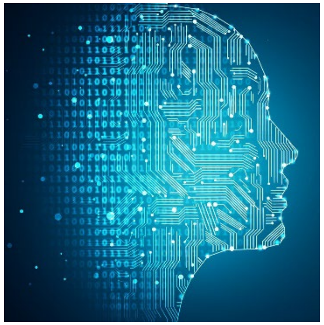
Principles of Machine Learning

Vector and Matrix Calculus

Kevin Moon (kevin.moon@usu.edu)

STAT/CS 5810/6655



Motivation

- Many machine learning algorithms require optimization
- Optimization typically requires calculus
- Thus we need vector and matrix calculus to understand how learning occurs

- The good news: we don’t need integrals!
- The bad news: matrix derivatives can get complicated enough so hold on…

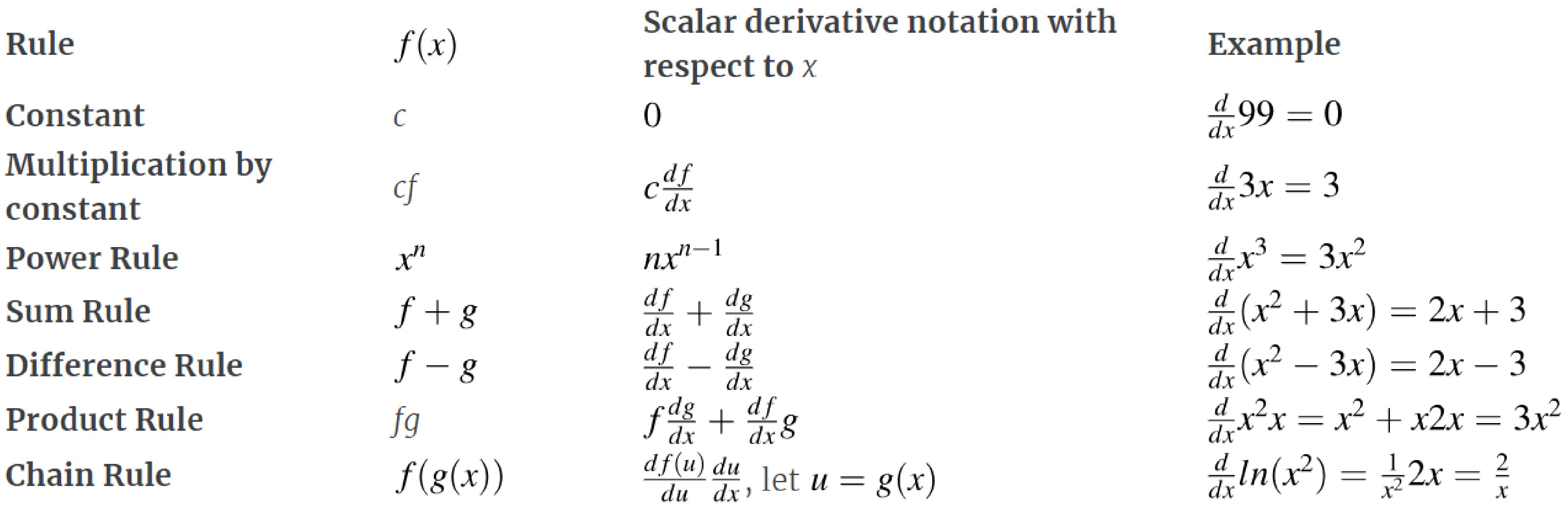
Outline

1. Univariate/scalar derivatives
2. Vector calculus and partial derivatives
3. Matrix calculus
4. Chain rules

What is a derivative?

- In words, a derivative measures how much a change in the input of a function changes the function output
- Mathematically, let $f: \mathbb{R} \rightarrow \mathbb{R}$
  - Above notation means f is a function that takes the real numbers ($\mathbb{R}$) as input and maps to the real numbers (the output)
- The amount a change in the input of f changes the output of f may not be constant across the entire input space
  - Thus the derivative of f is also a function of the input space
    - I.e., the derivative may be different for different inputs
  - Using the same above notation: $\frac{df}{dx}: \mathbb{R} \rightarrow \mathbb{R}$

Scalar derivative rules



Rule	$f(x)$	Scalar derivative notation with respect to x	Example
Constant	c	0	$\frac{d}{dx}99 = 0$
Multiplication by constant	cf	$c\frac{df}{dx}$	$\frac{d}{dx}3x = 3$
Power Rule	x^n	nx^{n-1}	$\frac{d}{dx}x^3 = 3x^2$
Sum Rule	$f+g$	$\frac{df}{dx} + \frac{dg}{dx}$	$\frac{d}{dx}(x^2+3x) = 2x+3$
Difference Rule	$f-g$	$\frac{df}{dx} - \frac{dg}{dx}$	$\frac{d}{dx}(x^2-3x) = 2x-3$
Product Rule	fg	$f\frac{dg}{dx} + \frac{df}{dx}g$	$\frac{d}{dx}x^2x = x^2 + x2x = 3x^2$
Chain Rule	$f(g(x))$	$\frac{df(u)}{du}\frac{du}{dx}$, let $u = g(x)$	$\frac{d}{dx}ln(x^2) = \frac{1}{x^2}2x = \frac{2}{x}$

- Can view $\frac{d}{dx}$ as an operator that maps a function of one parameter/variable to another function of that parameter

Vector calculus

- Machine learning functions can be multiple functions of multiple parameters
- Let's consider functions of two parameters
- **Example**: $f(x, y) = xy$
  - Notation: $f: \mathbb{R}^2 \to \mathbb{R}$
  - How does f change when we change x or y?
  - It depends on which we're changing
- We compute derivatives with respect to (wrt) one variable (parameter) at a time
  - I.e., we treat all other variables as constants
  - This gives us two **partial derivatives**
  - Written as $\frac{\partial}{\partial x}$ and $\frac{\partial}{\partial y}$

Example

Consider the multivariate function $f(x, y) = 3x^2y$

- When computing $\frac{\partial}{\partial x}$ we consider 3 and y to be constants
  - Thus $\frac{\partial f(x,y)}{\partial x} = 6xy$
- When computing $\frac{\partial}{\partial y}$ we consider 3 and x to be constants
  - Thus $\frac{\partial f(x,y)}{\partial y} = 3x^2$

The gradient

- Let’s combine the partial derivatives of f into a vector
  - This vector is called the **gradient** of f
  - The gradient of f is written as ∇f
- From the previous example, we have

$$\nabla f(x,y) = \begin{bmatrix} \dfrac{\partial f(x,y)}{\partial x} \\ \dfrac{\partial f(x,y)}{\partial y} \end{bmatrix} = \begin{bmatrix} 6yx \\ 3x^2 \end{bmatrix}$$

The Hessian

- We can also take the second partial derivatives of f
  - There are d^2 possible combinations: $\frac{\partial^2 f}{\partial x_1^2}, \frac{\partial^2 f}{\partial x_1 \partial x_2}, \ldots$
- The $d \times d$ matrix that contains all second partial derivatives is called the **Hessian***:*

$$\nabla^2 f = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_d} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & & \\ \vdots & & \ddots & \\ \frac{\partial^2 f}{\partial x_d \partial x_1} & \frac{\partial^2 f}{\partial x_d \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_d^2} \end{bmatrix}$$

The gradient and the Hessian

- In univariate calculus, you learned how to find critical points of a function by setting the derivative equal to zero and solving
  - In multivariate calculus, we can do the same thing by setting the gradient equal to the zero vector
- In univariate calculus, the second derivative tells you whether the critical point is a local minimum, local maximum, or a saddle point
  - The Hessian plays a similar role in multivariate calculus

Derivatives of multiple functions

- We sometimes need to take derivatives of multiple functions
- Taking derivatives of multiple functions moves us from vector calculus to matrix calculus
- Let's add the function $g(x, y) = 2x + y^8$
- The gradient of g is

$$\nabla g(x, y) = [2 \quad 8y^7]^T$$

- Gradients organize the partial derivatives for a specific scalar function
- With two functions, we organize the gradients into a matrix

The Jacobian matrix

- Consider the two functions from before: $f(x, y) = 3x^2y$ and $g(x, y) = 2x + y^8$
- The Jacobian consists of the gradients stacked in the columns
- Useful for organizing the derivatives
- Our example:

$$J = [\nabla f(x, y) \quad \nabla g(x, y)] = \begin{bmatrix} 6yx & 2 \\ 3x^2 & 8y^7 \end{bmatrix}$$

Generalization of the Jacobian

- Let’s vectorize the inputs:

$$\boldsymbol{x} = [x_1 \quad \dots \quad x_d]^T$$

- Let’s vectorize the functions to get $\boldsymbol{f}: \mathbb{R}^d \to \mathbb{R}^m$

$$\boldsymbol{f}(\boldsymbol{x}) = [f_1(\boldsymbol{x}) \quad \dots \quad f_m(\boldsymbol{x})]^T$$

- Previous example:

$$\boldsymbol{f}(\boldsymbol{x}) = \begin{bmatrix} 3x_1^2 x_2 \\ 2x_1 + x_2^8 \end{bmatrix}$$

- Often, we have $m = d$ because we have a scalar function for each component of the input vector: $\boldsymbol{f}: \mathbb{R}^d \to \mathbb{R}^d$

Generalization of the Jacobian

- The Jacobian matrix is the collection of all $d \times m$ partial derivatives
  - Equivalent to the collection of all gradients

$$J = [\nabla f_1(\boldsymbol{x}) \quad \nabla f_2(\boldsymbol{x}) \quad \dots \quad \nabla f_m(\boldsymbol{x})]$$

$$= \begin{bmatrix} \frac{\partial f_1(\boldsymbol{x})}{\partial x_1} & \frac{\partial f_2(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial f_m(\boldsymbol{x})}{\partial x_1} \\ \frac{\partial f_1(\boldsymbol{x})}{\partial x_2} & \frac{\partial f_2(\boldsymbol{x})}{\partial x_2} & \cdots & \frac{\partial f_m(\boldsymbol{x})}{\partial x_2} \\ \vdots & & \ddots & \vdots \\ \frac{\partial f_1(\boldsymbol{x})}{\partial x_d} & \frac{\partial f_2(\boldsymbol{x})}{\partial x_d} & \cdots & \frac{\partial f_m(\boldsymbol{x})}{\partial x_d} \end{bmatrix}$$

The Hessian and the Jacobian

- Consider a function $f\colon \mathbb{R}^d \to \mathbb{R}$
- Taking the gradient of f results in a vector of functions
  - I.e., ∇f is a d dimensional vector of functions
- The Hessian can be obtained by finding the Jacobian of ∇f

- The remaining slides go through some specific rules for calculating multidimensional derivatives
- These rules likely won’t be necessary for this course but are included for those that are interested
- Thus we will stop here

Element-wise binary operators

Many common vector operations, such as vector addition and multiplication of a vector by a scalar, can be represented as element-wise operations

- Element-wise means we apply the operator to the first element of the vector to get the first output, apply the operator to the second element of the vector to get the second output, etc.
- Deep learning examples: $\max(\boldsymbol{w}, \boldsymbol{x})$ and $\boldsymbol{w} > \boldsymbol{x}$

Element-wise binary operators

- General element-wise binary operator: $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x}) \odot \boldsymbol{g}(\boldsymbol{w})$
    - $\boldsymbol{y}: \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$
- We can break the Jacobian into two parts:

$$J_{\boldsymbol{x}} = \left[\nabla_{\boldsymbol{x}}\left(f_1(\boldsymbol{x}) \odot g_1(\boldsymbol{w})\right) \quad \dots \quad \nabla_{\boldsymbol{x}}\left(f_d(\boldsymbol{x}) \odot g_d(\boldsymbol{w})\right)\right]$$

$$J_{\boldsymbol{w}} = \left[\nabla_{\boldsymbol{w}}\left(f_1(\boldsymbol{x}) \odot g_1(\boldsymbol{w})\right) \quad \dots \quad \nabla_{\boldsymbol{w}}\left(f_d(\boldsymbol{x}) \odot g_d(\boldsymbol{w})\right)\right]$$

- $\nabla_{\boldsymbol{x}}$ is the gradient wrt $\boldsymbol{x}$, etc.

Element-wise binary operators

- This looks potentially nasty…
- Luckily, we can usually simplify this
- In many cases, the functions f_i and g_i are only functions of x_i and w_i, respectively
  - I.e., f_i is not a function of x_j for $j \neq i$
  - Example: $f_i =$ the identity function
- The Jacobians simplify to diagonal matrices in this case

Element-wise binary operators

- The Jacobians simplify to diagonal matrices in this case
- Example:

$$J_w = \begin{bmatrix} \frac{\partial}{\partial w_1}\left(f_1(x_1) \circledcirc g_1(w_1)\right) & 0 & \dots & 0 \\ 0 & \frac{\partial}{\partial w_2}\left(f_2(x_2) \circledcirc g_2(w_2)\right) & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & \frac{\partial}{\partial w_d}\left(f_d(x_d) \circledcirc g_d(w_d)\right) \end{bmatrix}$$

$$= diag\left(\frac{\partial}{\partial w_1}\left(f_1(x_1) \circledcirc g_1(w_1)\right), \frac{\partial}{\partial w_2}\left(f_2(x_2) \circledcirc g_2(w_2)\right), \dots, \frac{\partial}{\partial w_d}\left(f_d(x_d) \circledcirc g_d(w_d)\right)\right)$$

Element-wise binary operators

Example: vector addition $\boldsymbol{x} + \boldsymbol{w}$

$$J_{\boldsymbol{w}} = J_{\boldsymbol{x}} = I$$

- I is the $d \times d$ identity matrix

Jacobians of common operators

Op	Partial with respect to w
$+$	$\frac{\partial(\mathbf{w}+\mathbf{x})}{\partial\mathbf{w}} = diag(\ldots\frac{\partial(w_i+x_i)}{\partial w_i}\ldots) = diag(\vec{1}) = I$
$-$	$\frac{\partial(\mathbf{w}-\mathbf{x})}{\partial\mathbf{w}} = diag(\ldots\frac{\partial(w_i-x_i)}{\partial w_i}\ldots) = diag(\vec{1}) = I$
$\otimes$	$\frac{\partial(\mathbf{w}\otimes\mathbf{x})}{\partial\mathbf{w}} = diag(\ldots\frac{\partial(w_i\times x_i)}{\partial w_i}\ldots) = diag(\mathbf{x})$
$\oslash$	$\frac{\partial(\mathbf{w}\oslash\mathbf{x})}{\partial\mathbf{w}} = diag(\ldots\frac{\partial(w_i/x_i)}{\partial w_i}\ldots) = diag(\ldots\frac{1}{x_i}\ldots)$

Op	Partial with respect to x
$+$	$\frac{\partial(\mathbf{w}+\mathbf{x})}{\partial\mathbf{x}} = I$
$-$	$\frac{\partial(\mathbf{w}-\mathbf{x})}{\partial\mathbf{x}} = diag(\ldots\frac{\partial(w_i-x_i)}{\partial x_i}\ldots) = diag(-\vec{1}) = -I$
$\otimes$	$\frac{\partial(\mathbf{w}\otimes\mathbf{x})}{\partial\mathbf{x}} = diag(\mathbf{w})$
$\oslash$	$\frac{\partial(\mathbf{w}\oslash\mathbf{x})}{\partial\mathbf{x}} = diag(\ldots\frac{-w_i}{x_i^2}\ldots)$

- $\otimes$ and $\oslash$ are element-wise multiplication and division, respectively
- $\otimes$ is sometimes called the Hadamard product

Scalar expansion derivatives

- Example: adding scalar z to vector $\boldsymbol{x}$
  - Can write $y = \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{g}(z)$ where $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{x}$ and $\boldsymbol{g}(z) = \mathbf{1}z$
  - We can derive the Jacobians using the same rules given previously: $J_{\boldsymbol{x}} = I$ and $J_z = \mathbf{1}^T$
- Example: multiplying the vector $\boldsymbol{x}$ by the scalar z
  - Using similar ideas, we get $J_{\boldsymbol{x}} = Iz$ and $J_z = \boldsymbol{x}^T$

Vector sum reduction

- In deep learning, we often sum up the elements of vectors
  - E.g. the dot product
- Luckily, the gradient is a linear operator and so it can be taken inside the sum to get a sum of gradients:

$$\nabla\left(\sum_{i=1}^{d} f_i(\boldsymbol{x})\right) = \sum_{i=1}^{d} \nabla f_i(\boldsymbol{x})$$

Group Exercise

1. Compute the gradient wrt $\boldsymbol{x}$ of $\sum_{i=1}^{d} x_i$
2. Compute the gradient wrt $\boldsymbol{x}$ of $\sum_{i=1}^{d} z x_i$
3. Compute the gradient wrt z of $\sum_{i=1}^{d} z x_i$

1. $\mathbf{1}$
2. $z\mathbf{1}$
3. $\sum_{i=1}^{d} x_i$

Chain rules

- We can’t use our current rules to find the derivatives of complex functions
  - E.g. $sum(\boldsymbol{w} + \boldsymbol{x})$
- We need chain rules to do this

Univariate chain rule

- Let f and g be univariate functions
- What is the derivative of $f\big(g(x)\big)$?
- Define intermediate variables: $y = f\big(g(x)\big)$ & $u = g(x)$
- The chain rule is

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

- Example: $y = \sin x^2$
  - The derivative is $2x \cos x^2$

Univariate chain rule

- The chain rule can be extended to an arbitrary number of nested functions
- For example, let f_1, f_2, f_3, and f_4 all be univariate functions and let $y = f_4\left(f_3\left(f_2\left(f_1(x)\right)\right)\right)$
  - Introduce intermediate variables: $u_1 = f_1(x), u_2 = f_2(u_1)$, $u_3 = f_3(u_2), u_4 = f_4(u_3) = y$
- Apply the chain rule multiple times:

$$\frac{dy}{dx} = \frac{dy}{du_3}\frac{du_3}{du_2}\frac{du_2}{du_1}\frac{du_1}{dx}$$

Univariate total derivative chain rule

- The univariate chain rule has a limitation: all intermediate variables must be functions of single variables
- Consider the function $y = x + x^2$
  - We can use the addition rule to take the derivative of this function
  - Can we use the chain rule?
- Set $u_1(x) = x^2$. This gives $\mathrm{y} = \mathrm{u}_2(x, u_1) = x + u_1$
- When differentiating wrt x we need to consider how u_1 depends on x

Univariate total derivative chain rule

- To compute $\frac{dy}{dx}$, we need to sum up all possible contributions from changes in x to changes in y
- The total derivative wrt x assumes all variables are functions of x and potentially vary as x varies
- In the previous example:

$$\frac{dy}{dx} = \frac{\partial u_2(x, u_1)}{\partial x} = \frac{\partial u_2}{\partial x} + \frac{\partial u_2}{\partial u_1}\frac{\partial u_1}{\partial x} = 1 + 2x$$

Univariate total derivative chain rule

- This generalizes as follows:

$$\frac{\partial f(x, u_1, \dots, u_n)}{\partial x} = \frac{\partial f}{\partial x} + \frac{\partial f}{\partial u_1}\frac{\partial u_1}{\partial x} + \cdots + \frac{\partial f}{\partial u_n}\frac{\partial u_n}{\partial x}$$

$$= \frac{\partial f}{\partial x} + \sum_{i=1}^{n} \frac{\partial f}{\partial u_i}\frac{\partial u_i}{\partial x}$$

- The total derivative assumes all variables are potentially codependent while the partial derivative assumes all variables but x are constant

Vector chain rule

- Let $\boldsymbol{g}: \mathbb{R}^d \to \mathbb{R}^k$ and $\boldsymbol{f}: \mathbb{R}^k \to \mathbb{R}^m$
- What is the Jacobian of $\boldsymbol{f}\big(\boldsymbol{g}(\boldsymbol{x})\big)$?
- The vector chain rule:

$$\frac{\partial \boldsymbol{f}\big(\boldsymbol{g}(\boldsymbol{x})\big)}{\partial \boldsymbol{x}} = \begin{bmatrix} \frac{\partial g_1}{\partial x_1} & \cdots & \frac{\partial g_k}{\partial x_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_1}{\partial x_d} & \cdots & \frac{\partial g_k}{\partial x_d} \end{bmatrix} \begin{bmatrix} \frac{\partial f_1}{\partial g_1} & \cdots & \frac{\partial f_m}{\partial g_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_1}{\partial g_k} & \cdots & \frac{\partial f_m}{\partial g_k} \end{bmatrix}$$

- This is equivalent to multiplying the Jacobian of $\boldsymbol{g}$ by the Jacobian of $\boldsymbol{f}$

Group Exercise

1. Compute the gradient wrt $\boldsymbol{w} \in \mathbb{R}^d$ of the following function using the vector chain rule:

$$\sigma_{\boldsymbol{w}}(\boldsymbol{x}) = \frac{1}{1 + \exp\left(-\sum_{j=1}^{d} w_j x_j\right)}$$

Group Exercise

1. Let $f(z) = \frac{1}{1+\exp(-z)}$ and let $g(\boldsymbol{w}) = \sum_{j=1}^{d} w_j x_j$. Notice that $f: \mathbb{R} \to \mathbb{R}$ and $g: \mathbb{R}^d \to \mathbb{R}$.
   It can be shown that $\frac{df(z)}{dz} = \frac{e^{-z}}{(1+e^{-z})^2} = f(z)\big(1 - f(z)\big)$

   Furthermore, $\nabla_{\boldsymbol{w}} g(\boldsymbol{w}) = \boldsymbol{x}$.

   Applying the vector chain rule gives

$$\nabla_{\boldsymbol{w}} \sigma_{\boldsymbol{w}}(x) = f\left(\sum_{j=1}^{d} w_j x_j\right)\left(1 - f\left(\sum_{j=1}^{d} w_j x_j\right)\right)\boldsymbol{x}$$

Further Reading

- https://explained.ai/matrix-calculus/index.html#sec4