Pre-class Assignment-25

Reading:

(1) The condition $\lim_{K\to\infty} \sum_{t=0}^{K} \gamma_t = \infty$ is important in the context of algorithm that uses step-sizes in iterative process. Here's why it's necessary:

* This condition ensures that the algorithm makes sufficient progress over time. If the sum of the step sizes does not diverge, the cummulative changes made by the algorithm could be too small, preventing it from reaching the optimum.
* The condition $\lim_{K\to\infty} \gamma_K = 0$ ensures that step-size decrease to zero, allowing for finer adjustment near the optimum. Its complement $\sum_{t=0}^{K} \gamma_t = \infty$, guarantees that the cummulative effect is enough to reach any point in the space, despite decreasing step sizes.
* In the absence of this condition, step-sizes might remain too large or not decrease adequately, potentially causing the algorithm to oscillate or diverge from the optimal solution.

② For the function $f(\gamma) = f(x_{t-1} - \gamma \nabla f(x_{t-1}))$, the convexity of 'f' over its domain does not directly imply the convexity $f(\gamma)$ in terms of γ.

* The given function is a composition of the convex function 'f' with a linear transformation. The convexity of this composite function is not directly inferred from the convexity of 'f' alone.
* The term '$-\gamma \nabla f(x_{t-1})$' indicates a step against the gradient of 'f'. How this transformation influences the convexity of the overall function $f(\gamma)$ is complex and context-dependent.
* For $f(\gamma)$ to be convex, it must satisfy specific convexity conditions in its one-dimensional line search space, which is not automatically assured by the convexity of 'f' in its original domain.

③ Gradient descent with momentum tends to perform better than Standard gradient descent under certain Conditions related to the properties of the objective function.

* Gradient descent with momentum excels with objective functions having ill-conditioned curvature, aiding in Smoothing out frequent direction changes & accelerating Convergence in narrow valleys.
* It performs better in situations with noisy gradients, as it averages out noise over iterations, leading to a more stable descent.
* Momentum is benefical in traversing flat areas more efficiently and stabilizing the descent in deep, narrow minima, where Standard gradient descent might slow down or oscillate.
* Especially in non-convex landscapes common in deep learning, momentum can help achieve faster Convergence, reducing overall training time.

Exercises

① given, $f(x) = x^3 - 3x^2 - x + 3$ for all $x \geq 0$

i) Choosing $\gamma_t = \frac{1}{25t}$ for 3 steps $t = 1, 2, 3$

$$f'(x) = 3x^2 - 6x - 1$$

$$f'(x) = 0 \Rightarrow 3x^2 - 6x - 1 = 0 \Rightarrow x_1 = 2.15 \text{ \& } x_2 = -0.15$$

Considering $x_1 = 2.15$

for gradient descent, $x_{t+1} = x_t - \gamma_t \nabla f(x_t)$

Round 1: for $\gamma_t = \frac{1}{25t}$ where $t = 1, 2, 3$

$$x_0 = 0 \text{ \& } x_1 = x_0 - \frac{1}{25} f'(x_0) = 0 - \frac{1}{25}(-1) = \frac{1}{25}$$

Now, $\gamma_2 = \frac{1}{25}$

$$x_2 = x_1 - \frac{1}{50} \cdot f'(x_1)$$

$$x_2 = \frac{1}{25} - \frac{1}{50}\left(3\left(\frac{1}{25}\right)^2 - 6\left(\frac{1}{25}\right) - 1\right) = 0.0647$$

Now, $\gamma_3 = \frac{1}{75}$

$$x_3 = x_2 - \frac{1}{75} f'(x_2)$$

$$x_3 = 0.0647 - \frac{1}{75}\left(3(0.0647)^2 - 6(0.0647) - 1\right) = 0.0834$$

ii) Choosing $\gamma_t = \frac{1}{t}$ where $t = 1, 2, 3$

$\gamma_1 = 1$

$x_0 = 0$; $x_1 = x_0 - 1(f'(x)) = 0 - 1(-1) = 1$

Now, $\gamma_2 = \frac{1}{2}$

$$x_2 = x_1 - \frac{1}{2}(f'(x_1)) = 1 - \frac{1}{2}(3(1)^2 - 6(1) - 1) = 3$$

Now, $\gamma_3 = 1/3$

$$x_3 = x_2 - \frac{1}{3}(f'(x_2)) = 3 - \frac{1}{3}(3(3)^2 - 6(3) - 1) = -1.667$$

iii) Choosing $\gamma_t = \frac{1}{50t}$ where $t = 1, 2, 3$

$\gamma_1 = 1/50$

$$x_0 = 0 \; ; \; x_1 = x_0 - \frac{1}{50} f'(x_0) = 0 - \frac{1}{50}(-1) = \frac{1}{50}$$

Now, $\gamma_2 = 1/100$

$$x_2 = x_1 - \frac{1}{100} f'(x_1) = \frac{1}{50} - \frac{1}{100}\left(3\left(\frac{1}{50}\right)^2 - 6\left(\frac{1}{50}\right) - 1\right) = 0.031$$

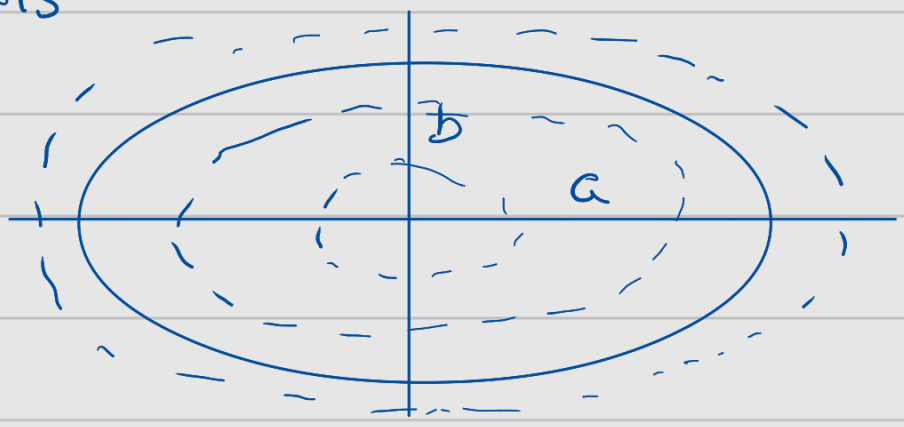
Now, $\gamma_3 = 1/150$

$$x_3 = x_2 - \frac{1}{150} f'(x_2) = 0.031 - \frac{1}{50}\left(3(0.031)^2 - 6(0.031) - 1\right)$$

$$= 0.0388$$

On Comparing the results with pre-class Assignment-24, the non-constant step sizes ensures the algorithm to converge faster by taking larger steps when gradient is steep & smaller steps when gradient is shallow, thus preventing any overshooting & oscillations.

(2) when 'a' is substantially larger than b it means ellipse is streched towards x-axis & less towards y-axis



$$f(x,y) = \frac{x^2}{a^2} + \frac{y^2}{b^2}$$

(x_i, y_i) are the points resulted from gradient descent with some constant step without momentum. The update is given by

$$(x_{i+1}, y_{i+1}) = (x_i, y_i) - \alpha \nabla f(x_i, y_i)$$

$$\nabla f(x,y) = \left(\frac{2x}{a^2}, \frac{2y}{b^2}\right)$$

b) We have $\nabla f(x,y) = \left(2x/a^2, 2y/b^2\right)$

level curve is given by

$$f(x,y) = f(x_i,y_i) = \frac{x^2}{a^2} + \frac{y^2}{b^2} = \frac{x^2}{a^2} + \frac{y^2}{b^2}$$

$$\text{optimal Step Size} = \sqrt{\left(\frac{\partial f}{\partial x}\right)^2 + \left(\frac{\partial f}{\partial y}\right)^2}$$

$$= \sqrt{\left(\frac{2x_i}{a^2}\right)^2 + \left(\frac{2y_i}{b^2}\right)^2} = \text{optimal Step}$$

when we update this gradient in exact search along -ve direction, we converge to minima ie

$$x_{i+1} = x_i - \text{optimal step} \cdot \frac{2x_i}{a^2}$$

$$y_{i+1} = y_i - \text{optimal step} \cdot \frac{2y_i}{b^2}$$

(3) $f(x) = x^4 - x^3 - x^2 + 1$ (4 rounds) $x_0 = -1$ &

$\gamma = 1/10$

i) $\alpha = 0.3$

$f'(x) = 4x^3 - 3x^2 - 2x$

$V_{t+1} = \alpha V_t + \gamma \nabla f(x_t)$ & $x_{t+1} = x_t - V_{t+1}$

$V_1 = \alpha V_0 + \gamma f'(x_2) = 0.3(0) + \frac{1}{10}(4(-1)^3 - 3(-1)^2 - 2(-1))$

$V_1 = = -0.5$

$x_1 = x_0 - v_1 = 0 - (-0.5) = 0.5$

Now, $v_2 = \alpha v_1 + \gamma f'(x_1) = -0.275$

$x_2 = x_1 - v_2 = 0.5 - (-0.275) = 0.775$

$v_3 = \alpha v_2 + \gamma f'(x_2) = -0.231$

$x_3 = x_2 - v_3 = 0.775 - (-0.231) = 1.006$

$v_4 = \alpha v_3 + \gamma f'(x_3) = -0.1693$

$x_4 = x_3 - v_4 = 1.006 - (-0.1693) = 1.1753$

ii) $\alpha = 0.8$

$v_1 = \alpha v_0 + \gamma f'(x_0) = 0.8(0) + \frac{1}{10}(4(-1)^3 - 3(-1)^2 - 2(-1))$

$v_1 = -0.5$

$x_1 = x_0 - v_1 = 0 - (-0.5) = 0.5$

$v_2 = \alpha v_1 + \gamma f'(x_1) = -0.525$

$x_2 = x_1 - v_2 = 0.5 - (-0.525) = 1.025$

$v_3 = \alpha v_2 + \gamma f'(x_2) = -0.5094$

$x_3 = x_2 - v_3 = 1.025 - (-0.5094) = 1.5344$

$v_4 = \alpha v_3 + \gamma f'(x_3) = -0.8386$

$x_4 = x_3 - v_4 = 1.5344 - (-0.8386)$

$x_4 = 2.373$

iii) $\alpha = 0$

$$v_1 = \alpha v_0 + \gamma f'(x_0) = 0 + 0.1\left(4(-1)^3 - 3(-1)^2 - 2(-1)\right)$$

$$v_1 = -0.5$$

$$x_1 = x_0 - v_1 = -1 - (-0.5) = -0.5$$

$$v_2 = \alpha v_1 + \gamma f'(x_1) = -0.025$$

$$x_2 = x_1 - v_2 = -0.5 - (-0.025) = -0.475$$

$$v_3 = \alpha v_2 + \gamma f'(x_2) = -0.0156$$

$$x_3 = x_2 - v_3 = -0.475 - (-0.0156) = -0.4594$$

$$v_4 = \alpha v_3 + \gamma f'(x_3) = -0.01$$

$$x_4 = -0.4594 - (-0.01) = 0.4494$$

iv)