# 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 $y = \mathbf{w}^T\mathbf{x}$, $\quad\mathbf{x} \in \mathbb{R}^2$

## Today

Quick trick to:

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

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

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

$\square = f(\,\boxed{\vdots}\,)$

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

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

$$\vdots \qquad \vdots$$

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

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

# Calculus with Linear Algebra

## Functions with linear algebra

$$y = \mathrm{f}(x)$$

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

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

$$\begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} = \mathrm{f}(x)$$

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

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

$$\begin{bmatrix} Y_{1,1} & \cdots & Y_{1,K} \\ \vdots & \ddots & \vdots \\ Y_{N,K} & \cdots & Y_{N,K} \end{bmatrix} = \mathrm{f}(x)$$

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

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

# Calculus with Linear Algebra

## One way to think of it: Bag of Derivatives

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

# Calculus with Linear Algebra

## One way to think of it: Bag of Derivatives

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

# Calculus with Linear Algebra

Jacobian: Vector in, vector out

Numerator-layout

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

# Calculus with Linear Algebra

## Numerator-layout vs denominator-layout

## Vector in, vector out

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

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

# 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 |
| $\frac{\partial y}{\partial \boldsymbol{x}} \in \mathbb{R}^{1 \times M}$ | $\frac{\partial y}{\partial \boldsymbol{x}} \in \mathbb{R}^{M \times 1}$ |
| | |

# Calculus with Linear Algebra

Scalar in, vector out

Numerator-layout

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

# Calculus with Linear Algebra

Gradient: Vector in, scalar out

Transpose of numerator-layout

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

# Calculus with Linear Algebra

Matrix in, scalar out

Keep same dimensions as matrix

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

# Exercise 1

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

$$f\left(\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}\right) \rightarrow \begin{bmatrix} x_1^2 \\ x_2^2 \\ x_3^2 \end{bmatrix} \qquad \text{Example: } f\left(\begin{bmatrix} 7 \\ 3 \\ 5 \end{bmatrix}\right) \rightarrow \begin{bmatrix} 49 \\ 9 \\ 25 \end{bmatrix}$$

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

# Calculus

## Chain rule

# Reminder: Calculus Chain Rule (scalar version)

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

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

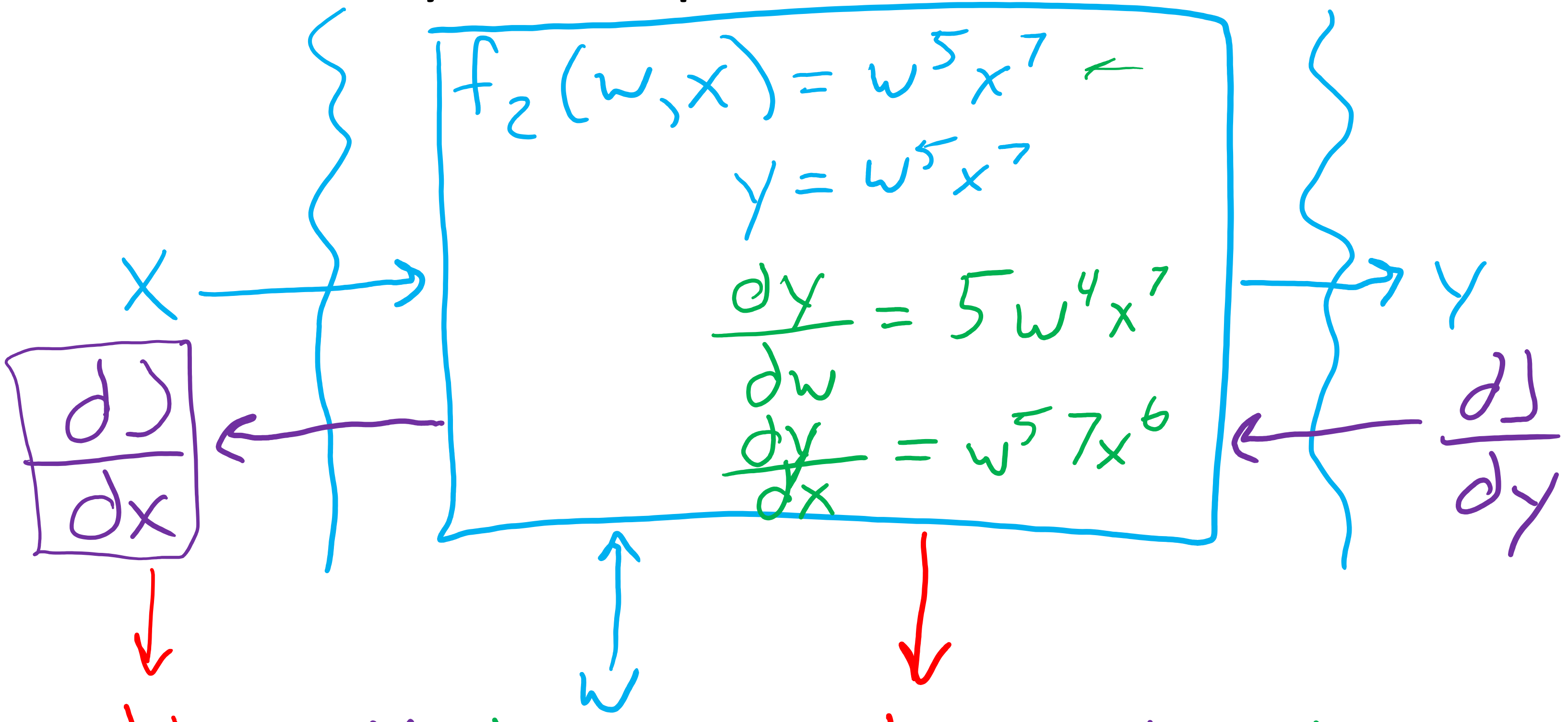
# Network Optimization: Layer Implementation

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

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

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

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




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

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

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

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

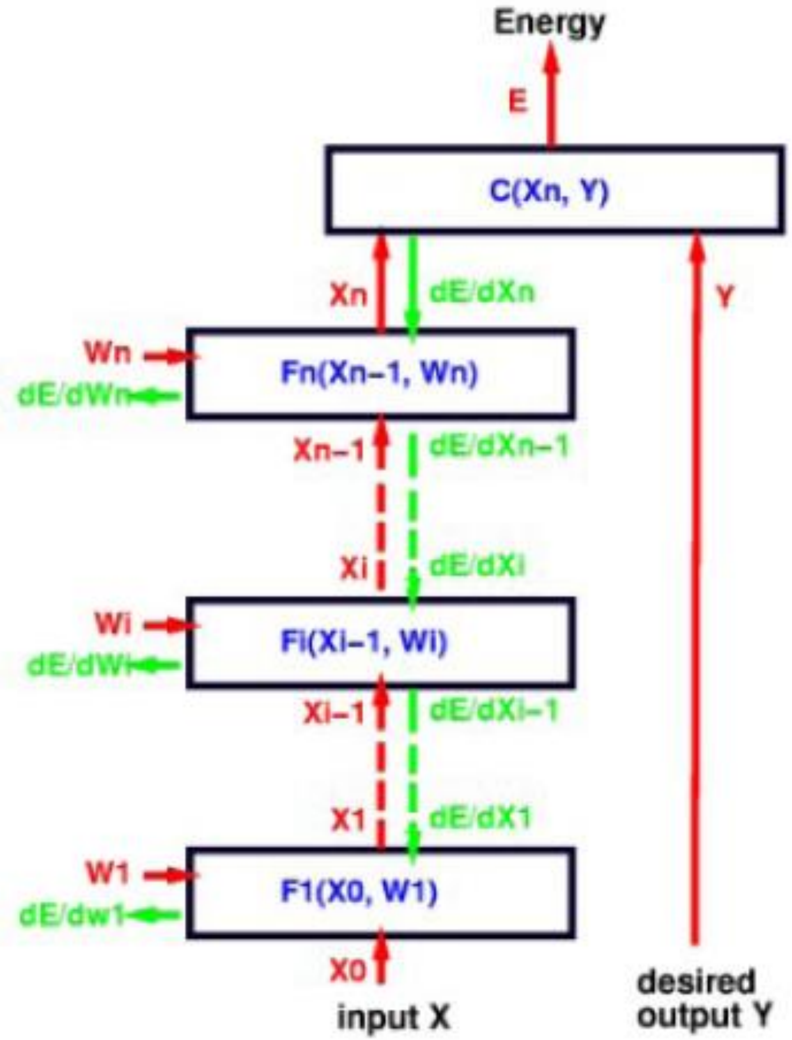
$$\frac{dJ}{dw} = \frac{dJ}{dy}\frac{dy}{dw}$$

Lots of repeated calculations

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?

$f(x) = 11x + 9$

$y = f(x)$

# What is a Derivative?

$g(x) = 3x$

$f(z) = 2z$

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

# What is a Derivative?

$$g_1(x) = 3x$$

$$g_2(x) = 5x$$

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

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

# Multivariate Chain Rule

$$z_1 = g_1(x)$$

$$\vdots$$

$$z_M = g_M(x)$$

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

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

# Calculus Chain Rule

## Scalar:

$$z = g(x)$$

$$y = f(z)$$

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

## Multivariate:

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

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

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

# Exercise 2

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

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

Multivariate chain rule

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

# 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 $\lambda$:

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

Team $x$:

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