Advanced Deep Learning for Physics
Exercise 2
Convergence rate and Momentum

Andrea Scaioli

May 2025

1 Convergence Rate

1.1 (a) Solution:

The loss function $L(x) = ax^2$ with $a > 0$. The algorithm to be convergent must satisfy the two following conditions:

$$\begin{cases} |x[i+1] - x^*| \leq c\,|x[i] - x^*| \\ 0 \leq c < 1 \end{cases}$$

where $x[i+1] = x[i] - \eta \left.\frac{\partial L}{\partial x}\right|_{x[i]}$.
If we assume $|x[i] - x^*| \neq 0$, or we have already found the minimum, we can compute:

$$0 \leq \frac{|x[i+1] - x^*|}{|x[i] - x^*|} < 1 \Leftrightarrow 0 \leq \frac{\left|x[i] - \eta \left.\frac{\partial L}{\partial x}\right|_{x[i]} - x^*\right|}{|x[i] - x^*|} < 1$$

Considering that $\frac{\partial L}{\partial x} = 2ax$ and $x^* = 0$ we obtain:

$$0 \leq \left|\frac{(1 - 2\eta a)\,x[i]}{x[i]}\right| < 1 \Leftrightarrow 0 \leq |1 - 2\eta a| < 1$$

So the convergence rate is $\boxed{c = |1 - 2\eta a|}$

$$\begin{cases} 0 \leq |1 - 2\eta a| \\ |1 - 2\eta a| < 1 \end{cases} \Leftrightarrow \begin{cases} \textbf{always true} \\ -1 < 1 - 2\eta a < 1 \end{cases}$$

so in the end we obtain:

$$0 < 2\eta a < 2 \Leftrightarrow 0 < \eta < \frac{1}{a}$$

The Learning rate interval that produce a convergent algorithm is: $\boxed{0 < \eta < \frac{1}{a}}$

Now let us study $c = 0$ to obtain the best learning rate η^* and the best convergence rate c^*:

$$|1 - 2\eta^* a| = 0 \Rightarrow \boxed{\eta^* = \frac{1}{2a}\textbf{and } c^* = 0}$$

1.2 (b) Solution:

The loss function $L(x_1, x_2) = ax_1^2 + bx_2^2$ with $0 < a < b$

Calculation of the convergence rate c and the learning rate interval:

$$0 \leq c < 1 \Leftrightarrow 0 \leq \frac{\|x[i] - \eta \left.\frac{\partial L}{\partial x}\right|_{x[i]} - x^*\|}{\|x[i] - x^*\|} < 1$$

Assuming $\|x[i] - x^*\| \neq 0$. $x^* = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ so we obtain:

$$0 \leq \frac{\|x[i] - \eta \left.\frac{\partial L}{\partial x}\right|_{x[i]}\|}{\|x[i]\|} < 1 \Leftrightarrow \|x[i] - \eta \left.\frac{\partial L}{\partial x}\right|_{x[i]} \| < \|x[i]\|$$

Because the norm is always positive ($\|\cdot\| \geq 0$ **always**).

$$\|\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} 2\eta a x_1 \\ 2\eta b x_2 \end{bmatrix}\| < \|\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\| \Leftrightarrow \|\begin{bmatrix} 1 - 2\eta a & 0 \\ 0 & 1 - 2\eta b \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\| < \|\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\| \Leftrightarrow \|\begin{bmatrix} (1 - 2\eta a)\, x_1 \\ (1 - 2\eta b)\, x_2 \end{bmatrix}\| < \|\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\|$$

From the previous line, applying the property $\boldsymbol{Cauchy - Schwart}$ of the norm, we can calculate c:

$$\|x[i] - \eta \left.\frac{\partial L}{\partial x}\right|_{x[i]} \| = \|\begin{bmatrix} 1 - 2\eta a & 0 \\ 0 & 1 - 2\eta b \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\| \leq \|\begin{bmatrix} 1 - 2\eta a & 0 \\ 0 & 1 - 2\eta b \end{bmatrix}\| \cdot \|\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\|$$

Comparing this equation with the definition of c results that:

$$\boxed{c = \|\begin{bmatrix} 1 - 2\eta a & 0 \\ 0 & 1 - 2\eta b \end{bmatrix}\| = \max\left[|1 - 2\eta a|, |1 - 2\eta b|\right]}$$

Computing the norm explicitly with the 2D Euclidean norm definition we obtain:

$$(1 - 2\eta a)^2 x_1^2 + (1 - 2\eta b)^2 x_2^2 < x_1^2 + x_2^2 \Leftrightarrow (a\eta - 1)\, \eta a x_1^2 + (b\eta - 1)\, \eta b x_2^2 < 0 \Rightarrow$$

$$\Rightarrow \begin{cases} 0 < \eta < \frac{1}{a} \\ 0 < \eta < \frac{1}{b} \end{cases} \Rightarrow \boxed{0 < \eta < \frac{1}{b}}$$

Taking into account that $0 < a < b$.

Now we want to calculate the best $\eta*$ and c^*:

$$c^* = \min_\eta \left[c\,[\eta]\right] = \min_\eta \left[\max\left[|1 - 2\eta a|, |1 - 2\eta b|\right]\right]$$

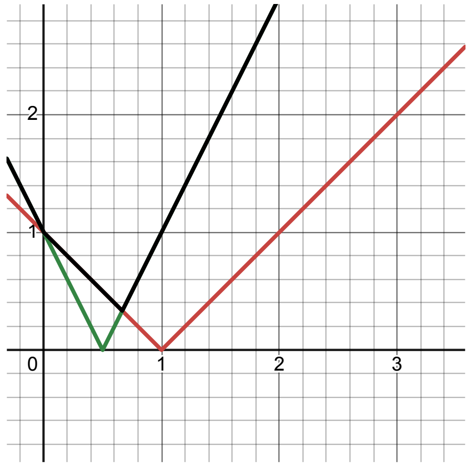

Figure 1: Qualitative graph of the two V-shaped function in red and green ad of the max function in black

This equation can be solved by searching the intersection between the two V-shaped curves as shown in the following figure above, so:

$$1 - 2\eta a = -1 + 2\eta b \Rightarrow \boxed{\eta * = \frac{1}{a+b}}$$

and

$$c^* = \| \begin{bmatrix} 1-2\eta^* a & 0 \\ 0 & 1-2\eta^* b \end{bmatrix} = \begin{bmatrix} \frac{b-a}{a+b} & 0 \\ 0 & \frac{a-b}{a+b} \end{bmatrix} \| = \frac{b-a}{a+b}$$

so: $\boxed{c^* = \frac{b-a}{b+a}}$

2 Gradient Descent and its Acceleration with Momentum

2.1 Solution for trajectory [A]

The points we will use are: $x[0] = \begin{bmatrix} -1 \\ -1 \end{bmatrix}$, $x[1] = \begin{bmatrix} ? \\ -0.5 \end{bmatrix}$, $x[2] = \begin{bmatrix} ? \\ 0.1 \end{bmatrix}$.
From the curvature of the trajectory, it seems that a momentum algorithm is been used, so we try to calculate η and m:

2.1.1 calculation of η:

$$x[1] = x[0] + v[1] \Leftrightarrow \begin{bmatrix} ? \\ 0.5 \end{bmatrix} = \begin{bmatrix} -1 \\ -1 \end{bmatrix} + \begin{bmatrix} v_{x_1}[1] \\ v_{x_2}[1] \end{bmatrix}$$

$$\begin{bmatrix} v_{x_1}[1] \\ v_{x_2}[1] \end{bmatrix} = \begin{bmatrix} ? \\ 0.5 \end{bmatrix} = m \cdot \begin{bmatrix} v_{x_1}[1] \\ v_{x_2}[1] \end{bmatrix} - \frac{\partial L}{\partial x}\bigg|_{x[0]} = \eta \begin{bmatrix} 2 \\ 8 \end{bmatrix}$$

$$\Rightarrow \boxed{\eta = \frac{0.5}{8} = 0.0625}$$

Knowing η we can compute
$v_{x_1}[1] = 0.0625 \cdot 2 = 0.125$ and $x_1[1] = -1 + 0,125 = -0.875$ so:

$$v[1] = \begin{bmatrix} 0.125 \\ 0.5 \end{bmatrix} \textbf{ and } x[1] = \begin{bmatrix} -0.875 \\ -0.5 \end{bmatrix}$$

This is consistent with the point on the grid.

2.1.2 calculation of m:

$$x[2] = x[1] + v[2] \Leftrightarrow \begin{bmatrix} ? \\ 0.1 \end{bmatrix} = \begin{bmatrix} -0.875 \\ -0.5 \end{bmatrix} + \begin{bmatrix} v_{x_1}[2] \\ v_{x_2}[2] \end{bmatrix}$$

$$\Rightarrow v_{x_2}[2] = 0.1 + 0.5 = 0.6$$

$$v[2] = mv[1] - \eta \frac{\partial L}{\partial x}\bigg|_{x[1]} = m \begin{bmatrix} 0.125 \\ 0.5 \end{bmatrix} + \begin{bmatrix} 0.109 \\ 0.25 \end{bmatrix}$$

$$\Rightarrow \boxed{m = \frac{0.6 - 0.25}{0.5} = 0.7}$$

To verify the correctness of out results let's try to compute the next step:

$$x[3] = x[2] + v[3] = x[2] + mv[2] - \eta \frac{\partial L}{\partial x}\bigg|_{x[2]}$$

$$x[3] = \begin{bmatrix} -0.678 \\ 0.1 \end{bmatrix} + 0.7 \begin{bmatrix} 0.197 \\ 0.6 \end{bmatrix} - 0.0625 \begin{bmatrix} 2 \cdot -0.678 \\ 8 \cdot 0.1 \end{bmatrix} = \begin{bmatrix} -0.455 \\ 0.47 \end{bmatrix}$$

This is consistent with the point on the grid.

2.2 Solution for trajectory [B]

The algorithm is a Gradient Descent, so we can compute η:

$$x[1] = x[0] - \eta \frac{\partial L}{\partial x}\bigg|_{x[0]} = \begin{bmatrix} -1 \\ 0.5 \end{bmatrix} + \eta \begin{bmatrix} 2 \\ -4 \end{bmatrix} \Leftrightarrow \begin{bmatrix} ? \\ 0 \end{bmatrix} = \begin{bmatrix} -1 \\ 0.5 \end{bmatrix} + \eta \begin{bmatrix} 2 \\ -4 \end{bmatrix}$$

$$\Rightarrow \boxed{\eta = \frac{-0.5}{-4} = 0.125}$$

Knowing η we can compute $x_1[1] = -1 + 0.125 = -0.75$ so:

$$x[1] = \begin{bmatrix} -0.75 \\ 0 \end{bmatrix}$$

This is consistent with the point on the grid. Let's try the next step:

$$x[2] = x[1] - \eta \left.\frac{\partial L}{\partial x}\right|_{x[1]} = \begin{bmatrix} -0.75 \\ 0 \end{bmatrix} - 0.125 \begin{bmatrix} 2 \cdot -0.75 \\ 0 \end{bmatrix} = \begin{bmatrix} -0.562 \\ 0 \end{bmatrix}$$

This point is consistent too.

2.3 Solution for trajectory [C]

We can do the same as for the trajectory [B], so we can compute η:

$$x[1] = x[0] - \eta \left.\frac{\partial L}{\partial x}\right|_{x[0]} \Rightarrow \begin{bmatrix} 0.6 \\ -0.6 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} - \eta \begin{bmatrix} 2 \\ 8 \end{bmatrix} \Rightarrow \eta \begin{bmatrix} 2 \\ 8 \end{bmatrix} = \begin{bmatrix} 0.4 \\ 1.6 \end{bmatrix}$$

$$\Rightarrow \boxed{\eta = \frac{0.4}{2} = \frac{1.6}{8} = 0.2}$$

The next step is:

$$x[2] = x[1] - \eta \left.\frac{\partial L}{\partial x}\right|_{x[1]} \Leftrightarrow \begin{bmatrix} 0.6 \\ -0.6 \end{bmatrix} - 0.2 \begin{bmatrix} 2 \cdot 0.6 \\ 8 \cdot -0.6 \end{bmatrix} = \begin{bmatrix} 0.36 \\ 0.36 \end{bmatrix}$$

This is consistent with the point on the grid.