Week 4 Tutorial Tasks

Sarat Moka

Theoretical Task: Adam's Effective Learning Rate

From lecture notes, recall that the Adam algorithm update rule, without biases corrections, is as follows: Starting with an initial point $\theta^{(0)}$ and with $v^{(0)} = s^{(0)} = 0$,

$$\theta^{(t+1)} = \theta^{(t)} - \alpha \frac{1}{\sqrt{s^{(t+1)}} + \epsilon} v^{(t+1)}, \quad \text{for } t = 1, 2, \ldots,$$

where

$$\begin{aligned} v^{(t+1)} &= \beta v^{(t)} + (1-\beta)\nabla C(\theta^{(t)}) \\ s^{(t+1)} &= \gamma s^{(t)} + (1-\gamma)\left(\nabla C(\theta^{(t)}) \odot \nabla C(\theta^{(t)})\right), \end{aligned}$$

where $\odot$ denotes the Hadamard or elementwise product between the vectors of the same size.

In this setting, consider a stationary gradients case: $\nabla C(\theta^{(t)}) \approx g$ for all t for a fixed vector g. Then,

1. Prove that for sufficiently large t the *effective* learning rate vector is

   $$\alpha_{\texttt{eff}} \approx \alpha \frac{1}{|g|},$$

   where the inverse and $|\cdot|$ are applied elementwise.

2. Based on the above expression, how gradient elements effect the learning rate of each element of θ?

Coding Task: Adam vs GD Learning Rate Sensitivity

Compare optimization paths of Adam and the basic gradient descent with different learning rates on following Rosenbrock function:

$$f(x, y) = (1-x)^2 + 10(y - x^2)^2,$$

which has a unique global minimum at (1,1). Take the initial point $(x^{(0)}, y^{(0)}) = (-1.5, 2.5)$ and vary the learning rate parameter α over $[0.5, 0.1, 0.05, 0.01, 0.005]$.

In Adam implementation, use the default parameter values provided in Remark 4.3.2 in the lecture notes.