

Last time: mixed up row and column vectors.

Matrix Calculus 101

$$x \in \mathbb{R}^{n \times 1}, y \in \mathbb{R}^{m \times 1}, A \in \mathbb{R}^{m \times n}$$

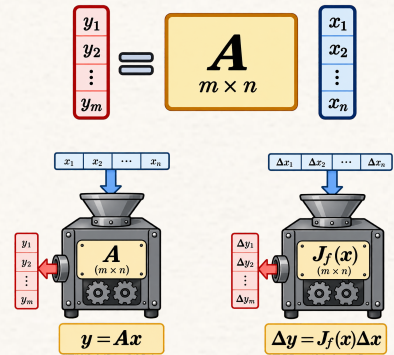
$$y = Ax \rightarrow \frac{\partial y}{\partial x} = A$$

$$y = x^T A \rightarrow \frac{\partial y}{\partial x} = A^T$$

$$y = x^T x \rightarrow \frac{\partial y}{\partial x} = 2x^T$$

$$y = x^T A x \rightarrow \frac{\partial y}{\partial x} = x^T (A + A^T) = 2x^T A \text{ if } A \text{ symmetric}$$

$$\text{Chain rule: } y = f(g(x)) \rightarrow \frac{\partial y}{\partial x} = \frac{\partial y}{\partial g} \frac{\partial g}{\partial x}$$



Note: for scalar outputs, these give $(1, n)$ row-vector Jacobians
transpose results to get $(n, 1)$ column gradients

$$X \in \mathbb{R}^{m \times n}, w \in \mathbb{R}^{n \times 1}, y \in \mathbb{R}^{m \times 1}$$

$$\nabla_w L(w) = \nabla_w \|Xw - y\|^2$$

$$= \nabla_w (Xw - y)^T (Xw - y)$$

$$= 2 (Xw - y)^T X \quad (1, n)$$

$$\text{Set } \nabla_w L(w) \text{ to } 0: \quad 2(Xw - y)^T X = 0 \quad (1, n)$$

$$2X^T (Xw - y) = 0 \quad (n, 1)$$

$$2X^T X w - 2X^T y = 0 \quad (n, 1)$$

$$w = (X^T X)^{-1} X^T y \quad (n, 1)$$

1.1 Definitions

$$f: \mathbb{R}^n \rightarrow \mathbb{R}, \quad g: \mathbb{R}^n \rightarrow \mathbb{R}^m$$

Gradient of f with respect to x evaluated at x :

$$\nabla_x f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{n \times 1}$$

Jacobian of g with respect to x evaluated at x :

$$J_f(x) \quad \nabla_x g(x) = \begin{bmatrix} \frac{\partial g_1(x)}{\partial x_1} & \dots & \frac{\partial g_1(x)}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial g_m(x)}{\partial x_1} & \dots & \frac{\partial g_m(x)}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{m \times n}$$

Hessian of f w.r.t. x evaluated at x :

$$H_f(x) \quad \nabla_x^2 f(x) = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \dots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \vdots & & \vdots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \dots & \frac{\partial^2 f(x)}{\partial x_n^2} \end{bmatrix} \in \mathbb{R}^{n \times n}$$

Assume gradient is a column vector.

1.2 Approximation

$$\begin{aligned} \text{Recall for } f: \mathbb{R} \rightarrow \mathbb{R}, \quad \frac{df}{dx} &= \lim_{\epsilon \rightarrow 0} \frac{f(x+\epsilon) - f(x)}{\epsilon} \\ &\approx \frac{f(x+\epsilon) - f(x)}{\epsilon} \quad \text{for small } \epsilon \\ \Rightarrow f(x+\epsilon) &\approx f(x) + \epsilon \frac{df}{dx} \end{aligned}$$

Extending to $f: \mathbb{R}^n \rightarrow \mathbb{R}$. if we perturb one coordinate:

$$f(x_1 + \epsilon, x_2, \dots, x_n) \approx f(x_1, \dots, x_n) + \epsilon \frac{\partial f}{\partial x_1}(x_1, \dots, x_n)$$

What if we perturb two coordinates?

$$\begin{aligned}
& f(x_1 + \epsilon_1, x_2 + \epsilon_2, \dots, x_n) \\
&= f(x_1, x_2 + \epsilon_2, \dots, x_n) + \epsilon_1 \frac{\partial f}{\partial x_1}(x_1, x_2 + \epsilon_2, \dots, x_n) \\
&= f(x_1, x_2, \dots, x_n) + \epsilon_2 \frac{\partial f}{\partial x_2}(x_1, x_2, \dots, x_n) + \\
&\quad \epsilon_1 \frac{\partial f}{\partial x_1}(x_1, x_2, \dots, x_n) + \epsilon_1 \epsilon_2 \frac{\partial^2 f}{\partial x_1 \partial x_2}(x_1, x_2, \dots, x_n) \quad \text{H.O.T} \\
&= f(x_1, x_2, \dots, x_n) + \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} \end{bmatrix} \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \end{bmatrix}
\end{aligned}$$

In general, $f(x + \epsilon) \approx f(x) + \nabla_x f(x)^T \epsilon$

(a) $f(x_1, x_2) = x_1^2 + e^{x_1 x_2} + 2 \log(x_2)$. $\nabla f(x_1, x_2) = ?$ $\nabla^2 f(x_1, x_2) = ?$

$$\nabla_x f(x_1, x_2) = \begin{bmatrix} 2x_1 + x_2 e^{x_1 x_2} \\ x_1 e^{x_1 x_2} + 2 \frac{1}{x_2} \end{bmatrix}$$

$$\nabla_x^2 f(x_1, x_2) = \begin{bmatrix} 2 + x_2^2 e^{x_1 x_2} & e^{x_1 x_2} + x_1 x_2 e^{x_1 x_2} \\ e^{x_1 x_2} + x_1 x_2 e^{x_1 x_2} & x_1^2 e^{x_1 x_2} - 2 \frac{1}{x_2^2} \end{bmatrix}$$

(b) What is Jacobian of $\nabla_x f$?

It's just $\nabla_x^2 f(x_1, x_2)$

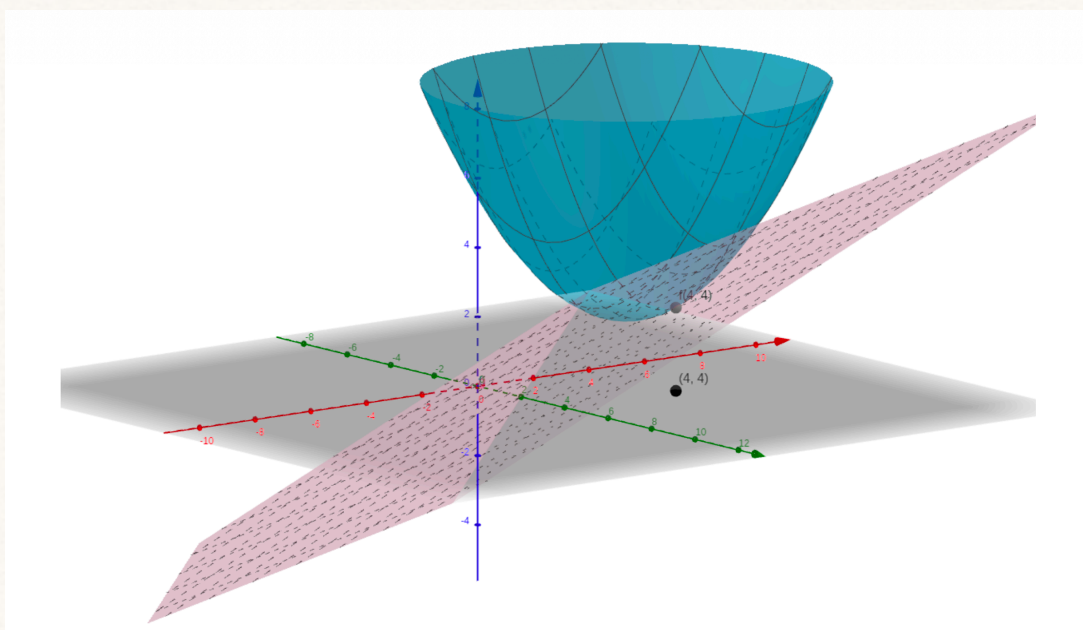
(c) $\nabla_x f(x) =$ best linear approx around $f(x)$
What does Jacobian $\nabla_x g(x)$ offer?

Best linear approx of a vector

$$g(x + \epsilon) \approx g(x) + \nabla_x g(x)^T \epsilon$$

(d) If we had $\nabla_x^2 f(x)$, can we do better?

Best quadratic approx of f : $f(x + \epsilon) \approx f(x) + \epsilon^T \nabla_x f(x) + \frac{1}{2} \epsilon^T \nabla_x^2 f(x) \epsilon$



(e) $f(x_1, x_2) = 2 + 0.2(x_1 - 3)^2 + 0.2(x_2 - 3)^2$

What is the function describing the tangent plane?

$$f(y) = f(x) + \nabla_x f(x)^T (y - x) = 2.4 + \begin{bmatrix} 0.4 \\ 0.4 \end{bmatrix}^T (y - \begin{bmatrix} 4 \\ 4 \end{bmatrix})$$

But wait! ... I heard gradient points towards direction of steepest ascent?

How is it a plane AND a vector?

Let's see: what is the ϵ ^{$\nearrow \|\epsilon\|_2 = 1$} that maximizes the increase in $f(x)$?

$$f(x + \epsilon) \approx f(x) + \epsilon^T \nabla_x f(x)$$

$$\epsilon^* = \underset{\epsilon, \|\epsilon\|_2 = 1}{\operatorname{argmax}} \epsilon^T \nabla_x f(x)$$



$$\begin{aligned} \text{By Cauchy - Schwartz, } |\epsilon^T \nabla_x f(x)| &\leq \underbrace{\|\epsilon\|_2}_{=1} \|\nabla_x f(x)\|_2 \\ &= \|\nabla_x f(x)\|_2 \end{aligned}$$

$$\text{Equality when } \epsilon = \frac{\nabla_x f(x)}{\|\nabla_x f(x)\|_2}$$

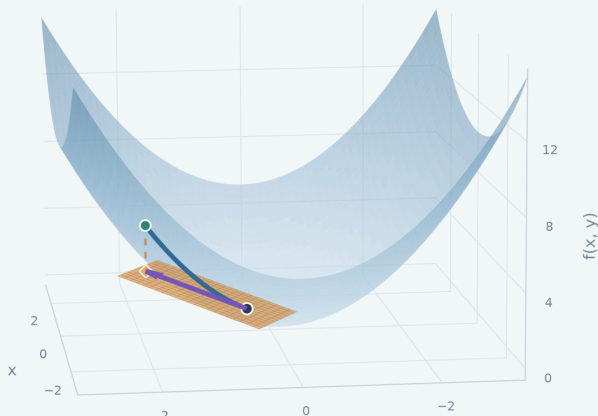
Duality of gradients:

Interpretation: Linear map
 $\Delta x \rightarrow \Delta f(x)$

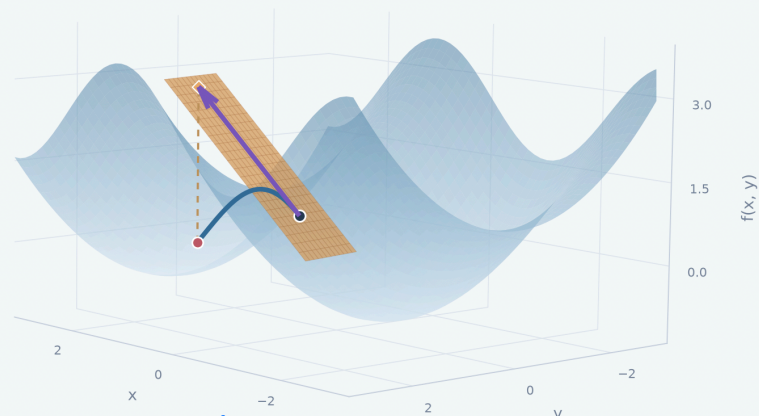
Direction of maximum ascent

Representation: $(1, n)$ row vector
 jacobian

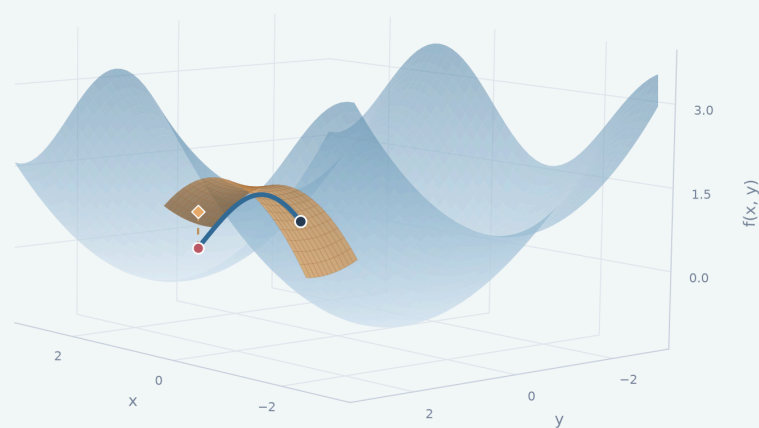
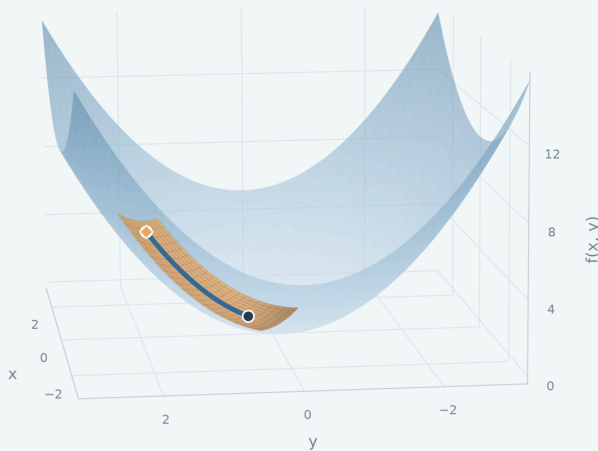
$(n, 1)$ column vector
 (same shape as x)



Gradient accurate



Gradient inaccurate



1.3 Algebra (Jacobian form)

$$\nabla_x c = 0$$

$$\nabla_x (af + bg)(x) = a \nabla_x f(x) + b \nabla_x g(x)$$

$$\nabla_x (f^T g)(x) = g(x)^T \nabla_x f(x) + f(x)^T \nabla_x g(x)$$

$$\nabla_x (f \circ g)(x) = \nabla_x f(g(x))^T \nabla_x g(x)$$

Chain rule $\rightarrow$ Product rule

$$y = f(x)^T g(x)$$

$$\text{Let } h(u, v) = u^T v$$

$$\text{then } y = h(f(x), g(x))$$

$$\begin{aligned}\frac{\partial y}{\partial x} &= \frac{\partial y}{\partial (f(x), g(x))} \frac{\partial (f(x), g(x))}{\partial x} \\&= \begin{bmatrix} \frac{\partial y}{\partial f(x)}^T & \frac{\partial y}{\partial g(x)}^T \end{bmatrix} \begin{bmatrix} \frac{\partial f(x)}{\partial x} \\ \frac{\partial g(x)}{\partial x} \end{bmatrix} \\&= \begin{bmatrix} g(x)^T & f(x)^T \end{bmatrix} \begin{bmatrix} \frac{\partial f(x)}{\partial x} \\ \frac{\partial g(x)}{\partial x} \end{bmatrix} \\&= g(x)^T \frac{\partial f(x)}{\partial x} + f(x)^T \frac{\partial g(x)}{\partial x}\end{aligned}$$

(c) - (d) : see solution

$$(e) \quad f(x) = x^T A x, \quad \text{derive } \nabla_x f(x)$$

$$\begin{aligned}\nabla_x f(x) &= x^T \nabla_x (Ax) + (Ax)^T \nabla_x (x^T) \\&= x^T A + (Ax)^T \\&= x^T A + x^T A^T \\&= x^T (A + A^T) \quad \rightarrow \text{gradient form } (A + A^T) x\end{aligned}$$

$$(f) \quad \text{what is } \nabla_x^2 f(x) ?$$

$$\begin{aligned}\nabla_x^2 f(x) &= \nabla_x (\nabla_x f(x)) \\&= \nabla_x ((A + A^T) x) \\&= A + A^T\end{aligned}$$

(9) Let $f(x) = (Ax - y)^T W (Ax - y)$. What is $\nabla_x f(x)$?

$$\text{Let } g(x) = Ax - y \text{ and } h(z) = z^T W z$$

$$\text{then } f(x) = (h \circ g)(x)$$

$$\nabla_x f(x) = \nabla_g h(g) \nabla_x g(x)$$

$$= (Ax - y)^T (W + W^T) A$$

$$\rightarrow \text{gradient form } A^T (W + W^T) (Ax - y)$$