Université Paris Saclay	**MSc AI & MSc Data Science.**
TC2-Optimization for Machine Learning	Faïcel Chamroukhi

TD/TP - 2 - November 2024.

Solution 1. Gradient Descent

1. Calculate the gradient:

   - Given the function $f(x) = x^2 + 4x + 4$, the gradient is computed as follows:

   $$\nabla f(x) = \frac{d}{dx}(x^2 + 4x + 4) = 2x + 4.$$

2. Perform two steps of Gradient Descent:

   - We start from an initial point $x_0 = 2$.
   - The update rule for gradient descent is given by

   $$x_{k+1} = x_k - \alpha \nabla f(x_k)$$

   , where α is the step size.
   - Setting $\alpha = 0.1$:

   $$x_1 = x_0 - \alpha \nabla f(x_0) = 2 - 0.1 \times (2 \times 2 + 4) = 2 - 0.1 \times 8 = 1.2,$$
   $$x_2 = x_1 - \alpha \nabla f(x_1) = 1.2 - 0.1 \times (2 \times 1.2 + 4) = 1.2 - 0.1 \times 6.4 = 0.56.$$

   - The sequence $x_0 = 2$, $x_1 = 1.2$, and $x_2 = 0.56$ shows the progression towards minimizing $f(x)$.

Solution 2. Least Squares Function

1. **Gradient of the Least Squares Function:**

   - We can write $f(\mathbf{w}) = \frac{1}{2}\|y - X\mathbf{w}\|^2$ as:

   $$\begin{aligned} f(\mathbf{w}) &= \frac{1}{2}(y - X\mathbf{w})^\top (y - X\mathbf{w}) \\ &= \frac{1}{2}\left(y^\top y - y^\top X\mathbf{w} - \mathbf{w}^\top X^\top y + \mathbf{w}^\top X^\top X\mathbf{w}\right) \\ &= \frac{1}{2}\left(y^\top y - 2y^\top X\mathbf{w} + \mathbf{w}^\top X^\top X\mathbf{w}\right) \qquad (1) \end{aligned}$$

   Since the two terms $-y^\top X\mathbf{w} = -\mathbf{w}^\top X^\top y$ are scalars and equal.
   - To compute the gradient $\nabla f(\mathbf{w}) = \frac{\partial f(\mathbf{w})}{\partial \mathbf{w}}$, we differentiate. We can differentiate (1) term by term:

- $\frac{\partial}{\partial \mathbf{w}}\left(\frac{1}{2}y^\top y\right) = 0$ as this term is a constant with respect to $\mathbf{w}$.
- $\frac{\partial}{\partial \mathbf{w}}\left(-y^\top X\mathbf{w}\right) = -X^\top y$ as this term is linear in $\mathbf{w}$, so its derivative with respect to $\mathbf{w}$ is $-X^\top y$.
  We have the property:

$$\frac{\partial(a^\top x)}{\partial x} = \frac{\partial(x^\top a)}{\partial x} = a.$$

- $\frac{\partial}{\partial \mathbf{w}}\left(\frac{1}{2}\mathbf{w}^\top X^\top X\mathbf{w}\right) = X^\top X\mathbf{w}$: This term is quadratic in $\mathbf{w}$, so its derivative with respect to $\mathbf{w}$ yields $X^\top X\mathbf{w}$.
  We have the property:

$$\frac{\partial(x^\top Ax)}{\partial x} = (A + A^\top)x.$$

  If A is symmetric (which is the case for $X^\top X$), then:

$$\frac{\partial(x^\top Ax)}{\partial x} = 2Ax.$$

Combining these results, we obtain:

$$\nabla f(\mathbf{w}) = -X^\top y + X^\top X\mathbf{w} = -X^\top(y - X\mathbf{w}) \tag{2}$$

2. **Hessian of the Least Squares Function:**

- The Hessian of $f(\mathbf{w})$ is given by the second derivative: By differentiating the gradient (2) w.r.t $\mathbf{w}$ we get

$$\nabla^2 f(\mathbf{w}) = X^\top X$$

  since

$$\frac{\partial(Ax)}{\partial x} = A$$

  where A is a matrix and x is a vector.
- Since $X^\top X$ is positive semi-definite ($\forall z, z^\top X^\top Xz = \|Xz\|^2 \geq 0$), the function $f(\mathbf{w})$ is convex. This confirms that the least squares problem is a convex optimization problem.
- If the matrix X is of full rank, then $X^\top X$ is not only positive semi-definite but also positive definite. This means $X^\top X$ has strictly positive eigenvalues, ensuring a unique global minimum for the least squares problem. The full rank condition implies that the columns of X are linearly independent, which guarantees that $X^\top X$ is invertible.