Stochastic gradient descent

Giulio Galvan

15 dicembre 2015

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 and there exists a compact level set of $f(\cdot)$, hence (1) has a unique optimal solution $\boldsymbol{x}_*$. Let also L be the Lipschitz constant of $\nabla f(\cdot)$. 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 $(\boldsymbol{x},\boldsymbol{\xi})$ return a stochastic descent direction $D(\boldsymbol{x},\boldsymbol{\xi})$ such that $d(\boldsymbol{x}) \triangleq \mathbb{E}[D(\boldsymbol{x},\boldsymbol{\xi})]$ satisfy:

$$-(\boldsymbol{x}-\boldsymbol{x}_*)^T(\nabla f(\boldsymbol{x}) - g(\boldsymbol{x})) \geq -\mu L \left\|\boldsymbol{x}_j - \boldsymbol{x}_*\right\|_2^2 \quad \forall \boldsymbol{x} \in \mathbb{R}^n, \tag{2}$$

  for some $\mu \in (0,1)$.

Consider the algorithm defined by

$$\boldsymbol{x}_{j+1} = \boldsymbol{x}_j - \gamma_j D(\boldsymbol{x}_j, \boldsymbol{\xi}_j). \tag{3}$$

Let $A_j \triangleq \left\|\boldsymbol{x}_j - \boldsymbol{x}_*\right\|_2^2$ and $a_j \triangleq \mathbb{E}[A_j]$. From (3) we get

$$\begin{aligned} A_{j+1} &= \frac{1}{2}\left\|\boldsymbol{x}_j - \gamma_j D(\boldsymbol{x}_j,\boldsymbol{\xi}_j) - \boldsymbol{x}_*\right\|_2^2 \\ &= A_j + \frac{1}{2}\gamma_j^2 \left\|D(\boldsymbol{x}_j,\boldsymbol{\xi}_j)\right\|_2^2 - \gamma_j(\boldsymbol{x}_j - \boldsymbol{x}_*)^T D(\boldsymbol{x}_j,\boldsymbol{\xi}_j). \end{aligned} \tag{4}$$

Since $\boldsymbol{x}_j = \boldsymbol{x}_j(\boldsymbol{\xi_{[j-1]}})$ is independent of $\boldsymbol{\xi}_j$ we have

$$\begin{aligned} \mathbb{E}[(\boldsymbol{x}_j - \boldsymbol{x}_*)^T D(\boldsymbol{x}_j,\boldsymbol{\xi}_j)] &= \mathbb{E}_{\boldsymbol{\xi}_{[j-1]}}[\mathbb{E}_{\boldsymbol{\xi}_j}[(\boldsymbol{x}_j - \boldsymbol{x}_*)^T D(\boldsymbol{x}_j,\boldsymbol{\xi}_j)]|\boldsymbol{\xi}_{[j-1]}] \\ &= \mathbb{E}_{\boldsymbol{\xi}_{[j-1]}}[(\boldsymbol{x}_j - \boldsymbol{x}_*)^T \mathbb{E}_{\boldsymbol{\xi}_j}[D(\boldsymbol{x}_j,\boldsymbol{\xi}_j)]|\boldsymbol{\xi}_{[j-1]}] \\ &= \mathbb{E}_{\boldsymbol{\xi}_{[j-1]}}[(\boldsymbol{x}_j - \boldsymbol{x}_*)^T d(\boldsymbol{x}_j)] \end{aligned} \tag{5}$$

Let now assume that there exists $M > 0$ such that

$$\mathbb{E}[D(\boldsymbol{x},\boldsymbol{\xi})] \leq M^2 \quad \forall \boldsymbol{x} \in \mathbb{R}^n. \tag{6}$$

Using (5) and 6 we obtain, taking expectation of both sides of (4)

$$a_{j+1} \leq a_j - \gamma_j \mathbb{E}_{\boldsymbol{\xi}_{[j-1]}}[(\boldsymbol{x}_j - \boldsymbol{x}_*)^T d(\boldsymbol{x}_j)] + \frac{1}{2}\gamma_j^2 M^2 \tag{7}$$

Since $f(\cdot)$ is strongly convex there exists $c > 0$ such that

$$(\boldsymbol{y} - \boldsymbol{x})^T (\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x})) \geq c \left\| \boldsymbol{y} - \boldsymbol{x} \right\|_2^2 \tag{8}$$

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

$$(\boldsymbol{x} - \boldsymbol{x}_*)^T \nabla f(\boldsymbol{x}_*) \geq 0 \quad \boldsymbol{x} \in \mathbb{R}^n. \tag{9}$$

Inequalities (9) and (8) implies

$$(\boldsymbol{x} - \boldsymbol{x}_*)^T \nabla f(\boldsymbol{x}) \geq c \left\| \boldsymbol{x} - \boldsymbol{x}_* \right\|_2^2 \quad \boldsymbol{x} \in \mathbb{R}^n. \tag{10}$$

Adding and subtracting the descent direction $g(\boldsymbol{x})$ we get

$$(\boldsymbol{x} - \boldsymbol{x}_*)^T (\nabla f(\boldsymbol{x}) - g(\boldsymbol{x}) + g(\boldsymbol{x})) \geq c \left\| \boldsymbol{x} - \boldsymbol{x}_* \right\|_2^2, \tag{11}$$

which can be rewritten as

$$(\boldsymbol{x} - \boldsymbol{x}_*)^T g(\boldsymbol{x}) \geq c \left\| \boldsymbol{x} - \boldsymbol{x}_* \right\|_2^2 - (\boldsymbol{x} - \boldsymbol{x}_*)^T (\nabla f(\boldsymbol{x}) - g(\boldsymbol{x})) \tag{12}$$

From (2), taking expectations of both side of (12) we obtain

$$\begin{aligned}
\mathbb{E}[(\boldsymbol{x}_j - \boldsymbol{x}_*)^T g(\boldsymbol{x}_j)] &\geq c\mathbb{E}[\|\boldsymbol{x}_j - \boldsymbol{x}_*\|_2^2)] - \mathbb{E}[(\boldsymbol{x}_j - \boldsymbol{x}_*)^T (\nabla f(\boldsymbol{x}_j) - g(\boldsymbol{x}_j))] && (13) \\
&\geq c(1 - \frac{\mu L}{c})\mathbb{E}[\|\boldsymbol{x}_j - \boldsymbol{x}_*\|_2^2)] && (14) \\
&= 2\bar{c}a_j, && (15)
\end{aligned}$$

with $\bar{c} = c(1 - \frac{\mu L}{c})$. Hence from (7) follows

$$a_{j+1} \leq (1 - 2\bar{c}\gamma_j)a_j + \frac{1}{2}\gamma_j 2M^2. \tag{16}$$

From here on the proof follows as in the standard case with $c(1 - \frac{\mu L}{c})$ in place of c.

Descent direction condition Condition 2 can be further elaborated. Let $\cos\theta_j$ be the angle between $\nabla f(\boldsymbol{x}_j)$ and $g(\boldsymbol{x}_j)$ and $\|g(\boldsymbol{x}_j)\| = \alpha_j \|\nabla f(\boldsymbol{x}_j)\|$ for some $\alpha_j > 0$. Then,

$$\begin{aligned}
\|\nabla f(\boldsymbol{x}_j) - g(\boldsymbol{x}_j)\|^2 &= \|\nabla f(\boldsymbol{x}_j)\|^2 + \|g(\boldsymbol{x}_j)\|^2 - 2\|\nabla f(\boldsymbol{x}_j)\| \|g(\boldsymbol{x}_j)\| \cos\theta_j && (17) \\
&= \|\nabla f(\boldsymbol{x}_j)\|^2 (1 + \alpha_j^2 - 2\alpha_j \cos\theta_j). && (18)
\end{aligned}$$

Since $\nabla f(\boldsymbol{x}_*) = 0$, using Lipschitz continuity of ∇f (with constant L) we get

$$\|\nabla f(\boldsymbol{x}_j) - g(\boldsymbol{x}_j)\| \leq L \|\boldsymbol{x}_j - \boldsymbol{x}_*\| (1 + \alpha_j^2 - 2\alpha_j \cos\theta_j)^{\frac{1}{2}} \tag{19}$$

Hence

$$(\boldsymbol{x}_j - \boldsymbol{x}_*)^T (\nabla f(\boldsymbol{x}_j) - g(\boldsymbol{x}_j)) \leq \|\boldsymbol{x}_j - \boldsymbol{x}_*\| \, \|\nabla f(\boldsymbol{x}_j) - g(\boldsymbol{x}_j)\| \tag{20}$$

$$\leq L \, \|\boldsymbol{x}_j - \boldsymbol{x}_*\|^2 \, (1 + \alpha_j^2 - 2\alpha_j \cos \theta_j)^{\frac{1}{2}}. \tag{21}$$

Hence condition **??** becomes

$$(1 + \alpha_j^2 - 2\alpha_j \cos \theta_j)^{\frac{1}{2}} \geq \mu \tag{22}$$