Gradient methods

Jaume Sánchez

Ex 3.1

$$f(x,y) = x^2 + xy + y^2 + 5\,; \qquad x_0 = (1,1)$$

a) Conjugate gradient descent:

$$\nabla f(x,y) = (2x+y,\ 2y+x)^T \Rightarrow \nabla f(x_0,y_0) = (3,3)^T$$

$$\nabla^2 f(x,y) = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} := A.$$

$$d_0 = -\nabla f(x_0,y_0) = (-3,-3)^T$$

$$\alpha_0 = \frac{(3,3)\cdot(3,3)^T}{(3,3)\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}(3,3)^T} = \frac{18}{54} = \frac{1}{3}$$

$$x_1 = (1,1)^T + \frac{1}{3}(-3,3)^T = (0,0)^T \Rightarrow \nabla f(0,0) = (0,0)$$

As $\nabla f(x_1) = 0$, $d_1 = 0 \Rightarrow$

$$\Rightarrow \quad x_2 = (0,0)^T$$

So after 2 steps (actually one) in the conjugate gradient descent we reach the point $(x,y) = (0,0)$

Solution.

b) Hypergradient descent :

Same function and starting point, that is:

$f(x,y) = x^2 + xy + y^2 + 5$; $x_0 = (1,1)$; $\nabla f(x,y) = (2x+y, 2y+x)$

The main difference in this method is that now the learning rate (α) will also be considered as a hyperparameter.

Let us start with $\alpha_0 = 0'1$ and set $\mu = 0'1$.

- $$x_1 = x_0 - \frac{\alpha_0 \nabla f(x_0)}{\|\nabla f(x_0)\|} = (1,1)^T - \frac{0'1\,(3,3)^T}{3\sqrt{2}} =$$
  $$\simeq (0'93, 0'93)$$
- $$\alpha_1 = \alpha_0 + \frac{\mu \cdot (\nabla f(x_1)) \cdot \nabla f(x_0)^T}{\|\nabla f(x_0)\|} =$$
  $$= 0'1 + \frac{0'1\,(2'79, 2'79) \cdot (3,3)^T}{3\sqrt{2}} \simeq 0'49$$
- $$x_2 = x_1 - \alpha_1 \cdot \frac{\nabla f(x_1)}{\|\nabla f(x_1)\|} =$$
  $$= (0'93, 0'93) - \frac{0'49\,(2'79, 2'79)}{3'95} = (0'58, 0'58)$$

So, after 2 steps in the hypergradient descent method we reach the point $(x,y) = (0'58, 0'58)$

Solution

Ex 3.2

$f(x,y) = (x+1)^2 + (y+3)^2 + 4$ starting at $(0,0)$

The Newton method is given by :

$$\boxed{x_{k+1} = x_k - \left[Hf(x_k)\right]^{-1} \cdot \nabla f(x_k)}$$

where Hf denotes the Hessian matrix of f.

- $\nabla f(x,y) = (2(x+1), 2(y+3))^T \Rightarrow \nabla f(0,0) = (2,6)^T$
- $Hf(x,y) = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$
- $\left[Hf(x,y)\right]^{-1} = \frac{1}{2} \cdot \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$

So, 1 step in the classical Newton method is:

$$x_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix} - \frac{1}{2} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 2 \\ 6 \end{pmatrix}$$

$$x_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix} - \begin{pmatrix} 1 \\ 3 \end{pmatrix} = \begin{pmatrix} -1 \\ -3 \end{pmatrix}$$

So, 1 step in the classical Newton method gives us $x_1 = (-1,-3)^T$ Solution.

EX 3.3

(remark: I think you meant $\frac{1}{|I_k|} \sum_{i \in |I_k|} \nabla f_i(x)$)

a) We choose $i(k)$ uniformly at random at every step. Suppose, $|I_k| = n$, then we have that

$$\mathbb{E}\left(\frac{1}{2} \sum_{i \in |I_k|} \nabla f_i(x)\right) = \frac{1}{2} \mathbb{E}\left(\nabla f_i(x) + \nabla f_j(x)\right) = \frac{1}{2} \cdot \frac{1}{n^2}\left[2n \nabla f_i(x) + 2n \nabla f_j(x)\right] =$$

$$= \frac{2n}{2 \cdot n^2}\left(\nabla f_i(x) + \nabla f_j(x)\right) = \frac{1}{n} \sum_{i \in |I_k|} \nabla f_i(x) = \nabla f(x)$$

Hence, it is a stochastic gradient $\square$

b) When $|I_k| = 2$, we have

$$\mathrm{Var}\left(\frac{1}{2}\left(\nabla f_i + \nabla f_j\right)\right) = \frac{1}{4} \mathrm{Var}\left(\nabla f_i + \nabla f_j\right) \overset{iid}{=}$$

$$= \frac{1}{4}\left[\mathrm{Var}(\nabla f_i) + \mathrm{Var}(\nabla f_j)\right].$$

Suppose $\mathrm{Var}(\nabla f_i) = \sigma^2 \; \forall i$

Then, $\mathrm{Var}\left(\frac{1}{2}\left(\nabla f_i + \nabla f_j\right)\right) = \frac{2\sigma^2}{4} = \frac{\sigma^2}{2}$.

Evidently, this is smaller than $\sigma^2 = \mathrm{Var}(\nabla f_i)$ $\square$

c) Suppose $|I_k| = n > 0$, then

$$\mathrm{Var}\left(\frac{1}{n} \sum_{i \in |I_k|} \nabla f_i\right) = \frac{1}{n^2} \sum_{i \in |I_k|} \mathrm{Var}(\nabla f_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} <$$

$$< \mathrm{Var}(\nabla f_i) = \sigma^2.$$

So this method actually works for other sizes of the minibatches. What's more, if $n_1 < n_2$,

$$\underset{|I_k| = n_1}{\mathrm{Var}\left(\frac{1}{n_1} \sum_{i \in |I_k|} \nabla f_i\right)} > \underset{|I_k'| = n_2}{\mathrm{Var}\left(\frac{1}{n_2} \sum_{j \in |I_k'|} \nabla f_j\right)}$$ $\square$

Solution.