On sliding gradient

Giulio Galvan

1 ottobre 2015

From Lipschitz continuity we get:

$$f(y) \leq f(x) + \nabla f(x)^T(y-x) + \frac{L}{2}\left\|x-y\right\|^2, \tag{1}$$

which choosing $x = x_k$ and $y = x_{k+1} = x_k + t_k d_k$ becomes:

$$f(x_{k+1}) - f(x_k) \leq t_k \nabla f(x_k)^t d_k + t_k^2 \frac{L}{2}\left\|d_k\right\|^2. \tag{2}$$

The previous inequality can be rewritten as:

$$f(x_k) - f(x_{k+1}) \geq -t_k \nabla f(x_k)^t d_k \cdot \left(1 + t_k \frac{L}{2} \frac{\nabla f(x_k)^t d_k}{\left\|\nabla f(x_k)\right\|^2 cos^2\theta_k}\right) \tag{3}$$

where

$$cos\theta_k = \frac{\nabla f(x_k)^t d_k}{\left\|\nabla f(x_k)\right\| \left\|d_k\right\|} \tag{4}$$

In a backtracking setting, as defined in algorithm 1, we search for a value of t_k such that:

$$f(x_k) - f(x_{k+1}) \geq \alpha t_k \nabla f(x_k)^t d_k. \tag{5}$$

When backtracking we have two possibilities: either $t_k = s$ satisfy inequality (5) or not. In the latter case it must hold:

$$f(x_k) - f(x_k + \frac{t_k}{\beta} d_k) < \alpha \frac{t_k}{\beta} \nabla f(x_k)^t d_k \tag{6}$$

Combining the latter with inequality 5 written for $t_k = \frac{t_k}{\beta}$ yields:

$$\alpha \frac{t_k}{\beta} \nabla f(x_k)^t d_k > -\frac{t_k}{\beta} \nabla f(x_k)^t d_k \cdot \left(1 + \frac{t_k}{\beta} \frac{L}{2} \frac{\nabla f(x_k)^t d_k}{\left\|\nabla f(x_k)\right\|^2 cos^2\theta_k}\right), \tag{7}$$

which in turn, being $\nabla f(x_k)^t d_k < 0$ since d_k is a descent direction, and $t_k, \beta > 0$ leads to:

$$t_k > \frac{2(\alpha+1)\beta}{L} \frac{\left\|\nabla f(x_k)\right\|^2 cos^2\theta_k}{\nabla f(x_k)^T d_k} \tag{8}$$

$$= \frac{2(\alpha+1)\beta}{L} \frac{\nabla f(x_k)^T d_k}{\|d_k^2\|} \tag{9}$$

If we impose

$$s \geq \frac{2(\alpha+1)\beta}{L} \frac{\nabla f(x_k)^T d_k}{\|d_k^2\|} \tag{10}$$

we can use 8 in 5 and get:

$$f(x_k) - f(x_{k+1}) > \frac{2\alpha(\alpha+1)\beta}{L} \|\nabla f(x_k)\|^2 cos^2\theta_k. \tag{11}$$

Summing over k, if f is bounded below, say by f^* and θ_k bounded away from 90 degrees we get:

$$f(x_0) - f^* \geq \sum_{k=0}^{N} f(x_k) - f(x_{k+1}) = f(x_0) - f(x_N) > C \sum_{k=0}^{N} \|\nabla f(x_k)\|^2, \tag{12}$$

hence we have convergence at the same rate as gradient descent.

We have made the following assumptions along the way:

- $f(\cdot)$ is bounded below
- d_k is a descent direction bounded away from 90 degrees w.r.t $\nabla f(x_k)$
- the initial guess of the backtracking algorithm $s \geq \frac{2(\alpha+1)\beta}{L} \frac{\nabla f(x_k)^T d_k}{\|d_k^2\|}$

Algorithm 1: Backtracking algorithm.

Data:
$s > 0$ initial step guess
$\alpha, \beta \in (0, 1)$

```
t_k ← s
while f(x_k) − f(x_{k+1}) < α t_k ∇f(x_k)^t d_k do
    t_k ← t_k / β
end
return t_k
```