3rd Assingment - Optimization - Gradient Methods

Exercise 3.1.: $f(x,y) = x^2 + xy + y^2 + 5$

- Starting at (1,1) write 2 steps of conjugate gradient descent method for f.

1st step:

$$X_0 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad \nabla f(x) = \begin{pmatrix} 2x + y \\ x + 2y \end{pmatrix} \quad \nabla f(X_0) = \begin{pmatrix} 3 \\ 3 \end{pmatrix}$$

$$Z_1 = -\nabla f(X_0) = \begin{pmatrix} -3 \\ -3 \end{pmatrix}$$

Minimizing $f(X_0 + a_1 Z_1)$ with respect to a_1 we get

$$a_1 = -\frac{(Z_1)^T \nabla f(X_0)}{(Z_1)^T A (Z_1)} \quad \text{where} \quad A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \text{ the Hessian.}$$

$$a_1 = -\frac{(-3 \ -3)\begin{pmatrix} 3 \\ 3 \end{pmatrix}}{(-3 \ -3)\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}\begin{pmatrix} -3 \\ -3 \end{pmatrix}} = -\frac{-18}{(-9 \ -9)\begin{pmatrix} -3 \\ -3 \end{pmatrix}} = -\frac{-18}{54} = \frac{1}{3}$$

$$X_1 = X_0 + a_1 Z_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} + \frac{1}{3}\begin{pmatrix} -3 \\ -3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} //$$

- For the same starting point and function, do 2 steps for hypergradient descent ~~p~~ method.

<u>1^{st} step:</u>

$$\begin{pmatrix} x_0 \\ y_0 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad \nabla f(x_0, y_0) = \begin{pmatrix} 3 \\ 3 \end{pmatrix} \quad \nabla f(x,y) = \begin{pmatrix} 2x + y \\ x + 2y \end{pmatrix}$$

$$d_1 = -\nabla f(x_0, y_0) = \begin{pmatrix} -3 \\ -3 \end{pmatrix} \qquad A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$$

$$(x_1, y_1) = (x_0, y_0) - \alpha_0 A^{-1} \nabla f(x_0, y_0)$$

$$A^{-1} = \frac{1}{4-1} \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} = \begin{pmatrix} 2/3 & -1/3 \\ -1/3 & 2/3 \end{pmatrix}$$

For $\alpha_0 = 0.01 = \frac{1}{100}$

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} - \frac{1}{100} \begin{pmatrix} 2/3 & -1/3 \\ -1/3 & 2/3 \end{pmatrix} \begin{pmatrix} 3 \\ 3 \end{pmatrix}$$

$$= \begin{pmatrix} 1 \\ 1 \end{pmatrix} - \frac{1}{100} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 99/100 \\ 99/100 \end{pmatrix}$$

<u>2^{nd} step:</u>

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} 99/100 \\ 99/100 \end{pmatrix} \quad \nabla f(x_1, y_1) = \begin{pmatrix} 297/100 \\ 297/100 \end{pmatrix}$$

$$(x_2, y_2) = (x_1, y_1) - \alpha_1 A^{-1} \nabla f(x_1, y_1)$$

$$= \begin{pmatrix} 99/100 \\ 99/100 \end{pmatrix} - \alpha_1 \begin{pmatrix} 2/3 & -1/3 \\ -1/3 & 2/3 \end{pmatrix} \begin{pmatrix} 297/100 \\ 297/100 \end{pmatrix} = \begin{pmatrix} 99/100 \\ 99/100 \end{pmatrix} - \alpha_1 \begin{pmatrix} 99/100 \\ 99/100 \end{pmatrix}$$

To calculate α_1: $\frac{\partial}{\partial \alpha_1} \underbrace{f\left[(x_1, y_1) - \alpha_1 A^{-1} \nabla f(x_1, y_1)\right]}_{\underset{g(\alpha_1)}{\parallel}}$

$$\Rightarrow \frac{\partial}{\partial \alpha_1} g(\alpha_1) = 0$$

So, $g(\alpha_1) = 3\left[\begin{pmatrix} 99/100 \\ 99/100 \end{pmatrix} - \alpha_1 \begin{pmatrix} 99/100 \\ 99/100 \end{pmatrix}\right]^2 + 5$

$$\frac{\partial}{\partial \alpha_1} g(\alpha_1) = -6\left[\begin{pmatrix} 99/100 \\ 99/100 \end{pmatrix} - \alpha_1 \begin{pmatrix} 99/100 \\ 99/100 \end{pmatrix}\right] \begin{pmatrix} 99/100 \\ 99/100 \end{pmatrix} = 0 \Rightarrow \boxed{\alpha_1 = 1}$$

Therefore,

$$(x_2, y_2) = \begin{pmatrix} 99/100 \\ 99/100 \end{pmatrix} - 1 \begin{pmatrix} 99/100 \\ 99/100 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad //$$

Exercise 3.2: Write down 1 step in the classical Newton method for the function $f(x,y)=(x+1)^2+(y+3)^2+4$ starting at $(0,0)$

$$\nabla f(x,y)=\begin{pmatrix}2(x+1)\\2(y+3)\end{pmatrix} \qquad H(x,y)=\begin{pmatrix}2 & 0\\0 & 2\end{pmatrix}$$ the hessian matrix

$$x_0=\begin{pmatrix}0\\0\end{pmatrix} \qquad \nabla f(0,0)=\begin{pmatrix}2\\6\end{pmatrix} \qquad g_1=-\nabla f(0,0)=\begin{pmatrix}-2\\-6\end{pmatrix}$$

$$X_1=\overbrace{x_0+[H(x_0)]^{-1}g_1}^{\text{Newton method}}=\begin{pmatrix}0\\0\end{pmatrix}+\left[\begin{pmatrix}2 & 0\\0 & 2\end{pmatrix}\right]^{-1}\begin{pmatrix}-2\\-6\end{pmatrix}$$

$$\hookrightarrow \left[\begin{pmatrix}2 & 0\\0 & 2\end{pmatrix}\right]^{-1}=\frac{1}{4-0}\begin{pmatrix}2 & 0\\0 & 2\end{pmatrix}=\begin{pmatrix}1/2 & 0\\0 & 1/2\end{pmatrix}$$

In general: $A=\begin{pmatrix}a & b\\c & d\end{pmatrix} \qquad A^{-1}=\frac{1}{ad-cb}\begin{pmatrix}d & -b\\-c & a\end{pmatrix}$

$$X_1=\begin{pmatrix}1/2 & 0\\0 & 1/2\end{pmatrix}\begin{pmatrix}-2\\-6\end{pmatrix}=\begin{pmatrix}-1\\-3\end{pmatrix}$$

Exercise 3.3: $x^{(k+1)} = x^{(k)} - \frac{a^{(k)}}{|I_k|} \sum_{i \in I_k} \nabla f_i(x^{(k)})$

size of minibatch equal to 2. Show that:

▶ $\sum_{i \in I_k} \nabla f_i(x)$ is a stochastic gradient

This expression is the sum of the individual functions $f_i(x)$ for i in the mini-batch I_k. Therefore, $\sum_{i \in I_k} \nabla f_i(x)$ is an estimate of the true gradient and it is stochastic because is calculated based on a randomly selected subset of the data.

$|I_k| = 2$, I_k a set with two elements randomly selected from $\{1, 2, .., n\}$

$$g_k = \frac{1}{|I_k|} \sum_{i \in I_k} \nabla f_i(x) = \frac{1}{2}(\nabla f_{i_1}(x) + \nabla f_{i_2}(x))$$

$$\mathbb{E}(g_k) = \frac{1}{2} \mathbb{E}(\nabla f_{i_1}(x) + \nabla f_{i_2}(x)) = \underbrace{\frac{1}{2}\mathbb{E}(\nabla f_{i_1}(x)) + \frac{1}{2}\mathbb{E}(\nabla f_{i_2}(x))}_{\hookrightarrow \text{because are independent}} = \frac{1}{2} \cdot 2\, \mathbb{E}(\nabla f_{i_1}(x)) = \nabla f_i(x)$$

▶ The variance of this gradient is smaller than the variance of the standard stochastic gradient $\nabla_{i(k)}(x)$ (size of minibatch, 1)

Let's write down the variances for each gradient

- $Var(g_k) = Var\left(\frac{1}{2}(\nabla f_{i_1}(x) + \nabla f_{i_2}(x))\right)$

  $= \frac{1}{4} Var(\nabla f_{i_1}(x)) + \frac{1}{4} Var(\nabla f_{i_2}(x)) + \frac{1}{2} Cov$

  ↳ because $\nabla f_{i_1}(x)$ and $\nabla f_{i_2}(x)$ are independent

- $Var(\nabla_{i_k}(x))$

The variance of g_k is the average of the variances of two individual gradients in the mini-batch and it is smaller than the variance of $\nabla_{i_k}(x)$ when the mini-batch size is 1.

▶ Does your argument work for other sizes of the mini-batches?

Yes, because the main idea is that while mini-batch size increases, the variance of the stochastic gradient tends to decrease compaired to the variance of the individual gradients in the mini-batch. This is because averaging over some samples tends to smooth out the noise accosiated with individual noise.

For example a mini-batch size m: • $Var(g_k) = \frac{1}{m^2} \sum_{i \in I_k} Var(\nabla f_i(x))$

• $Var(\nabla_{i_k}(x))$