Week 2 Optimisation for Machine Learning

Neimhin Robinson Gunning, 16321701

March 13, 2024

Let

$$f(x,y) = 3(x-5)^4 + 10(y-9)^2 \tag{1}$$

and

$$g(x,y) = \max(x-5,0) + 10|y-9| \tag{2}$$

Using `sympy` we find the derivatives:

$$\nabla f = [\frac{df}{dx}, \frac{df}{dy}] = [12(x-5)^3, 20y - 180]$$

$$\nabla g = [\frac{dg}{dx}, \frac{dg}{dy}] = [\text{Heaviside}(x-5), 10\text{sign}(y-9)]$$

Clearly, the minimum of $f(x,y)$ is 0 and they is minimized by $x=5$, $y=9$. The other function $g(x,y)$ also has minimum 0 but is minized by any of $x \in [-\infty, 5]$ and $y = 9$.

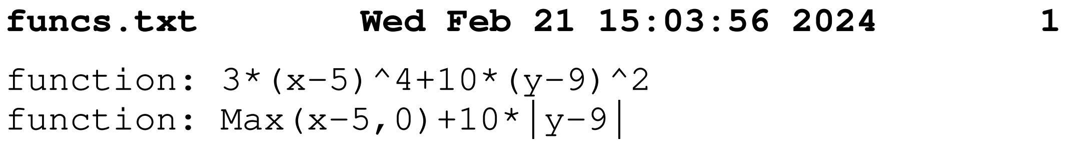
1 (a)

1.1 (a) (i) Polyak

The Polyak step size is

$$\alpha_{\text{Polyak}} = \frac{f(x) - f^*}{\nabla f(x)^T \nabla f(x)} \tag{3}$$

where x is the parameter vector, $f(x)$ is the function to optimise, and $f^* \approx \min_x f(x)$.



```
funcs.txt          Wed Feb 21 15:03:56 2024          1

function: 3*(x-5)^4+10*(y-9)^2
function: Max(x-5,0)+10*|y-9|
```

Figure 1: Two bivariate functions downloaded from `https://www.scss.tcd.ie/Doug.Leith/CS7DS2/week4.php`

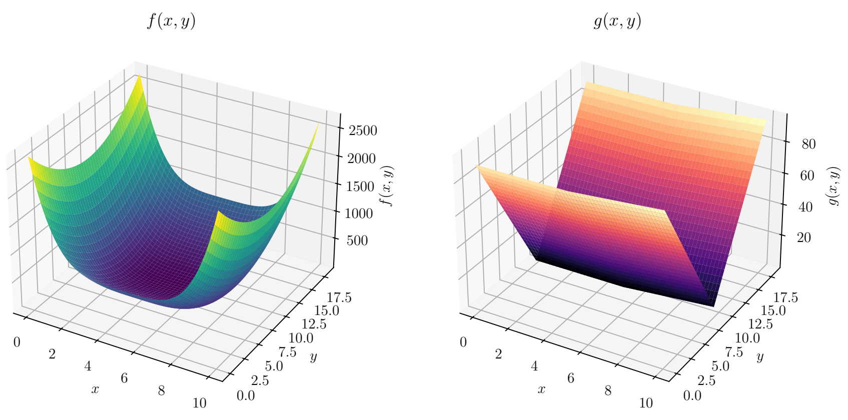


Figure 2

Gradient descent iteration with Polyak step size is implemented in Listing 1. The function is evaluated at the current value for x and the numerator is calculated: $f(x) - f^*$. A reasonable estimate for the minimum of the function, f^*, is required, here assumed to be 0. The dot product of the gradient is taken as the denominator. The step size is $\frac{f(x)-f^*}{\nabla f(x)^T \nabla f(x)}$. We multiply the step size by the gradient and subtract the result from the current x.

Listing 1: An implementation of the update step of gradient descent using Polyak step size.

src/polyak.py   Tue Mar 12 15:12:27 2024   1

```python
import numpy as np

def iterate(self):
    self._x_value = self._start
    self._old_x_value = None
    self._f_star = 0
    self._iteration = 0
    self._converged_value = False
    self._grad_value = self._gradient(self._x_value)

    yield self.state_dict()

    while not self._converged_value:
        if self._max_iter > 0 and self._iteration > self._max_iter:
            break
        numerator = self._function(self._x_value) - self._f_star
        self._grad_value = self._gradient(self._x_value)
        denominator = np.dot(self._grad_value, self._grad_value) # sum of element-wise products
        self._old_x_value = self._x_value
        step = numerator/denominator
        self._x_value = self._x_value - step * self._grad_value
        self._converged_value = self._converged(self._x_value, self._old_x_value)
        yield self.state_dict()
```

1.2 (a) (ii) RMSProp

The RMSProp step size at iteration t is

$$\alpha_t = \frac{\alpha_0}{\epsilon + \sqrt{(1-\beta)\sum_{i=0}^{t-1} \beta^{t-i} (\nabla f(x_i))^2}} \tag{4}$$

and the update rule is

$$x_{t+1} := x_t - \alpha_t * \nabla f(x_t) \tag{5}$$

where ϵ is some small value to prevent divide by zero, α_0 and β are hyperparameters to be set, noting that $0 < \beta \leq 1$. The result is that previous gradients influence the current step size, but are gradually forgotten due to the β^{t-i} term.

A Python implementation of the update step is provided in Listing 2. The term inside the square root can be calculated iteratively, as in line 25 of Listing 2.

Listing 2: An implementation of the update step of gradient descent using RMSProp step size.

```
src/rmsprop.py        Tue Mar 12 18:04:47 2024        1
     1: def iterate(self):
     2:     import numpy as np
     3:     self._x_value = self._start
     4:     old_x_value = None
     5:     self._iteration = 0
     6:     self._sum = np.zeros(self._x_value.shape)
     7:     alpha_n = np.zeros(self._x_value.shape)
     8:     alpha_n.fill(self._step_size)
     9:     self._converged_value = False
    10:     self._grad_value = self._gradient(self._x_value)
    11:
    12:     yield self.state_dict()
    13:
    14:     while not self._converged_value:
    15:         self._iteration += 1
    16:         if self._max_iter > 0 and self._iteration > self._max_iter:
    17:             break
    18:         self._grad_value = self._gradient(self._x_value)
    19:         old_x_value = self._x_value
    20:         self._x_value = self._x_value - alpha_n * self._grad_value
    21:         self._sum = self._beta * self._sum + (1-self._beta) * (self._grad_value**2)
    22:         alpha_n = self._step_size / (self._sum**0.5+self._epsilon)
    23:         self._converged_value = self._converged(self._x_value, old_x_value)
    24:         yield self.state_dict()
```

1.3 (a) (iii) Heavy Ball

The Heavy Ball step is

$$z_{t+1} = \beta z_t + \alpha \nabla f(x_t) \tag{6}$$

with the update rule

$$x_{t+1} = x_t - z_{t+1} \tag{7}$$

where t is the current iteration (starting at 0), $z_0 = 0$, and x_0, α, and β have to be set.

A Python implementation of the update step is provided in Listing 3.

Listing 3: An implementation of the update step of gradient descent using Heavy Ball step size.

```
src/heavy_ball.py        Tue Mar 12 14:57:31 2024        1
import lib


def iterate(self):
    self._x_value = self._start
    self._old_x_value = None
    self._iteration = 0
    self._converged_val = False
    self._grad_value = self.gradient(self._x_value)
    self._z = 0
    yield self.state_dict()  # yield initial values

    while not self._converged_val:
        self._iteration += 1
        if self._max_iter > 0 and self._iteration > self._max_iter:
            break
        self._grad_value = self._gradient(self._x_value)
        self._old_x_value = self._x_value
        self._z = self._beta * self._z + self._step_size * self._grad_value
        self._x_value = self._x_value - self._z
        self._converged_val = self._converged(self._x_value, self._old_x_value)
        yield self.state_dict()
```

1.4 (a) (iv) Adam

The Adam step size is calculated in terms of

$$m_{t+1} = \beta_1 m_t + (1-\beta_1)\nabla f(x_t) \tag{8}$$

and

$$v_{t+1} = \beta_2 v_t + (1-\beta_2)[\nabla f(x_t) \circ \nabla f(x_t)] \tag{9}$$

from which we get

$$\hat{m} = \frac{m_{t+1}}{(1-\beta_1^t)} \tag{10}$$

and

$$\hat{v} = \frac{v_{t+1}}{(1-\beta_2^t)} \tag{11}$$

which are used in the update step as

$$x_{t+1} = x_t - \alpha[\frac{\hat{m}_1}{\epsilon + \sqrt{\hat{v}_1}}, \dots, \frac{\hat{m}_n}{\epsilon + \sqrt{\hat{v}_n}}] \tag{12}$$

where t is the iteration, α, β_1, and β_2 are hyperparameters, and ϵ is some small value to prevent divide-by-zero. The element-wise product is denoted with $\circ$.

A Python implementation of the update step is provided in Listing 4.

2 (b)

In this section we look at the influence of the hyperparameters on the behaviour gradient descent for each of RMSProp, Heavy Ball (HB), and Adam, on the functions $f(x, y)$ and $g(x, y)$. In each

Listing 4: An implementation of the update step of gradient descent using Adam step size.

```
src/adam.py        Tue Mar 12 14:57:31 2024        1

import lib
import numpy as np
import json


def iterate(self):
    self._x_value = self._start
    self._old_x_value = None
    self._iteration = 0
    self._m = np.zeros(self._x_value.shape, dtype=np.float64)
    self._v = np.zeros(self._x_value.shape, dtype=np.float64)
    self._converged_value = False
    self._grad_value = self._gradient(self._x_value)

    yield self.state_dict()

    while not self._converged_value:
        if self._max_iter > 0 and self._iteration > self._max_iter:
            break
        self._grad_value = self._gradient(self._x_value)
        self._m = self._beta * self._m + (1-self._beta)*self._grad_value
        # grad_value * grad_value gives element-wise product of np array
        self._v = self._beta2 * self._v + (1-self._beta2) * (self._grad_value*self._grad_value)
        self._old_x_value = self._x_value
        self._iteration += 1
        m_hat = self._m / (1-(self._beta ** self._iteration))
        v_hat = np.array(self._v / (1-(self._beta2 ** self._iteration)))
        v_hat_aug = v_hat**(0.5) + self._epsilon
        adam_grad = m_hat / v_hat_aug
        self._x_value = self._x_value - self._step_size * adam_grad
        self._converged_value = self._converged(self._x_value, self._old_x_value)
        yield self.state_dict()
```

case the starting point is $(x, y) = (6, 8)$. The convergence condition is when the maximum change goes below 0.001; $max(|x_n - x_{n+1}|, |y_n - yn + 1|) < 0.001$.

A feature of the $g(x, y)$ function is that its derivative w.r.t. x is 0 for all $x \leq 5$, which means that for starting values with $x < 5$, the gradient descent updates only move in one direction. These cases are not discussed further here.

2.1 (b) (i) RMSProp hyperparameters

The RMSProp step size has two hyperparameters, α, and β (ignoring ϵ for this discussion). The α parameter is the unadjusted step size, and the β parameter controls how quickly previous gradients decay in the calculation of subsequent step sizes. A lower β means faster decay, so in the limit as β goes to 0, RMSProp approaches the constant step size algorithm. As we increase β we assign more weight to the gradients from older iterations in terms of the next step size.

Various runs of gradient descent with RMSProp are presented in Figure 3, with function f on the left and g on the right. The top two plots show the progression of $f(x, y)$ and $g(x, y)$ respectively through iterations. We find that with $\beta = 0.25$ we can push α up to 0.4 without diverging, but with $\beta = 0.9$ we diverge, for f. Similarly for g the lower $\beta = 0.25$ allows us to push α up to 0.5 without diverging, whereas the higher $\beta = 0.9$ causes divergence.

An interesting behaviour of the RMSProp step size is that it can suddenly explode when the gradient is very small on successive iterations. This is caused by the denominator of α_t becoming very small. This means that when the estimate gets very close to the true minimum (and hence the gradient is very small for f) the step size can explode, causing the algorithm to start exploring again.

2.2 (b) (ii) Heavy Ball hyperparameters

Heavy Ball also has two hyperparameters, α and β, but does include the normalization term $\frac{1}{(1-\beta)}$, as in RMSProp. As the number of iterations increases the step can be factored to approximately $(1+\beta+\beta^2+...) = 1/(1-\beta)$, which means we are effectively scaling the baseline α parameter: $\alpha/(1-\beta)$, which means a larger β results in larger steps, once we have warmed up with some iterations. In the top left plot of Figure 4 we see that the steps stay larger for longer when $\beta = 0.9$ compared to $\beta = 0.25$.

Looking at the green line in the top right plot of Figure 4 we see that the movement starts of going left, but after x is less than 5 the gradient in the x axis is 0 and so the momentum gradually decreases.

2.3 (b) (iii) Adam hyperparameters

Various runs of gradient descent with different Adam hyperparameters are presented in Figure 5. The β_1 parameter controls how quickly the momentum term decays, i.e. how long the memory is for calculating momentum. The momentum for each dimension i is divided by another term $\epsilon + \sqrt{\hat{v}_i}$, which is reminiscent of RMSProp. This term is also calculated in terms of previous gradients, with length of memory controlled by β_2.

Of the parameters tested we find the best choice (while fixing $\alpha = 0.01$) for optimizing f is $\beta_1 = 0.25$ and $\beta_2 = 0.25$. Increasing β_1 to 0.9 seems promising (initially optimizing faster than $\beta_1 = 0.25$), but is cut off by our aggressive convergence condition.

For optimizing g the choice of β_1 and β_2 has little impact.

Increasing α is generally beneficial, so long as it does not cause divergence, and the choice of β_1 and β_2 influences how high we can push α without diverging.

3 (c)

3.1 (c) (i) ReLU with $x = -1$

The derivative of the ReLU function is 0 for all $x < 0$. Therefore all of the gradient descent step sizes, RMSProp, Adam, Heavy Ball, result in the optimization procedure converging instantly, regardless of hyperparameters. As it happens $x = -1$ also minimizes ReLU.

3.2 (c) (ii) ReLU with $x = 1$

When we start at $x = 1$ the gradient is non-zero so we actually have some movement. For each of the algorithms the first step is small, and get's larger on successive iterations, until the point

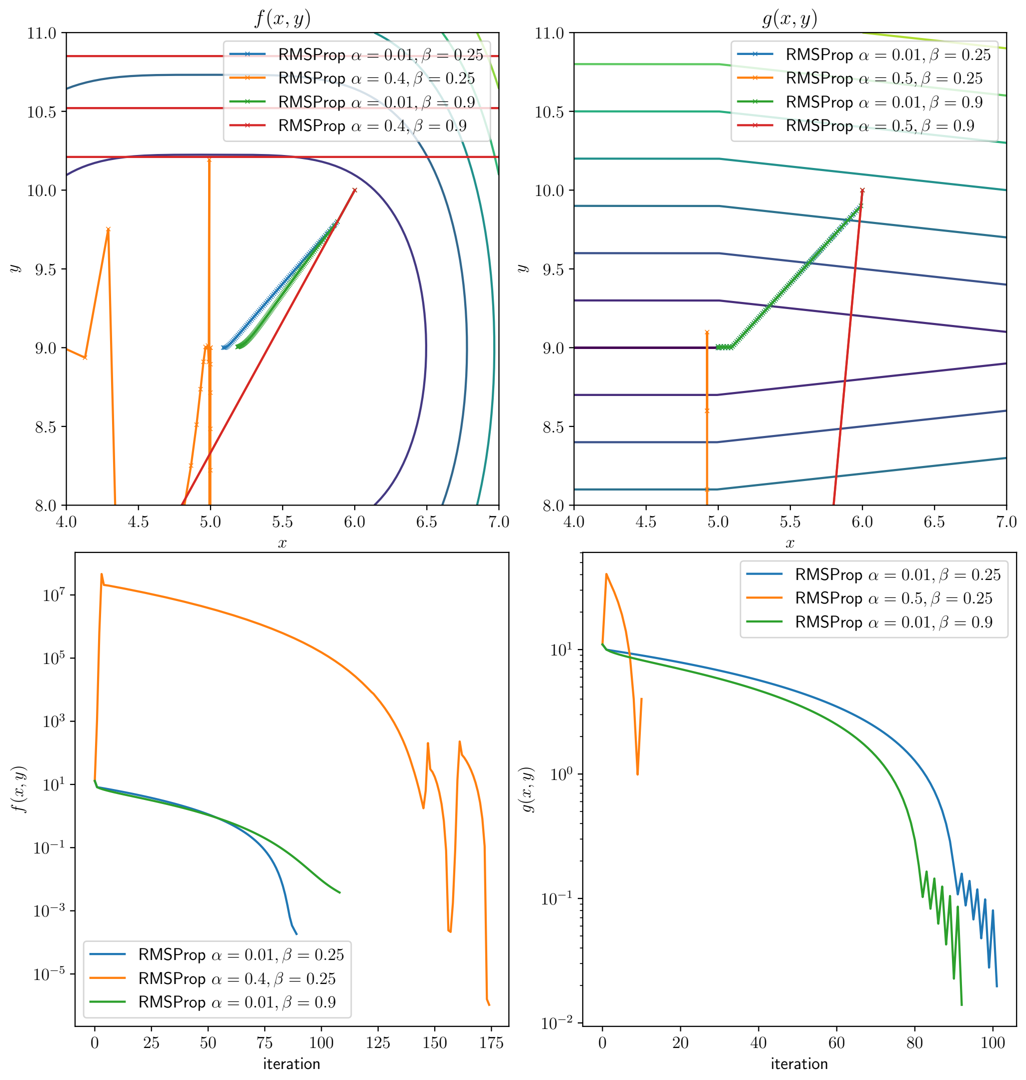


Figure 3: Various runs of gradient descent with different RMSProp hypers.

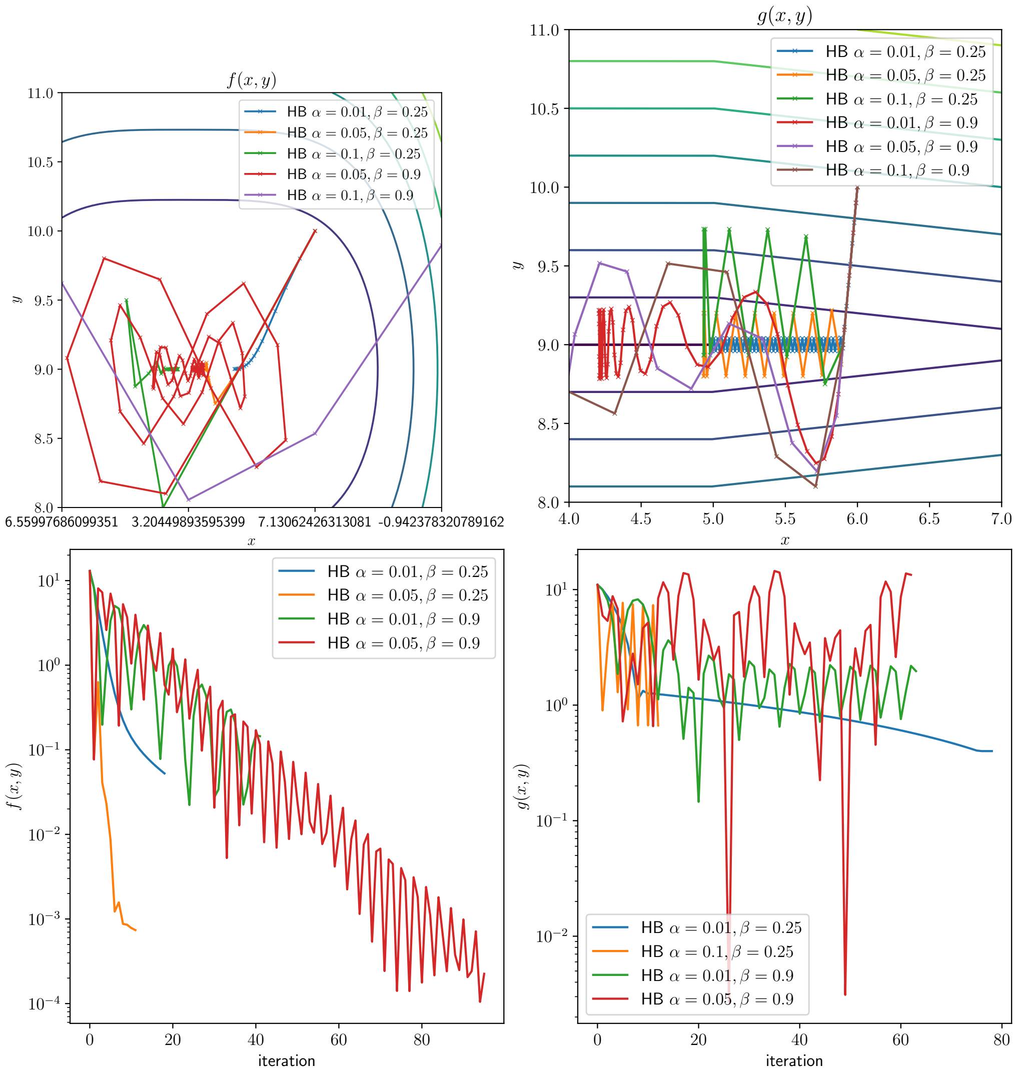


Figure 4: Various runs of gradient descent with different Heavy Ball hypers.

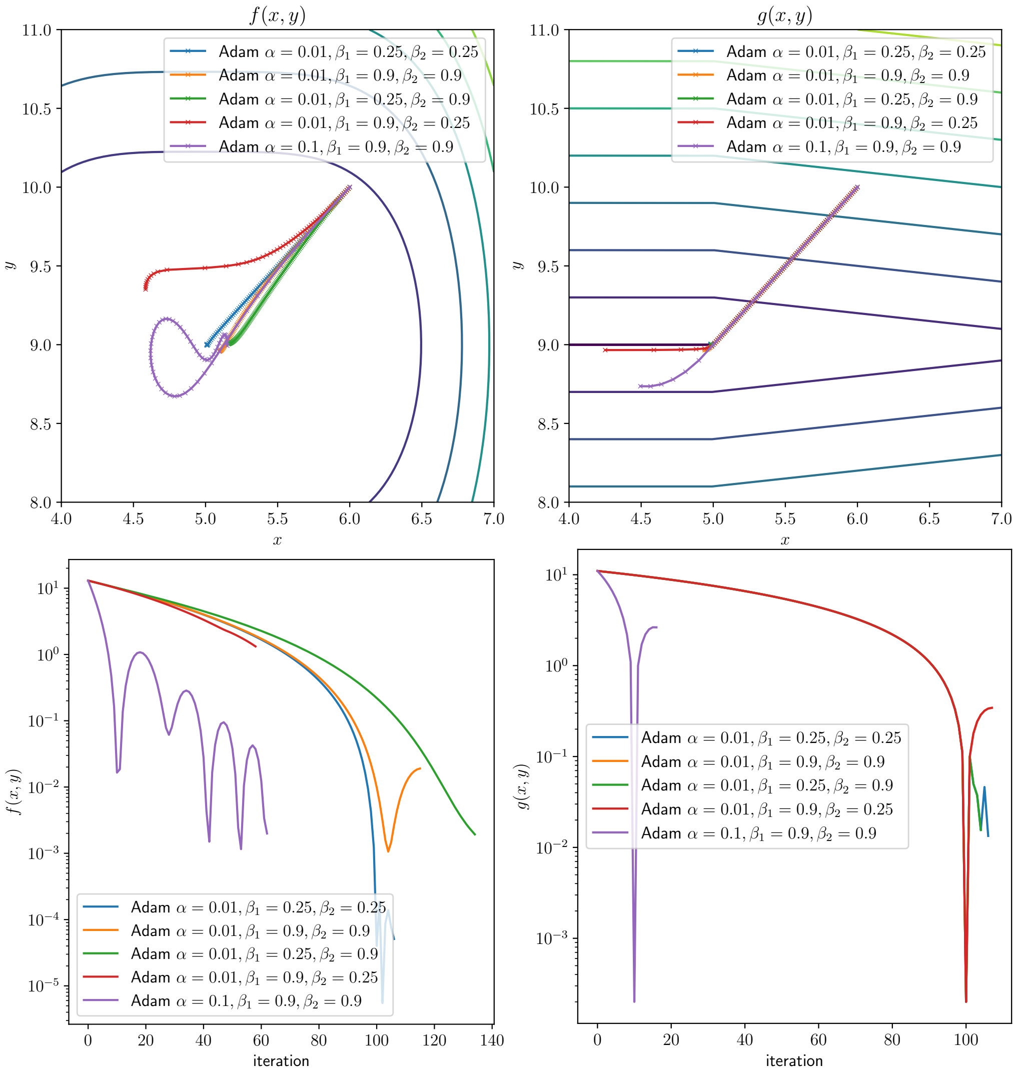


Figure 5: Various runs of gradient descent with different Adam hypers.

where $x < 0$. At this point the steps start to get smaller because each of the algorithms more heavily discounts gradients further in the past. Since each successive gradient after $x < 0$ will be 0 the step reduces until the convergence condition is met.

In the case of RMSProp with $\beta = 0.9$ and $\alpha = 0.1$, the total number of iterations is 8, but there is only one extra iteration after $x < 0$.

For Heavy Ball with $\beta = 0.9$ and $\alpha = 0.1$, the function is minimized in only 5 iterations, but it takes 63 iterations for the convergence condition to be met and for the algorithm to terminate. The final estimate with this Heavy Ball configuration is ≈ -4. The ReLU function is acting like a ramp down which a ball rolls, giving it momentum to glide some distance along the flat surface ($x < 0$).

The Adam run with $\beta_1 = 0.9$ and $\beta_2 = 0.25$, $\alpha = 0.1$, and $x_0 = 1$, starts off with essentially constant step sizes, taking a step of ≈ 0.1 for the first 10 steps. In this configuration it takes 139 iterations to reach the convergence condition, with x monotonically decreasing, and the final estimate being -1746. The reason for this is that the v term becomes very small amplifying the step size, but because β_1 and β_2 are different, the m term is not decreasing as quickly. Eventually m becomes small enough to counteract the miniscule v.

3.3 (c) (iii) ReLU with $x = 100$

In the case of RMSProp with $\beta = 0.9$, $\alpha = 0.1$, and $x_0 = 100$, the total number of iterations is 992, and again there is only one extra iteration after $x < 0$. The final estimate is -0.004. The steps taken on each iteration are approximately constant at 0.1.

For Heavy Ball with $\beta = 0.9$, $\alpha = 0.1$, and $x = 100$, the function is minimized after only 109 iterations, much faster than RMSProp with the same start. However, it takes 175 iterations for the convergence condition to be met and for the algorithm to terminate. The final estimate with this Heavy Ball configuration is ≈ -9. The step size starts out small at 0.1 and increases to about 1 by iteration 109, which is the derivative of ReLU for $x > 0$.

The Adam run with $\beta_1 = 0.9$ and $\beta_2 = 0.25$, $\alpha = 0.1$, and $x_0 = 100$, takes 1000 iterations until $x < 0$. starts off with essentially constant step sizes, taking a step of ≈ 0.1 for the first 1000 steps. In this configuration it takes 1133 iterations to reach the convergence condition, with x monotonically decreasing, and the final estimate being -2431. The reason for this is that the v term becomes very small amplifying the step size, but because β_1 and β_2 are different, the m term is not decreasing as quickly. Eventually m becomes small enough to counteract the miniscule v. The final value of m is about 10^{-6}, whereas the final value of v is about 10^{-80}.