As You Walk In

Please complete survey

- See Piazza for link

Announcements

HW1

- Due Today, 11:59 pm
- Online + Written components

Probability

- Reach out for help on HW1 probability before Wednesday!

Quiz

- First quiz on Wed 9/22
- Details posted on Piazza

Mathematical Foundations for Machine Learning

Calculus & Lagrange Multipliers

Instructor: Pat Virtue

Linear Regression

Last time

Calculus needed to:

- Find $\mathbf{w}$ that minimizes MSE with model $y = \mathbf{w}^T\mathbf{x}$, $\mathbf{x} \in \mathbb{R}^2$

Today

Quick trick to:

- Find $\mathbf{w}, b$ that minimizes MSE with model $y = \mathbf{w}^T\mathbf{x} + b$, $\mathbf{x} \in \mathbb{R}^M$

Calculus needed for neural networks

- Partial derivatives of various sizes
- Multivariate chain rule

Constrained optimization and Lagrange multipliers

Linear Regression

Quick trick to:

- Find $\mathbf{w}, b$ that minimizes MSE with model $y = \mathbf{w}^T\mathbf{x} + b, \quad \mathbf{x} \in \mathbb{R}^M$

Jump to Lecture 5 slides

Calculus

Partial derivatives

Calculus with Linear Algebra

$f(x, u, z, v)$

Vector in, scalar out

Gradient

$y = f(\mathbf{x}) \qquad y \in \mathbb{R}, \quad \mathbf{x} \in \mathbb{R}^M$

$\square = f\left(\begin{bmatrix} \\ \\ \\ \end{bmatrix}\right)$

$$\frac{\partial f}{\partial x_1} = 0$$

$$\frac{\partial f}{\partial x_2} = 0$$

$$\vdots$$

$$\frac{\partial f}{\partial x_m} = 0$$

$$\nabla_{\mathbf{x}} f = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_m} \end{bmatrix}$$

Calculus with Linear Algebra

Functions with linear algebra

input dim ++

out dim ++

$$y = \mathrm{f}(x) \qquad y = \mathrm{f}\left(\begin{bmatrix} x_1 \\ \vdots \\ x_M \end{bmatrix}\right) \qquad y = \mathrm{f}\left(\begin{bmatrix} X_{1,1} & \cdots & X_{1,L} \\ \vdots & \ddots & \vdots \\ X_{M,1} & \cdots & X_{M,L} \end{bmatrix}\right)$$

$$\begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} = \mathrm{f}(x) \qquad \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} = \mathrm{f}\left(\begin{bmatrix} x_1 \\ \vdots \\ x_M \end{bmatrix}\right) \qquad \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} = \mathrm{f}\left(\begin{bmatrix} X_{1,1} & \cdots & X_{1,L} \\ \vdots & \ddots & \vdots \\ X_{M,1} & \cdots & X_{M,L} \end{bmatrix}\right)$$

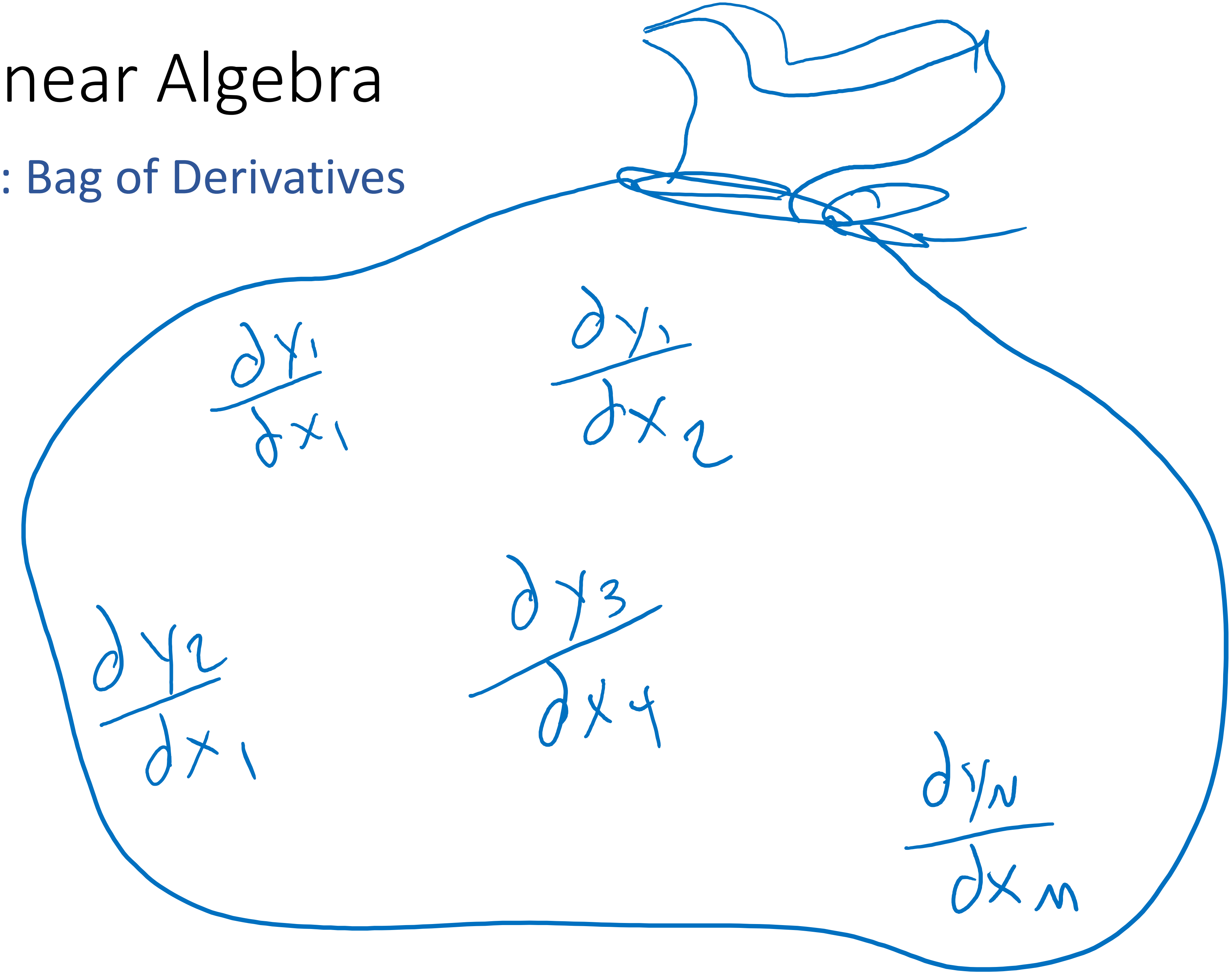
$$\begin{bmatrix} Y_{1,1} & \cdots & Y_{1,K} \\ \vdots & \ddots & \vdots \\ Y_{N,K} & \cdots & Y_{N,K} \end{bmatrix} = \mathrm{f}(x) \qquad \begin{bmatrix} Y_{1,1} & \cdots & Y_{1,K} \\ \vdots & \ddots & \vdots \\ Y_{N,K} & \cdots & Y_{N,K} \end{bmatrix} = \mathrm{f}\left(\begin{bmatrix} x_1 \\ \vdots \\ x_M \end{bmatrix}\right) \qquad \begin{bmatrix} Y_{1,1} & \cdots & Y_{1,K} \\ \vdots & \ddots & \vdots \\ Y_{N,K} & \cdots & Y_{N,K} \end{bmatrix} = \mathrm{f}\left(\begin{bmatrix} X_{1,1} & \cdots & X_{1,L} \\ \vdots & \ddots & \vdots \\ X_{M,1} & \cdots & X_{M,L} \end{bmatrix}\right)$$

Calculus with Linear Algebra

One way to think of it: Bag of Derivatives

$$\begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} = f\left(\begin{bmatrix} x_1 \\ \vdots \\ x_M \end{bmatrix}\right)$$

$\#: N \cdot M$

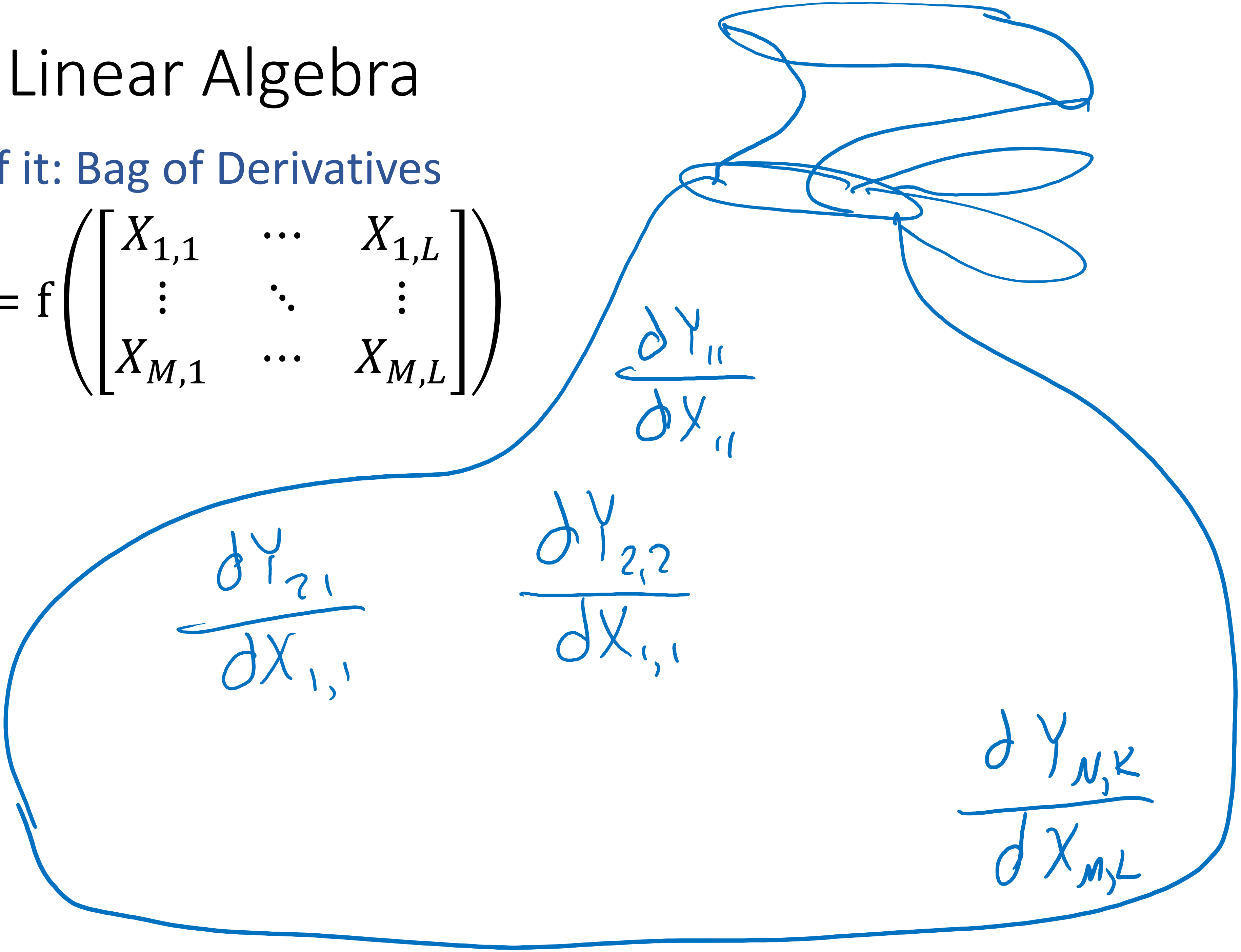


Calculus with Linear Algebra

One way to think of it: Bag of Derivatives

$$\begin{bmatrix} Y_{1,1} & \cdots & Y_{1,K} \\ \vdots & \ddots & \vdots \\ Y_{N,1} & \cdots & Y_{N,K} \end{bmatrix} = \mathrm{f}\left(\begin{bmatrix} X_{1,1} & \cdots & X_{1,L} \\ \vdots & \ddots & \vdots \\ X_{M,1} & \cdots & X_{M,L} \end{bmatrix}\right)$$

#: $N \cdot K \cdot M \cdot L$



Calculus with Linear Algebra

Jacobian: Vector in, vector out

Numerator-layout

$$y = f(x) \qquad y \in \mathbb{R}^N, \quad x \in \mathbb{R}^M, \quad \frac{\partial y}{\partial x} \in \mathbb{R}^{N\times M}$$

$$[\;] = f([\;])$$

$N > M$

$$\frac{\partial \vec{y}}{\partial \vec{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_M} \\ \frac{\partial y_2}{\partial x_1} & \ddots & & \\ \vdots & & & \\ \frac{\partial y_N}{\partial x_1} & & & \frac{\partial y_N}{\partial x_M} \end{bmatrix}$$

Calculus with Linear Algebra

Numerator-layout vs denominator-layout

Vector in, vector out

$\boldsymbol{y} = f(\boldsymbol{x}) \quad \boldsymbol{y} \in \mathbb{R}^N, \ \boldsymbol{x} \in \mathbb{R}^M$, assume $N \gg M$ for illustrative purposes

Numerator layout	Denominator layout
Number of outputs × number of inputs	Number of inputs × number of outputs
$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} \in \mathbb{R}^{N \times M}$	$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} \in \mathbb{R}^{M \times N}$

Calculus with Linear Algebra

Vector in, scalar out

Numerator-layout

$y = f(\boldsymbol{x}) \qquad y \in \mathbb{R}, \quad \boldsymbol{x} \in \mathbb{R}^{M}, \quad \frac{\partial y}{\partial \boldsymbol{x}} \in \mathbb{R}^{1 \times M}$

$$\frac{\partial y}{\partial \vec{x}} = \left[\frac{\partial y}{\partial x_1}, \quad \cdots \quad, \frac{\partial y}{\partial x_M} \right]$$

Calculus with Linear Algebra

Numerator-layout vs denominator-layout

Vector in, scalar out

$y = f(\boldsymbol{x}) \qquad y \in \mathbb{R}, \quad \boldsymbol{x} \in \mathbb{R}^M$

Numerator layout	Denominator layout
Number of outputs × number of inputs	Number of inputs × number of outputs
$\frac{\partial y}{\partial \boldsymbol{x}} \in \mathbb{R}^{1\times M}$	$\frac{\partial y}{\partial \boldsymbol{x}} \in \mathbb{R}^{M\times 1}$
$\begin{bmatrix} \quad \end{bmatrix}$	$\begin{bmatrix} \\ \\ \end{bmatrix}$

Calculus with Linear Algebra

Scalar in, vector out

Numerator-layout

$\boldsymbol{y} = f(x) \quad \boldsymbol{y} \in \mathbb{R}^N, \quad x \in \mathbb{R}, \quad \frac{\partial \boldsymbol{y}}{\partial x} \in \mathbb{R}^{N \times 1}$

$$\frac{\partial \vec{y}}{\partial x} = \begin{bmatrix} \frac{\partial y_1}{\partial x} \\ \vdots \\ \frac{\partial y_N}{\partial x} \end{bmatrix}$$

$[\;] = f(x)$

Calculus with Linear Algebra

Gradient: Vector in, scalar out

Transpose of numerator-layout

$y = f(\boldsymbol{x}) \qquad y \in \mathbb{R},\ \ \boldsymbol{x} \in \mathbb{R}^M,\ \ \frac{\partial y}{\partial \boldsymbol{x}} \in \mathbb{R}^{1\times M},\ \ \nabla_{\boldsymbol{x}} f \in \mathbb{R}^{M\times 1}$

$$\nabla_{\vec{x}} f = \left(\frac{\partial f}{\partial \vec{x}}\right)^T$$

$$\begin{bmatrix} \\ \\ \end{bmatrix} = \left(\begin{bmatrix} & & \end{bmatrix}\right)^T$$

Calculus with Linear Algebra

Matrix in, scalar out

Keep same dimensions as matrix

$y = f(\boldsymbol{X})$ $\quad y \in \mathbb{R},\ \boldsymbol{X} \in \mathbb{R}^{N\times M},\ \frac{\partial y}{\partial \boldsymbol{X}} \in \mathbb{R}^{N\times M}$

$$\frac{\partial y}{\partial X} = \begin{bmatrix} \frac{\partial y}{\partial X_{1,1}} & \cdots & \frac{\partial y}{\partial X_{1,M}} \\ \vdots & & \vdots \\ \frac{\partial y}{\partial X_{N,1}} & & \frac{\partial y}{\partial X_{N,M}} \end{bmatrix}$$

Exercise 1

Suppose we have a function that takes in a vector and squares each element individually, returning another vector, $\boldsymbol{y} = f(\boldsymbol{x})$.

$$f\left(\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}\right) \to \begin{bmatrix} x_1^2 \\ x_2^2 \\ x_3^2 \end{bmatrix} \quad y_1$$

Example: $f\left(\begin{bmatrix} 7 \\ 3 \\ 5 \end{bmatrix}\right) \to \begin{bmatrix} 49 \\ 9 \\ 25 \end{bmatrix}$

What is $\partial \boldsymbol{y} / \partial \boldsymbol{x}$? (use numerator layout)

$$\frac{\partial y}{\partial x} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \frac{\partial y_1}{\partial x_3} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \frac{\partial y_2}{\partial x_3} \\ \frac{\partial y_3}{\partial x_1} & \frac{\partial y_3}{\partial x_2} & \frac{\partial y_3}{\partial x_3} \end{bmatrix} = \begin{bmatrix} 2x_1 & 0 & 0 \\ 0 & 2x_2 & 0 \\ 0 & 0 & 2x_3 \end{bmatrix}$$

Calculus

Chain rule

Reminder: Calculus Chain Rule (scalar version)

$$x \rightarrow \boxed{g} \xrightarrow{z} \boxed{f} \rightarrow y$$

$$y = f(z)$$
$$z = g(x)$$

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

$$\frac{dy}{dx} = \frac{dy}{dz}\frac{dz}{dx}$$

$$\frac{df}{dx} = \frac{df}{dg}\frac{dg}{dx}$$

Network Optimization: Layer Implementation

$$J(\mathbf{w}) = z_3$$

$$z_3 = f_3(w_3, z_2)$$

$$z_2 = f_2(w_2, z_1)$$

$$z_1 = f_1(w_1, x)$$

$$\frac{\partial J}{\partial w_3} = \frac{\partial J}{\partial z_3}\frac{\partial z_3}{\partial w_3}$$

$$\boxed{\frac{\partial J}{\partial w_2}} = \boxed{\frac{\partial J}{\partial z_3}\frac{\partial z_3}{\partial z_2}}\boxed{\frac{\partial z_2}{\partial w_2}}$$

$$\frac{\partial J}{\partial w_1} = \frac{\partial J}{\partial z_3}\frac{\partial z_3}{\partial z_2}\frac{\partial z_2}{\partial z_1}\frac{\partial z_1}{\partial w_1}$$

Lots of repeated calculations

$f_2(w, x) = w^5 x^7$

$y = w^5 x^7$

$\frac{dy}{dw} = 5w^4x^7$

$\frac{dy}{dx} = w^5 7x^6$

x

y

w

$\boxed{\frac{dJ}{dx}}$

$\frac{dJ}{dy}$

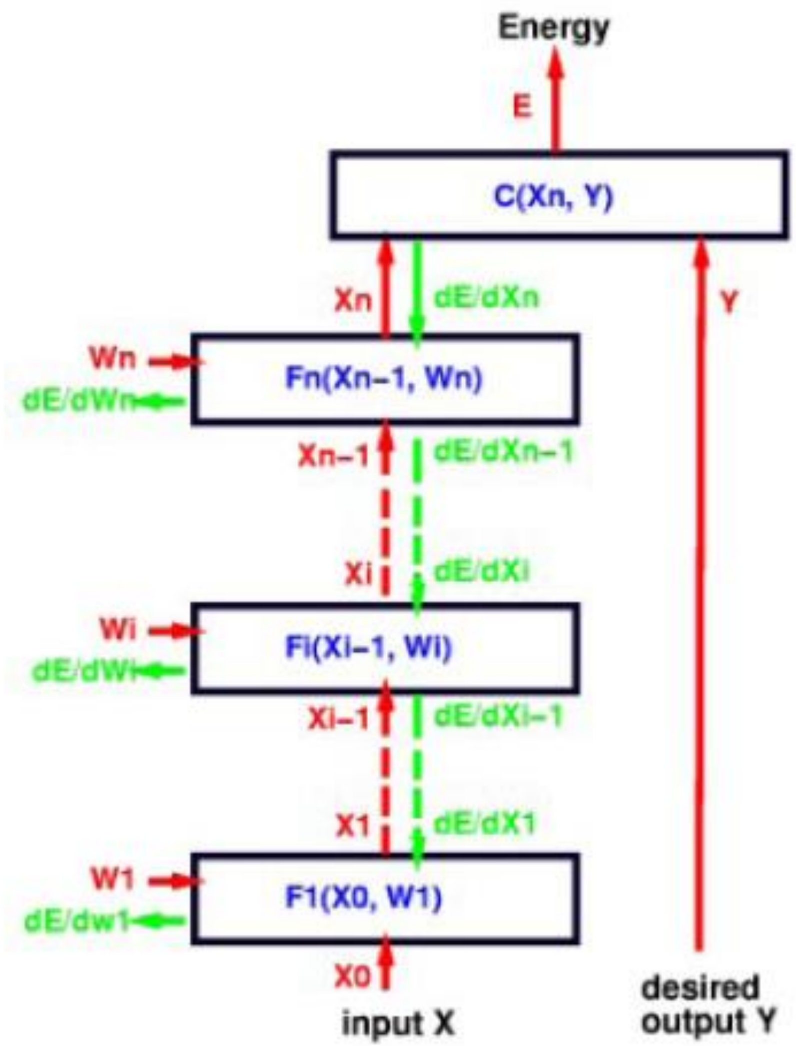
$\frac{dJ}{dx} = \frac{dJ}{dy}\frac{dy}{dx}$

$\frac{dJ}{dw} = \frac{dJ}{dy}\frac{dy}{dw}$

Example slide from 10-601

Backpropagation (continue)

Example slide from 10-701



- $\frac{\partial E}{\partial X_n} = \frac{\partial C(X_n,Y)}{\partial X_n}$
- $\frac{\partial E}{\partial X_{n-1}} = \frac{\partial E}{\partial X_n} \boxed{\frac{\partial F_n(X_{n-1},W_n)}{\partial X_{n-1}}}$
- $\boxed{\frac{\partial E}{\partial W_n} = \boxed{\frac{\partial E}{\partial X_n}} \boxed{\frac{\partial F_n(X_{n-1},W_n)}{\partial W_n}}}$
- $\frac{\partial E}{\partial X_{n-2}} = \frac{\partial E}{\partial X_{n-1}} \frac{\partial F_{n-1}(X_{n-2},W_{n-1})}{\partial X_{n-2}}$
- $\boxed{\frac{\partial E}{\partial W_{n-1}} = \boxed{\frac{\partial E}{\partial X_{n-1}}} \frac{\partial F_{n-1}(X_{n-2},W_{n-1})}{\partial W_{n-1}}}$
- ....etc, until we reach the first module.
- we now have all the $\frac{\partial E}{\partial W_i}$ for $i \in [1, n]$.

$\frac{\partial \sigma}{\partial X_{n-1}}$ $\sigma($ ... W_n

Say: $F_n = \sigma$

$\frac{\partial \sigma}{\partial W_n} \sigma(1-\sigma) x_{n-1}$

What is a Derivative?

$f(x) = 11x + 9$

$y = f(x)$

$$x \longrightarrow \boxed{f} \longrightarrow y$$

$$x + \epsilon \longrightarrow \boxed{f} \longrightarrow f(x + \epsilon)$$

$$\frac{df}{dx}$$

$$x = 7 \longrightarrow \boxed{11x + 9} \longrightarrow 86$$

$$7.1 \longrightarrow \boxed{11x + 9} \longrightarrow 87.1$$

$$\frac{\Delta y}{\Delta x} = \frac{1.1}{0.1} = 11$$

What is a Derivative?

$g(x) = 3x$

$f(z) = 2z$

$y = f(\underline{g(x)})$

$$x \longrightarrow \overset{g}{\boxed{3x}} \xrightarrow{z} \overset{f}{\boxed{2z}} \longrightarrow y$$

$$\frac{.6}{.1} = 6$$

$$\frac{\partial y}{\partial x} = \frac{\partial f}{\partial z} \frac{\partial g}{\partial x}$$

What is a Derivative?

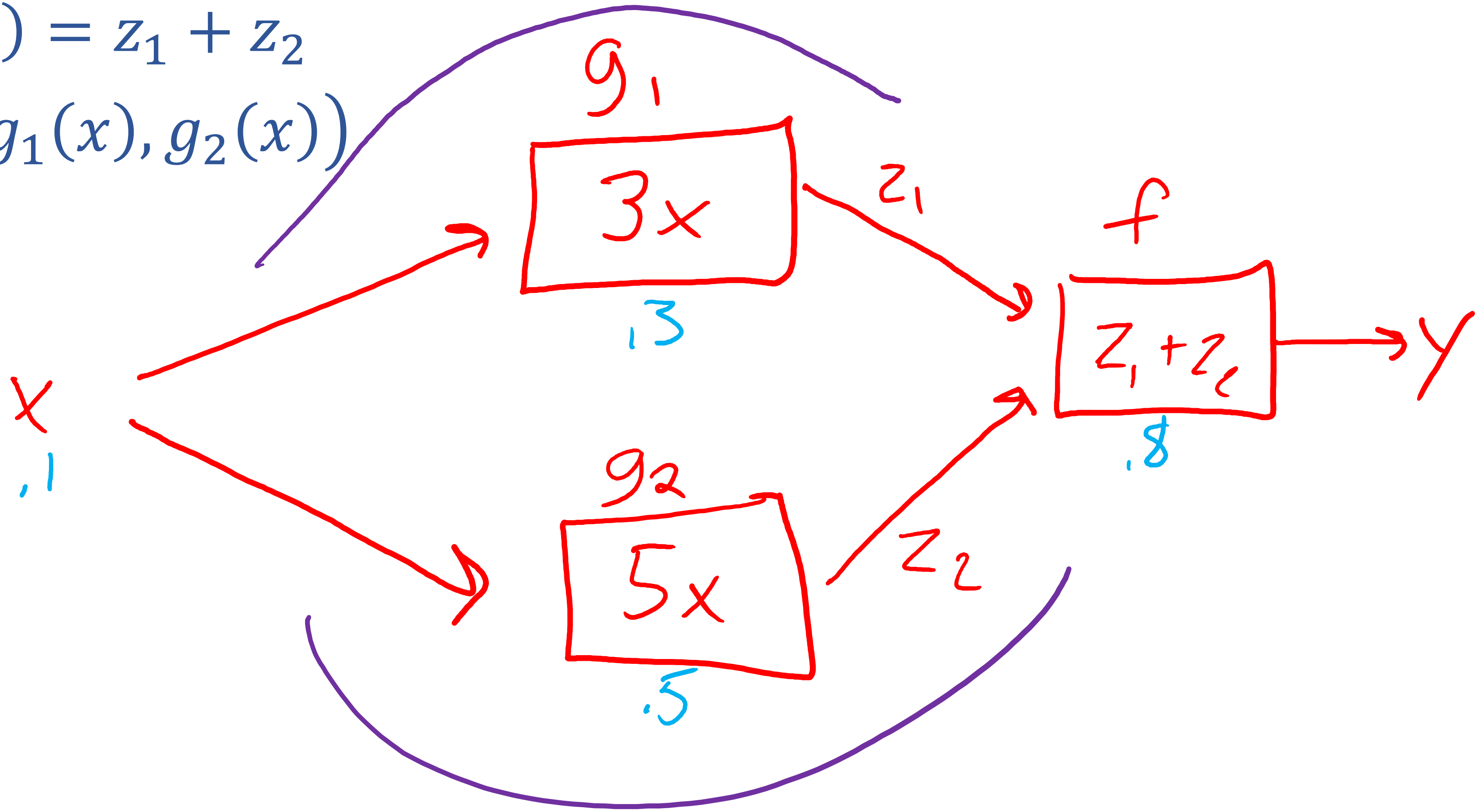
$$\frac{dy}{dx} = \frac{df}{dg_1}\frac{dg_1}{dx} + \frac{df}{dg_2}\frac{dg_2}{dx}$$

$g_1(x) = 3x$

$g_2(x) = 5x$

$f(z_1, z_2) = z_1 + z_2$

$y = f\big(g_1(x), g_2(x)\big)$



Multivariate Chain Rule

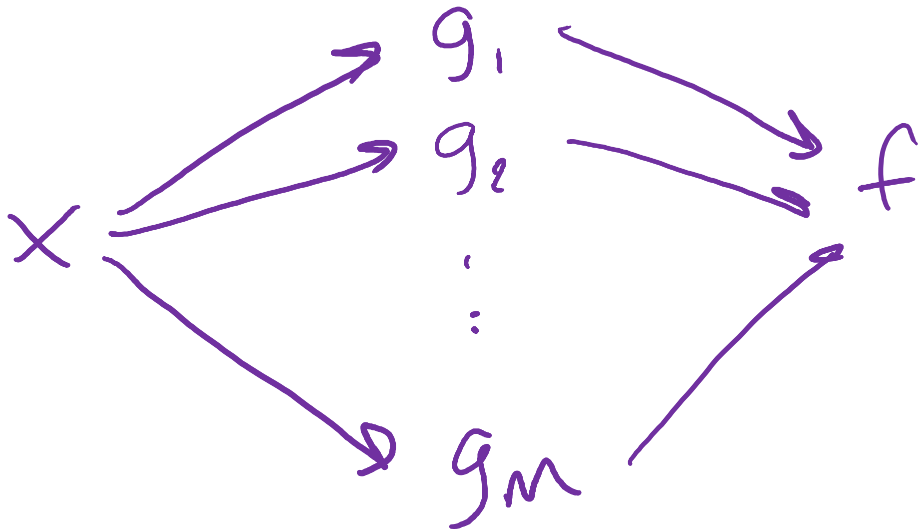
$$z_1 = g_1(x)$$

$$\vdots$$

$$z_M = g_M(x)$$

$$y = f(z_1, \cdots, z_M)$$

$$\boxed{\frac{\partial y}{\partial x} = \sum_{j=1}^{M} \frac{\partial y}{\partial z_j} \frac{\partial z_j}{\partial x}}$$



Calculus Chain Rule

Scalar:

$z = g(x)$

$y = f(z)$

$$\frac{dy}{dx} = \frac{dy}{dz}\frac{dz}{dx}$$

Multivariate:

$\mathbf{z} = g(x)$

$y = f(\mathbf{z})$

$$\frac{dy}{dx} = \sum_j \frac{\partial y}{\partial z_j}\frac{\partial z_j}{\partial x}$$

Exercise 2

$$z_1 = g_1(x) = \sin(x)$$
$$z_2 = g_2(x) = x^3$$
$$y = f(z_1, z_2) = z_1^4 e^{z_2} + 5z_1 + 7z_2$$

Multivariate chain rule

$$\mathbf{z} = g(x)$$
$$y = f(\mathbf{z})$$
$$\frac{dy}{dx} = \sum_j \frac{\partial y}{\partial z_j} \frac{\partial z_j}{\partial x}$$

$$\frac{\partial y}{\partial x} = \frac{\partial y}{\partial z_1}\frac{\partial z_1}{\partial x} + \frac{\partial y}{\partial z_2}\frac{\partial z_2}{\partial x}$$

$$= \left(4z_1^3 e^{z_2} + 5\right)\cdot\cos(x) + \left(z_1^4 e^{z_2} + 7\right)3x^2$$

Exercise 3

$$z_1 = g_1(x) = \sin(x)$$
$$z_2 = g_2(x) = x^3$$
$$y = f(z_1, z_2) = z_1 z_2$$

$$\frac{\partial y}{\partial x} =$$

<u>Multivariate chain rule</u>

$$\boldsymbol{z} = g(x)$$
$$y = f(\boldsymbol{z})$$
$$\frac{dy}{dx} = \sum_j \frac{\partial y}{\partial z_j} \frac{\partial z_j}{\partial x}$$

Calculus Chain Rule

Scalar:

$z = g(x)$

$y = f(z)$

$$\frac{dy}{dx} = \frac{dy}{dz}\frac{dz}{dx}$$

Multivariate:

$\mathbf{z} = g(x)$

$y = f(\mathbf{z})$

$$\frac{dy}{dx} = \sum_j \frac{\partial y}{\partial z_j}\frac{\partial z_j}{\partial x}$$

Multivariate:

$\mathbf{z} = g(\mathbf{x})$

$\mathbf{y} = f(\mathbf{z})$

$$\frac{dy_{\mathrm{i}}}{dx_{\mathrm{k}}} = \sum_j \frac{\partial y_i}{\partial z_j}\frac{\partial z_j}{\partial x_{\mathrm{k}}}$$

Linear Regression

Last time

Calculus needed to:

- Find $\mathbf{w}$ that minimizes MSE with model $y = \mathbf{w}^T\mathbf{x}$, $\mathbf{x} \in \mathbb{R}^2$

Today

Quick trick to:

- Find $\mathbf{w}, b$ that minimizes MSE with model $y = \mathbf{w}^T\mathbf{x} + b$, $\mathbf{x} \in \mathbb{R}^M$

Calculus needed for neural networks

- Partial derivatives of various sizes
- Multivariate chain rule

Constrained optimization and Lagrange multipliers

Constrained Optimization

Method of Lagrange multipliers

Exercise

Mini-max game

$$L(x, \lambda) = 2x + 9 - \lambda(x^2 - 2)$$

Two teams

Team λ:

- Goes first
- Chooses a value for λ in attempt to *maximize* $L(x, \lambda)$

Team x:

- Goes second
- Chooses a value for x in attempt to *minimize* $L(x, \lambda)$