymbol{\xi}_j)] | \boldsymbol{\xi}_{[j-1]}] \tag{7}$$

$$= \mathbb{E}_{\boldsymbol{\xi}_{[j-1]}}[(\boldsymbol{x}_j - \boldsymbol{x}_*)^T d(\boldsymbol{x}_j)]. \tag{8}$$

Equation (6) is given by the law of total expectation, (7) holds because $\boldsymbol{x}_j = \boldsymbol{x}_j(\boldsymbol{\xi_{[j-1]}})$ is not function of $\boldsymbol{\xi}_j$, hence independent of it. Using (3) and (8) we obtain, taking expectation on both sides of (5)

$$a_{j+1} \leq a_j - \gamma_j \mathbb{E}_{\boldsymbol{\xi}_{[j-1]}}[(\boldsymbol{x}_j - \boldsymbol{x}_*)^T d(\boldsymbol{x}_j)] + \frac{1}{2}\gamma_j^2 M^2. \tag{9}$$

Since $f(\cdot)$ is strongly convex with constant $c > 0$,

$$(\boldsymbol{x} - \boldsymbol{y})^T (f' - g')) \geq c \left\|\boldsymbol{x} - \boldsymbol{y}\right\|_2^2, \quad \forall f' \in \partial f(\boldsymbol{x}), g' \in \partial f(\boldsymbol{y}). \tag{10}$$

By optimality of $\boldsymbol{x}_*$ we have

$$(\boldsymbol{x} - \boldsymbol{x}_*)^T f' \geq 0 \quad \forall \boldsymbol{x} \in \mathbb{R}^n, \forall f' \in \partial f(\boldsymbol{x}_*). \tag{11}$$

Inequalities (10) and (11) together imply

$$(\boldsymbol{x} - \boldsymbol{x}_*)^T f' \geq c \left\|\boldsymbol{x} - \boldsymbol{x}_*\right\|_2^2 \quad \forall \boldsymbol{x} \in \mathbb{R}^n, \forall f' \in \partial f(\boldsymbol{x}). \tag{12}$$

Adding and subtracting the oracle direction $d(\boldsymbol{x})$ we get

$$(\boldsymbol{x} - \boldsymbol{x}_*)^T (f' - d(\boldsymbol{x}) + d(\boldsymbol{x})) \geq c \left\|\boldsymbol{x} - \boldsymbol{x}_*\right\|_2^2, \tag{13}$$

which can be rewritten as

$$(\boldsymbol{x} - \boldsymbol{x}_*)^T d(\boldsymbol{x}) \geq c \left\|\boldsymbol{x} - \boldsymbol{x}_*\right\|_2^2 - (\boldsymbol{x} - \boldsymbol{x}_*)^T (f' - d(\boldsymbol{x})). \tag{14}$$

From Assumption (2), and by taking expectations (from now on we will write $\mathbb{E}$ in place of $\mathbb{E}_{\boldsymbol{\xi}_{[j-1]}}$ for ease of notation) on both side of (14), we obtain

$$\begin{aligned}
\mathbb{E}[(\boldsymbol{x}_j - \boldsymbol{x}_*)^T (\boldsymbol{x}_j)] &\geq c\mathbb{E}[\|\boldsymbol{x}_j - \boldsymbol{x}_*\|_2^2)] - \mathbb{E}[(\boldsymbol{x}_j - \boldsymbol{x}_*)^T (f_j' - d(\boldsymbol{x}_j))] && (15)\\
&\geq c\left(1 - \frac{\mu L}{c}\right)\mathbb{E}[\|\boldsymbol{x}_j - \boldsymbol{x}_*\|_2^2] && (16)\\
&= 2\bar{c} a_j, && (17)
\end{aligned}$$

with $\bar{c} = c(1 - \frac{\mu L}{c})$ and $f_j' \in \partial f(\boldsymbol{x}_j)$. Hence from (9) it follows

$$a_{j+1} \leq (1 - 2\bar{c}\gamma_j) a_j + \frac{1}{2}\gamma_j^2 M^2. \tag{18}$$

Choosing now the stepsizes as $\gamma_j = \frac{\beta}{j}$ for some constant $\beta > \frac{1}{2\bar{c}}$ we get

$$a_{j+1} \leq (1 - 2\bar{c}\gamma_j) a_j + \frac{1}{2}\frac{\beta^2 M^2}{j^2}. \tag{19}$$

It follows by induction that

$$\mathbb{E}[\|\boldsymbol{x}_j - \boldsymbol{x}_*\|_2^2] = 2a_j \leq \frac{Q(\beta)}{j}, \tag{20}$$

where

$$Q(\beta) = \max \left\{ \frac{\beta^2 M^2}{2\bar{c} - 1}, \|\boldsymbol{x}_1 - \boldsymbol{x}_*\|_2^2 \right\}. \tag{21}$$

When ∇f is Lipschitz continuous we also have

$$f(\boldsymbol{x}) \leq f(\boldsymbol{x}_*) + \frac{1}{2} L \|\boldsymbol{x} - \boldsymbol{x}_*\|_2^2, \quad \forall \boldsymbol{x} \in \mathbb{R}^n, \tag{22}$$

hence we can get a bound also on the function value:

$$\mathbb{E}[f(\boldsymbol{x}_j) - f(\boldsymbol{x}_*)] \leq \frac{1}{2} L \, \mathbb{E}[\|\boldsymbol{x}_j - \boldsymbol{x}_*\|_2^2] \leq \frac{1}{2} L \, Q(\beta). \tag{23}$$

Assumption (2) can be further elaborated. Let θ be the angle between $f' \in \partial f(\boldsymbol{x})$ and $g(\boldsymbol{x})$. Write $\|g(\boldsymbol{x}_j)\| = \alpha \|\nabla f(\boldsymbol{x}_j)\|$ for some $\alpha > 0$, then

$$\begin{aligned} \|f' - g(\boldsymbol{x}_j)\|^2 &= \|f'\|^2 + \|g(\boldsymbol{x}_j)\|^2 - 2\|f'\| \|g(\boldsymbol{x}_j)\| \cos\theta_j & (24) \\ &= \|f'\|^2 (1 + \alpha_j^2 - 2\alpha_j \cos\theta_j). & (25) \end{aligned}$$

Hence

$$\begin{aligned} (\boldsymbol{x} - \boldsymbol{x}_*)^T (f' - g(\boldsymbol{x})) &\leq \|\boldsymbol{x} - \boldsymbol{x}_*\| \|f' - g(\boldsymbol{x})\| & (26) \\ &= \|\boldsymbol{x} - \boldsymbol{x}_*\| \|f'\| (1 + \alpha_j^2 - 2\alpha_j \cos\theta_j)^{\frac{1}{2}} & (27) \end{aligned}$$

Assume

$$\|f'\|_2 \leq L \|\boldsymbol{x} - \boldsymbol{x}_*\|_2. \tag{28}$$

Note that Equation (28) is implied simply by Lipschitz continuity in the differentiable case. A sufficient condition is thus

$$1 + \alpha^2 - 2\alpha\cos\theta_j \leq \left(\frac{\mu}{L}\right)^2. \tag{29}$$