Nabla Operator: Gradients in $\mathbb{R}^n$ are conveniently expressed with the *Nabla operator*:

$$\nabla \equiv \begin{pmatrix} \partial/\partial x_1 \\ \partial/\partial x_2 \\ \vdots \\ \partial/\partial x_n \end{pmatrix} \quad \text{or applied to some } f(\boldsymbol{x}) \quad \nabla f(\boldsymbol{x}) = \begin{pmatrix} \partial f/\partial x_1 \\ \partial f/\partial x_2 \\ \vdots \\ \partial f/\partial x_n \end{pmatrix}$$

Hessian Matrix: The Hessian is a square matrix of 2nd order partial derivatives of a scalar-valued function and a measure of local curvature:

$$H_f = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1\,\partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1\,\partial x_n} \\ \frac{\partial^2 f}{\partial x_2\,\partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2\,\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n\,\partial x_1} & \frac{\partial^2 f}{\partial x_n\,\partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{pmatrix}$$

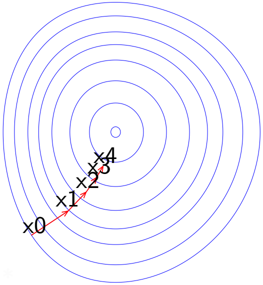
Gradient Descent

Gradient Descent (GD) is a first-order iterative optimization algorithm for finding a local minimum of a differentiable function:

- Assume $F(\boldsymbol{x})$ is defined and locally differentiable around $\boldsymbol{a}$
- Then $F(\boldsymbol{x})$ decreases fastest in the direction of the *negative* gradient of F at a, which is $-\nabla F(\boldsymbol{a})$
- For a sufficiently small step size (learning rate) $\gamma \in \mathbb{R}^+$, we have

$$\boldsymbol{a}_{n+1} = \boldsymbol{a}_n - \gamma \nabla F(\boldsymbol{a}_n) \quad \text{and} \quad F(\mathbf{a_n}) \geq F(\mathbf{a_{n+1}})$$

Iterating the above we get the



Gradient Descent Update Rule:

$$\boldsymbol{x}_{n+1} = \boldsymbol{x}_n - \gamma_n \nabla F(\boldsymbol{x}_n), \; n \geq 0$$

where $F(\boldsymbol{x}_0) \geq F(\boldsymbol{x}_1) \geq F(\boldsymbol{x}_2) \geq \ldots$ is a monotonic sequence