

# Stochastic gradient descent

Giulio Galvan

15 gennaio 2016

Consider the stochastic optimization problem

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

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  $\mathbf{x}_*$ . Let also  $L$  be the Lipschitz constant of  $\nabla f$ . 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  $(\mathbf{x}, \boldsymbol{\xi})$  return a stochastic direction  $D(\mathbf{x}, \boldsymbol{\xi})$  such that  $d(\mathbf{x}) \triangleq \mathbb{E}[D(\mathbf{x}, \boldsymbol{\xi})]$  satisfies:

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

for some  $\mu \in (0, \frac{c}{L})$ .

Consider the algorithm defined by

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

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

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

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

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

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

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

$$\mathbb{E}[\|D(\mathbf{x}, \boldsymbol{\xi})\|_2^2] \leq M^2 \quad \forall \mathbf{x} \in \mathbb{R}^n. \quad (6)$$

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

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

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

$$(\mathbf{y} - \mathbf{x})^T (\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})) \geq c \|\mathbf{y} - \mathbf{x}\|_2^2 \quad (8)$$

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

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

Inequalities (9) and (8) implies

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

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

$$(\mathbf{x} - \mathbf{x}_*)^T (\nabla f(\mathbf{x}) - g(\mathbf{x}) + g(\mathbf{x})) \geq c \|\mathbf{x} - \mathbf{x}_*\|_2^2, \quad (11)$$

which can be rewritten as

$$(\mathbf{x} - \mathbf{x}_*)^T g(\mathbf{x}) \geq c \|\mathbf{x} - \mathbf{x}_*\|_2^2 - (\mathbf{x} - \mathbf{x}_*)^T (\nabla f(\mathbf{x}) - g(\mathbf{x})) \quad (12)$$

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

$$\mathbb{E}[(\mathbf{x}_j - \mathbf{x}_*)^T g(\mathbf{x}_j)] \geq c \mathbb{E}[\|\mathbf{x}_j - \mathbf{x}_*\|_2^2] - \mathbb{E}[(\mathbf{x}_j - \mathbf{x}_*)^T (\nabla f(\mathbf{x}_j) - g(\mathbf{x}_j))] \quad (13)$$

$$\geq c(1 - \frac{\mu L}{c}) \mathbb{E}[\|\mathbf{x}_j - \mathbf{x}_*\|_2^2] \quad (14)$$

$$= 2\bar{c}a_j, \quad (15)$$

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

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

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

$$a_{j+1} \leq (1 - 2\bar{c}\gamma_j)a_j + \frac{1}{2} \frac{\beta^2 M^2}{j^2}. \quad (17)$$

It follows by induction that

$$\mathbb{E}[\|\mathbf{x}_j - \mathbf{x}_*\|_2^2] = 2a_j \leq \frac{Q(\beta)}{j}, \quad (18)$$

where

$$Q(\beta) = \max \left\{ \frac{\beta^2 M^2}{2\bar{c} - 1}, \|\mathbf{x}_1 - \mathbf{x}_*\|_2^2 \right\}. \quad (19)$$

Hence, since

$$f(\mathbf{x}) \leq f(\mathbf{x}_*) + \frac{1}{2}L \|\mathbf{x} - \mathbf{x}_*\|_2^2, \quad \forall \mathbf{x} \in \mathbb{R}^n, \quad (20)$$

we obtain

$$\mathbb{E}[f(\mathbf{x}_j) - f(\mathbf{x}_*)] \leq \frac{1}{2}L\mathbb{E}[\|\mathbf{x}_j - \mathbf{x}_*\|_2^2] \leq \frac{1}{2}LQ(\beta) \quad (21)$$

**Sufficient stochastic direction condition** Assumption 2 can be further elaborated. Let  $\theta$  be the angle between  $\nabla f(\mathbf{x})$  and  $g(\mathbf{x})$  and  $\|g(\mathbf{x})\| = \alpha \|\nabla f(\mathbf{x})\|$  for some  $\alpha > 0$ . Then,

$$\|\nabla f(\mathbf{x}_j) - g(\mathbf{x}_j)\|^2 = \|\nabla f(\mathbf{x}_j)\|^2 + \|g(\mathbf{x}_j)\|^2 - 2\|\nabla f(\mathbf{x}_j)\| \|g(\mathbf{x}_j)\| \cos \theta_j \quad (22)$$

$$= \|\nabla f(\mathbf{x}_j)\|^2 (1 + \alpha_j^2 - 2\alpha_j \cos \theta_j). \quad (23)$$

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

$$\|\nabla f(\mathbf{x}_j) - g(\mathbf{x})\| \leq L \|\mathbf{x} - \mathbf{x}_*\| (1 + \alpha^2 - 2\alpha \cos \theta)^{\frac{1}{2}} \quad (24)$$

Hence

$$(\mathbf{x} - \mathbf{x}_*)^T (\nabla f(\mathbf{x}) - g(\mathbf{x})) \leq \|\mathbf{x} - \mathbf{x}_*\| \|\nabla f(\mathbf{x}) - g(\mathbf{x})\| \quad (25)$$

$$\leq L \|\mathbf{x} - \mathbf{x}_*\|^2 (1 + \alpha^2 - 2\alpha \cos \theta)^{\frac{1}{2}}. \quad (26)$$

Hence a sufficient condition for assumption 2 to hold is

$$1 + \alpha^2 - 2\alpha \cos \theta_j \leq \mu^2 \quad (27)$$