Linear Algebra

Gradient Descent

Michael Ruddy

The Gradient

$f: \mathbb{R}^N \to \mathbb{R}$ $\quad (J_f^T)$

$f(x_1, x_2, \ldots, x_N)$

The gradient of f at a $(a_1, \ldots, a_N)$

is $\nabla f(a) = \begin{bmatrix} \partial f/\partial x_1(a) \\ \partial f/\partial x_2(a) \\ \vdots \\ \partial f/\partial x_N(a) \end{bmatrix}$

The Gradient $f(x, y) = z$

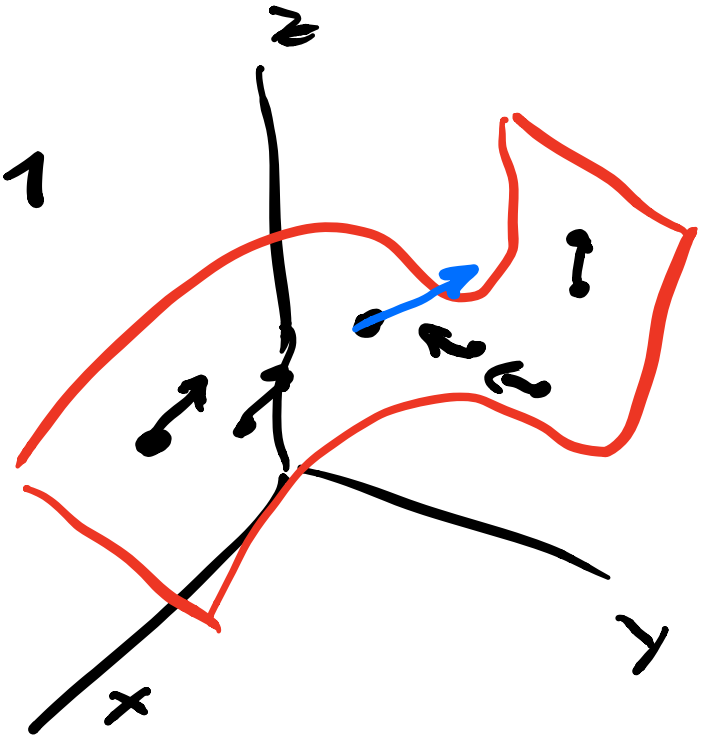
"steepest slope"

$$\nabla f = \begin{bmatrix} \partial f/\partial x \\ \partial f/\partial y \end{bmatrix}_{2 \times 1}$$

gradient $= \vec{0}$

$\updownarrow$

critical pt.



The Gradient

$$f(x,y,z) = 3x^2y + zy - x$$

at $(1,1,1)$

$$\nabla f = \begin{bmatrix} 6xy - 1 \\ 3x^2 + z \\ y \end{bmatrix} \qquad \nabla f(1,1,1) = \begin{bmatrix} 5 \\ 4 \\ 1 \end{bmatrix}$$

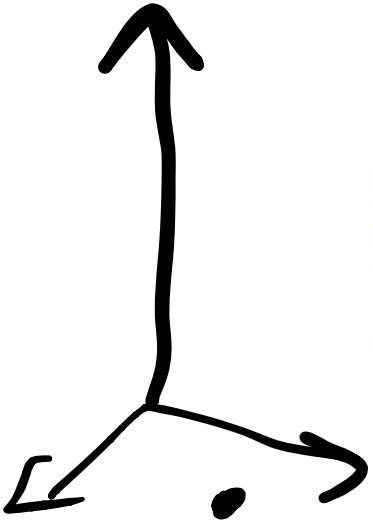
Gradient Descent

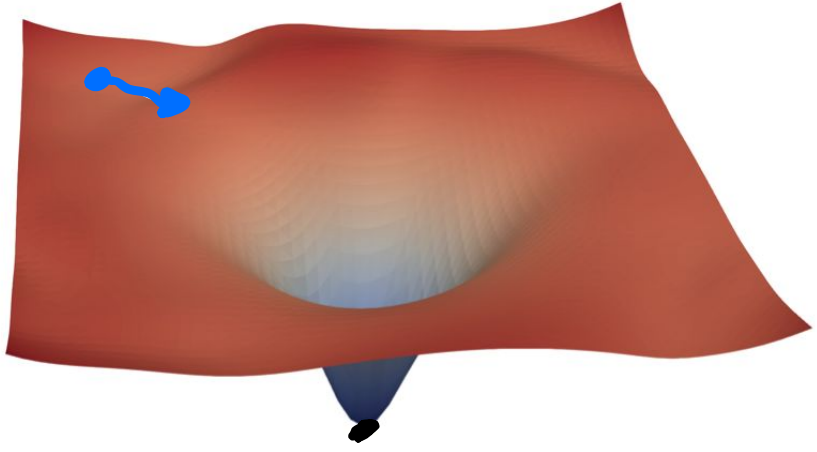
Idea: $f(x_1, \ldots, x_N)$

- Want to minimize f
- Negative gradient is fastest decrease.
- Starts w/ a value, use gradient to take steps towards the minimum.

Gradient Descent

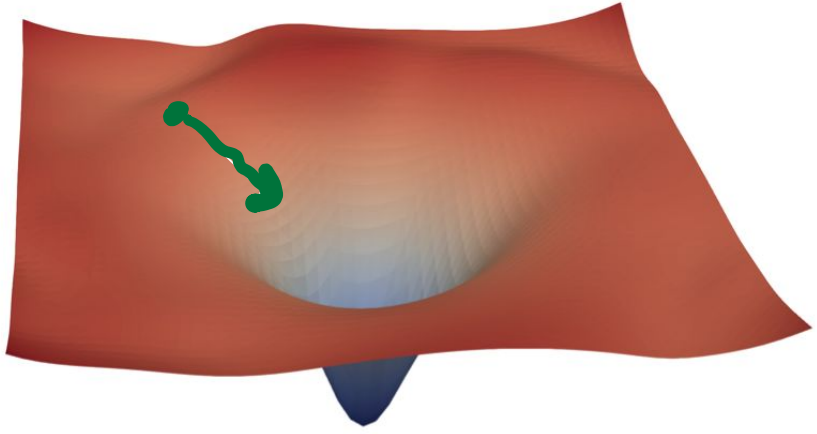
$f(x, y) = z$





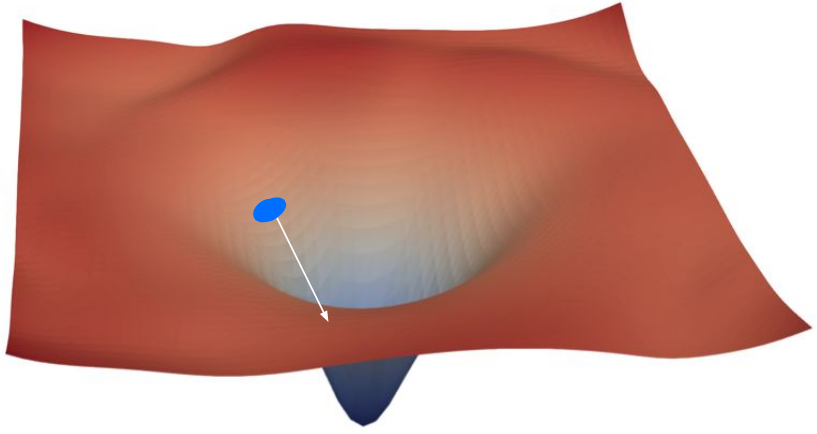
From *Visualizing the Loss Landscape of Neural =ets* by
H. Li, Z. Xu, G. Taylor, C. Studer, T. Goldstein

Gradient Descent



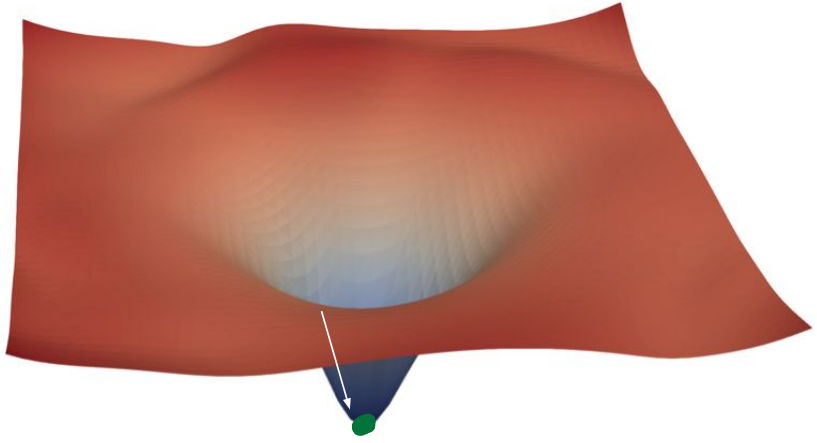
From *Visualizing the Loss Landscape of Neural =ets* by
H. Li, Z. Xu, G. Taylor, C. Studer, T. Goldstein

Gradient Descent



From *Visualizing the Loss Landscape of Neural Nets* by
H. Li, Z. Xu, G. Taylor, C. Studer, T. Goldstein

Gradient Descent



From *Visualizing the Loss Landscape of Neural Nets* by
H. Li, Z. Xu, G. Taylor, C. Studer, T. Goldstein

Gradient Descent

$$a_N = a_{N-1} - \gamma \nabla f(a_{N-1})$$

- Gamma γ is step size. (learning rate)

a_0 - best guess OR random starting point.

Example

Two iterations of gradient descent

$f(x,y) = 3x^2 + y^2$ Starting at $(1,1)$ a_0

w/ $\gamma = 0.1$

$$\nabla f = \begin{bmatrix} 6x \\ 2y \end{bmatrix}$$

$$a_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} - 0.1 \begin{bmatrix} 6 \\ 2 \end{bmatrix} = \begin{bmatrix} 0.4 \\ 0.8 \end{bmatrix}$$

$$a_2 = \begin{bmatrix} 0.4 \\ 0.8 \end{bmatrix} - 0.1 \begin{bmatrix} 2.4 \\ 1.6 \end{bmatrix} = \begin{bmatrix} 0.16 \\ 0.64 \end{bmatrix}$$

Example