Introduction to Neural Networks

Johns Hopkins University
Engineering for Professionals Program
605-447.71/625-438.71
Dr. Mark Fleischer

Module 2.1: Mathematical Review-Linear Algebra

This Sub-Module Covers …

- Some mathematical review of Linear Algebra.
- Some of the essential elements of matrix/vector algebra.
- This sub-module is then followed by a short quiz.

Mathematical Review

- Preliminaries: Notational conventions

Summation $\sum_{j=1}^{n} a_{ij} = a_{i1} + a_{i2} + \cdots + a_{in}$

Product $\prod_{j=1}^{n} a_{ij} = a_{i1} \cdot a_{i2} \cdot \ldots \cdot a_{in}$

Vectors If a row vector $\vec{a} = (a_1, a_2, \ldots, a_n)$ then $\vec{a}^T = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}$

Vector Operations

Vector Addition:

Given two vectors **a** and **b** *of the same size,* **a** + **b =** ($a_1 + b_1$, $a_2 + b_2$, …, $a_n + b_n$)

For subtraction, the "+" is substituted with a "-".

Vector Multiplication: (inner product, the dot product)

$$\vec{a} \cdot \vec{b} = (\vec{a}, \vec{b}) = \left\langle \vec{a}, \vec{b} \right\rangle = \sum_{i=1}^{n} a_i b_i = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n$$

= some scalar quantity.

Matrix-Vector Operations

The inner product ---reprise:

$$\mathbf{ab}^T = (a_1, a_2, \ldots, a_n)\begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}$$

This operation results in a single scalar value. If $\mathbf{a}^T\mathbf{b} = 0$ when both vectors **a** and **b** are non-zero vectors, then the vectors **a** and **b** are said to be *orthogonal*.

A *non-zero vector* is a vector in which at least one element is not zero.

(0, 0, 0, 1.3, 0, 0)

The Outer Product

$$\mathbf{a}^T\mathbf{b} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} \begin{pmatrix} b_1, b_2, \ldots, b_n \end{pmatrix} = \begin{pmatrix} a_1b_1 & a_1b_2 & \cdots & a_1b_n \\ a_2b_1 & a_2b_2 & \cdots & a_2b_n \\ \vdots & \vdots & \ddots & \vdots \\ a_nb_1 & a_nb_2 & \cdots & a_nb_n \end{pmatrix}$$

Matrix Multiplication

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}\begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \end{bmatrix}$$

2 × 2 2 × 2

2 × 2

~~[3 × 2] × [3 × 2]~~

[4 × 10] × [10 × 2] = [4 × 2]

$$\begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \end{bmatrix}\begin{bmatrix} \mathbf{b}_1 & \mathbf{b}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{a}_1 \cdot \mathbf{b}_1 & \mathbf{a}_1 \cdot \mathbf{b}_2 \\ \mathbf{a}_2 \cdot \mathbf{b}_1 & \mathbf{a}_2 \cdot \mathbf{b}_2 \end{bmatrix}$$

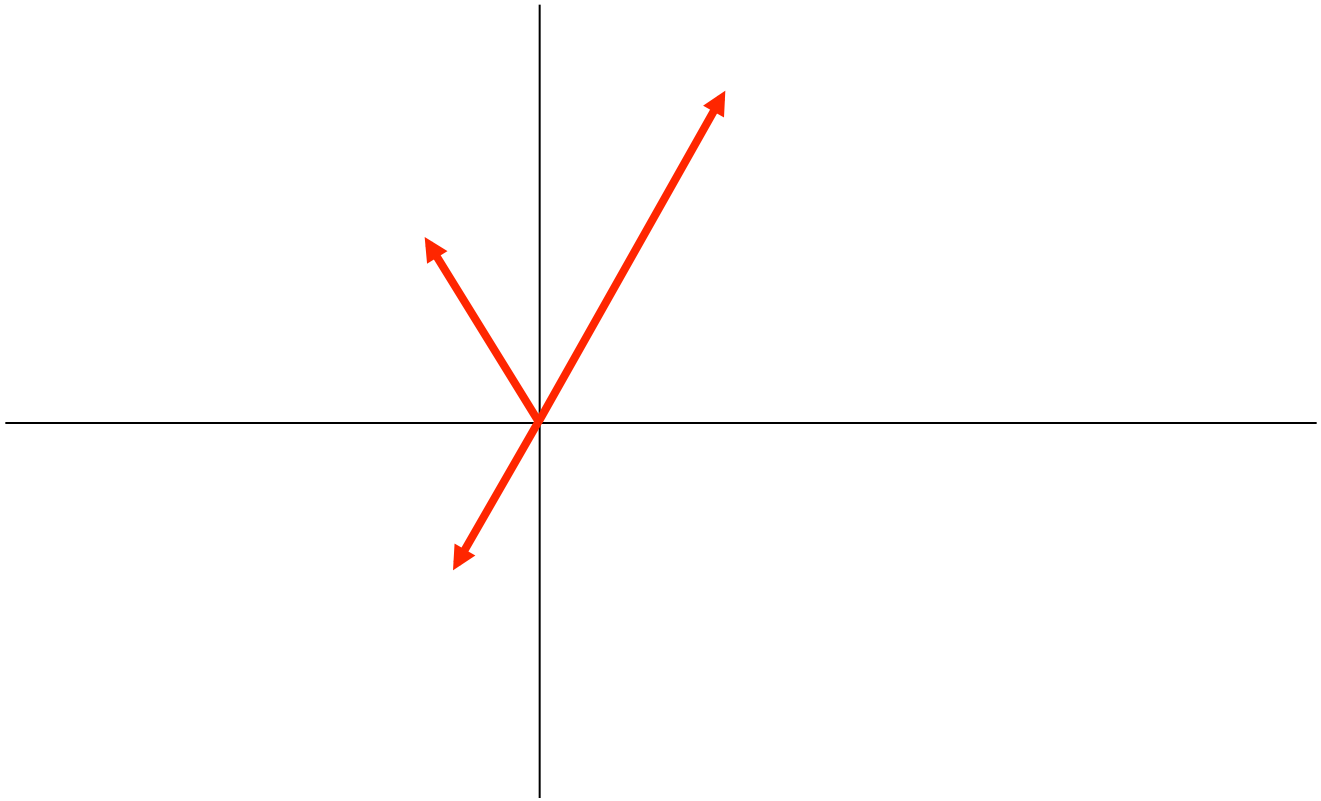
Linear Independence

A set of non-zero vectors $\mathbf{v}_i$, $i = 1, \ldots, n$ is said to be *linearly independent* where $\sum_{i=1}^{n} a_i \mathbf{v}_i = \mathbf{0}$ if and only if $a_i = 0$ for all i.

$$a_1 \begin{bmatrix} v_{11} \\ v_{12} \\ v_{13} \end{bmatrix} + a_1 \begin{bmatrix} v_{21} \\ v_{22} \\ v_{23} \end{bmatrix} + a_1 \begin{bmatrix} v_{31} \\ v_{32} \\ v_{33} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$$a_1 \begin{bmatrix} 1 \\ 0 \end{bmatrix} + a_2 \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Linear Independence



Introduction to Neural Networks

Johns Hopkins University
Engineering for Professionals Program

605-447.71/625-438.71

Dr. Mark Fleischer

Module 2.2: Mathematical Review-Differential Calculus

This Sub-Module Covers …

- Review of differential calculus:
  - derivatives, partial derivatives, gradients
  - directional derivatives
- The next sub-module covers:
  - Calculus-based optimization methods and related material:
    - First order necessary conditions.
    - Second order sufficiency conditions.
    - Definition of convexity.

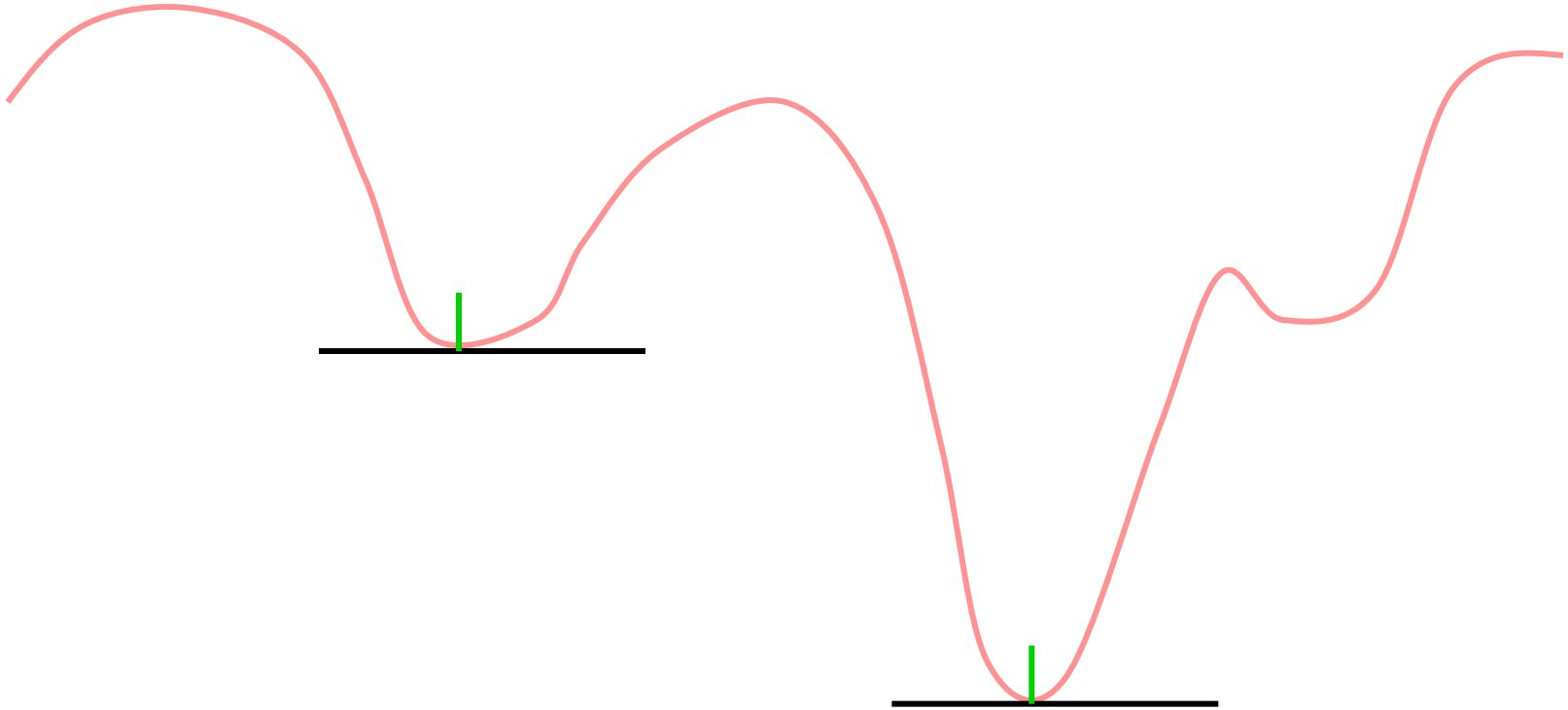
Optimization and Learning

Many types of optimization problems.

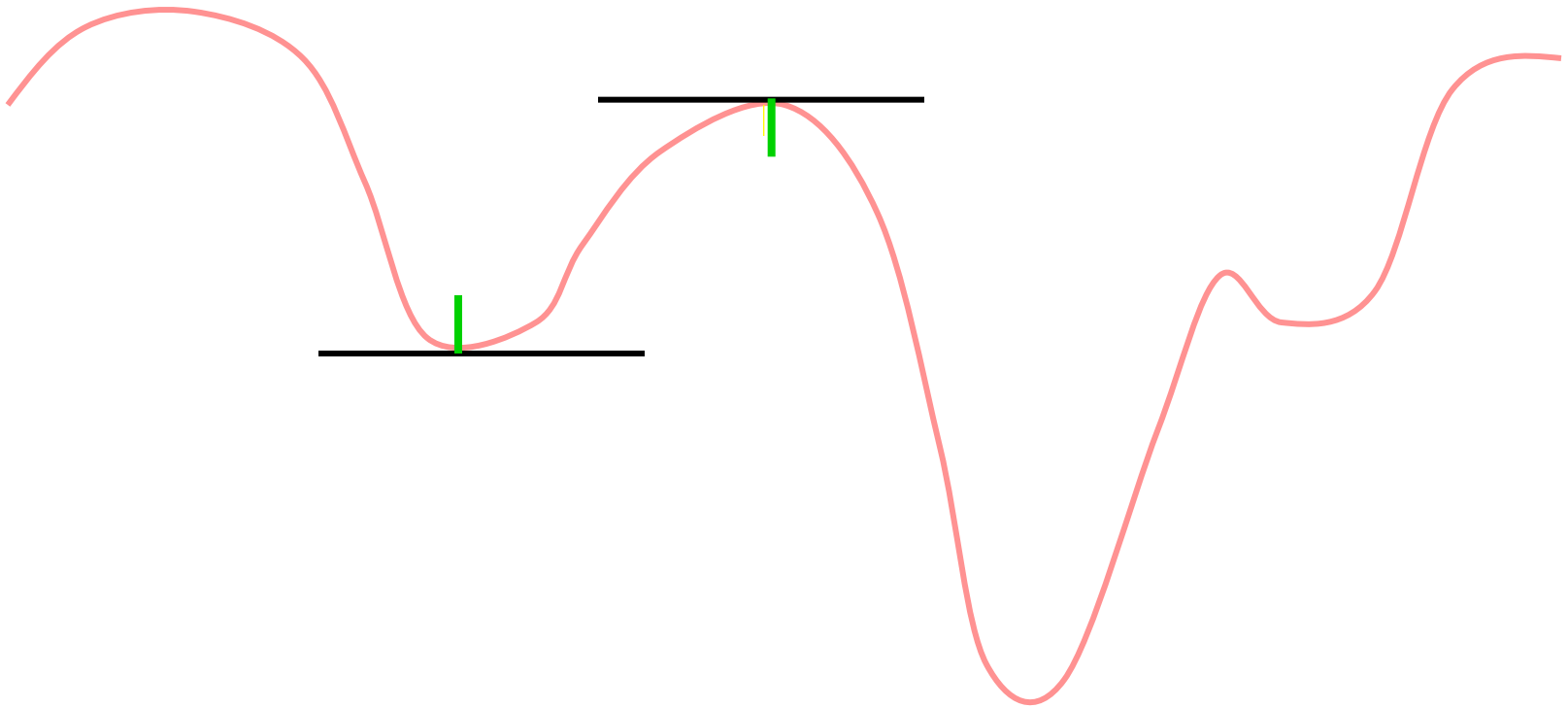
- For continuous functions, we use calculus-based methods.
- Problems with linear objective functions, linear constraint equations can be solved using Linear Programming methods
- Problems with non-linear objective functions, non-linear constraint equations can be solved using Non-Linear Programming methods.
- For now, we will not worry about constraints.

Training a neural network involves solving an optimization problem!

Calculus Based Optimization



Calculus Based Optimization



Derivatives

$$\frac{dy}{dx} = \frac{df(x)}{dx} = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

$$\frac{\partial f(x)}{\partial x_i} = \frac{f(x_1, \cdots, x_i + h, \cdots, x_n) - f(x_1, \cdots, x_i, \cdots, x_n)}{h}$$

$$\nabla f(\mathbf{x}) = \left(\frac{\partial f(x_1)}{\partial x_1}, \frac{\partial f(x_2)}{\partial x_2}, \cdots, \frac{\partial f(x_n)}{\partial x_n} \right)$$

Derivatives

$$f(x) = ax^n \Rightarrow \frac{df(x)}{dx} = f'(x) = nax^{n-1}$$

$$f(x_1, x_2) = ax_1x_2^n \Rightarrow \frac{\partial f(x_1, x_2)}{\partial x_2} = nax_1x_2^{n-1}$$

Directional Derivatives

- In partial derivatives, we consider the slope of a surface in the direction along one variable axis.
- In directional derivatives, we consider the slope of a surface in a specified direction.

Important for training neural networks!

Directional Derivatives

$$\frac{df(x,y,z)}{d\mathbf{d}} = \lim_{t \to 0^+} \frac{f(x+td_1, y+td_2, z+td_3) - f(x,y,z)}{t}$$

Each independent variable x, y and z is perturbed by a component of the direction vector **d** and multiplied by a factor *t*.

Directional Derivatives

$$\begin{aligned}
\frac{d\, f(x,y,z)}{d\mathbf{d}} &= \frac{\partial f}{\partial x}\frac{dx}{dt} + \frac{\partial f}{\partial y}\frac{dy}{dt} + \frac{\partial f}{\partial z}\frac{dz}{dt} \\
&= \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}\right) \cdot \left(\frac{dx}{dt}, \frac{dy}{dt}, \frac{dz}{dt}\right) \\
&= \nabla f(x,y,z) \cdot \mathbf{d}
\end{aligned}$$

Directional Derivatives

Introduction to Neural Networks

Johns Hopkins University
Engineering for Professionals Program

605-447.71/625-438.71

Dr. Mark Fleischer

Module 2.3: Mathematical Review-Calculus Based Optimization

This Sub-Module Covers …

- Calculus-based optimization methods and related material:
  - First order necessary conditions.
  - Second order sufficiency conditions.
  - Definition of convexity.
- Sets the stage for further mathematical review by exploring Metric Spaces in the next sub-module.

First-Order Necessary Conditions

TFAE

$$\frac{df(x^*)}{dx} = 0$$

$$\nabla f(\mathbf{x}^*) = \mathbf{0} = (0,0,\ldots,0)$$

$$\forall \mathbf{d}, \quad \nabla f(\mathbf{x}^*) \cdot \mathbf{d} = 0$$

$$\text{There exists a } t' > 0, \text{such that for all } t, \text{ where } 0 < t < t', \text{and for all non-zero vectors } \mathbf{d}$$

$$(\exists t' > 0, \ni \forall 0 < t < t' \wedge \mathbf{d} > \mathbf{0}),$$

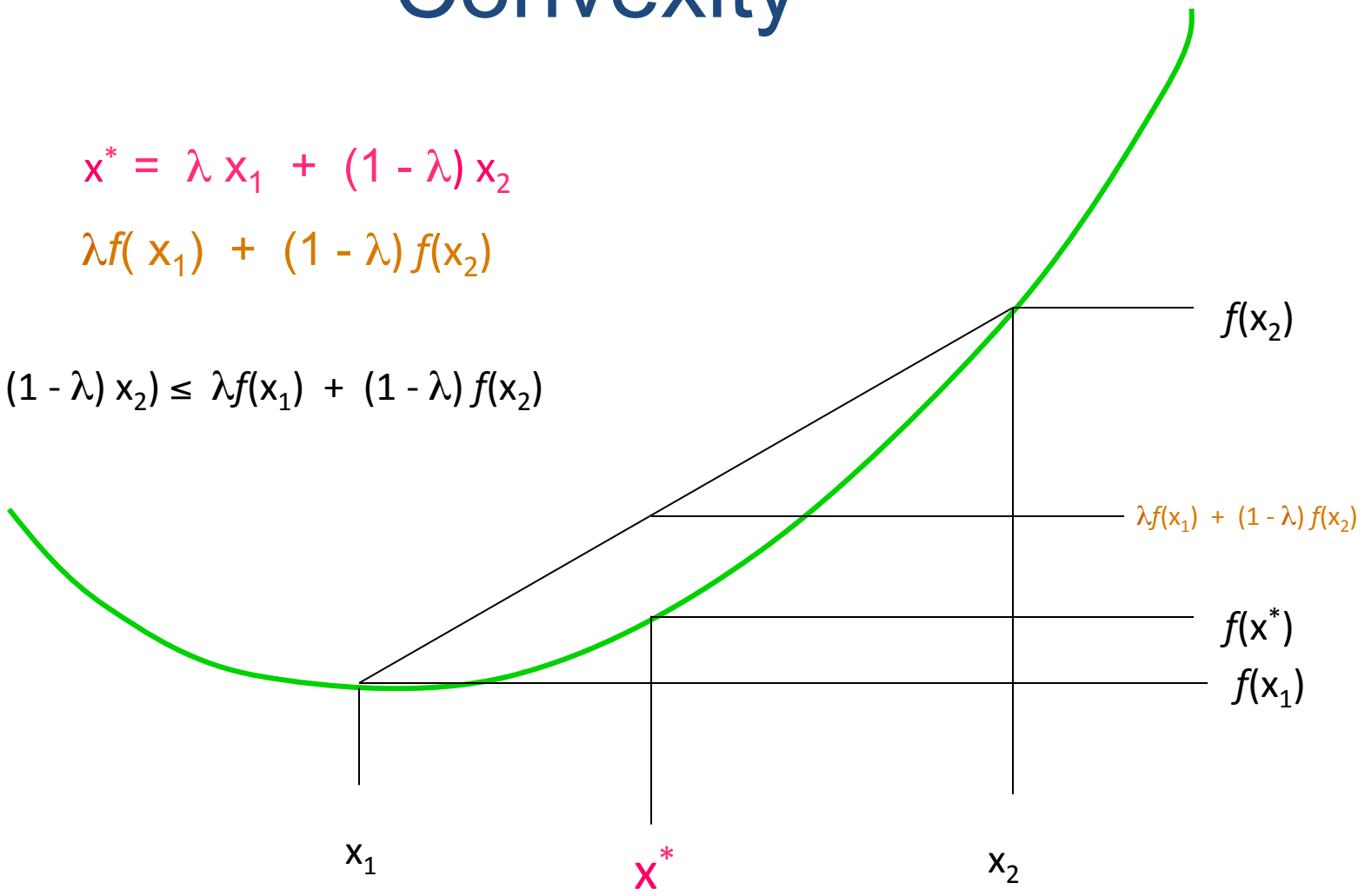
$$f(\mathbf{x}^*) < f(\mathbf{x}^* + t\mathbf{d})$$

Convexity

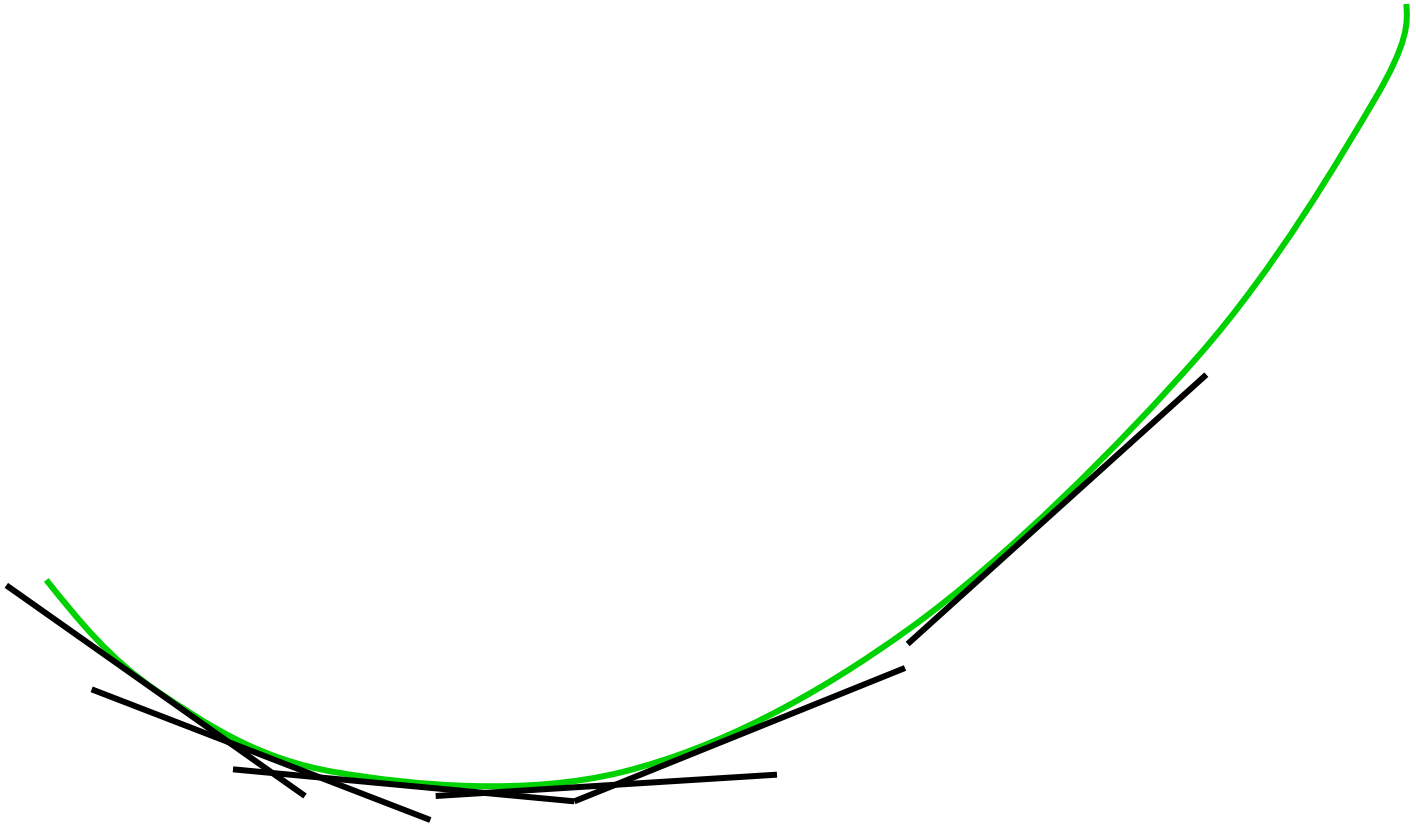
$$x^* = \lambda x_1 + (1 - \lambda) x_2$$

$$\lambda f(x_1) + (1 - \lambda) f(x_2)$$

$$f(\lambda x_1 + (1 - \lambda) x_2) \leq \lambda f(x_1) + (1 - \lambda) f(x_2)$$



Second Derivatives



Second-Order Sufficiency Conditions

TFAE (for determining minima)

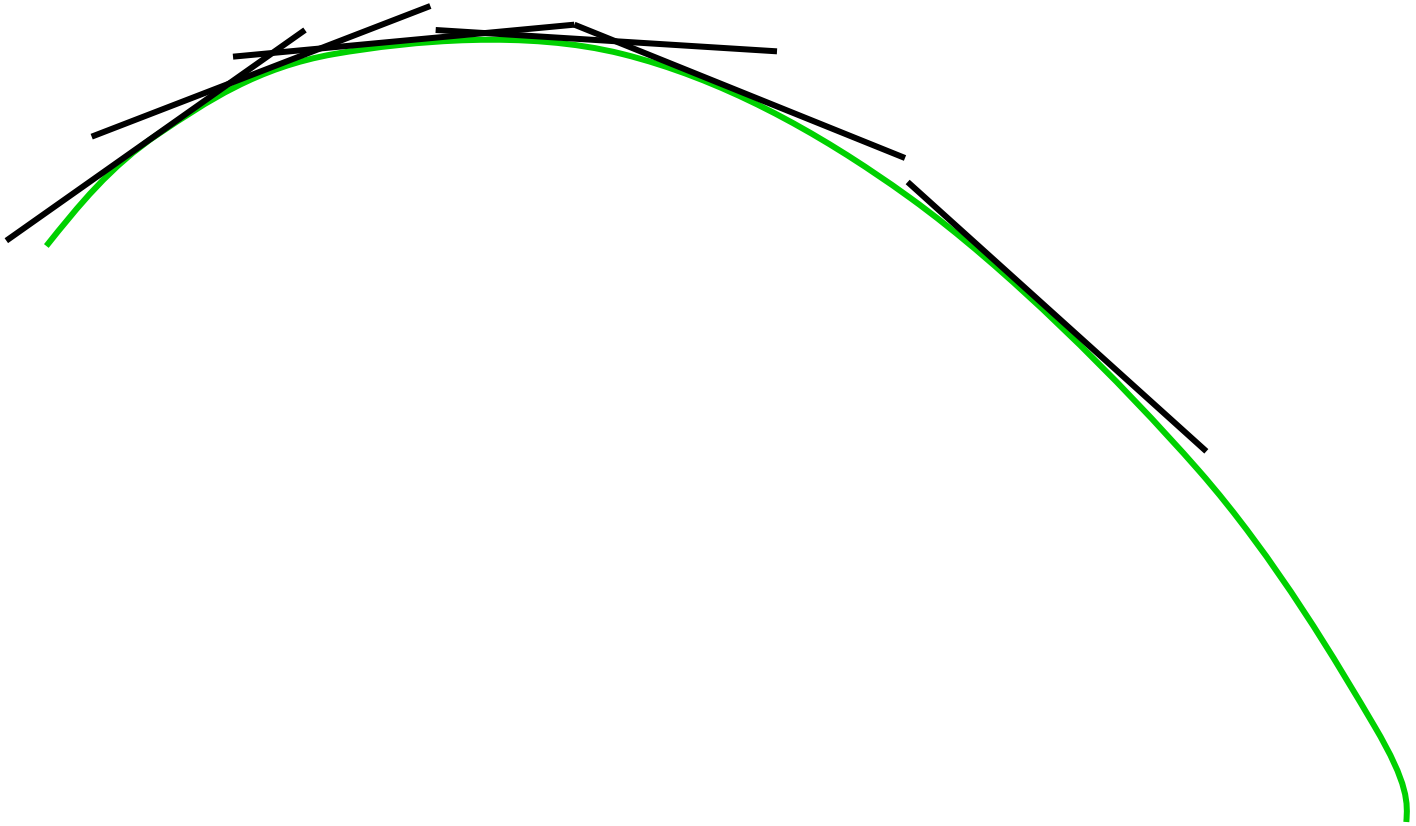
1. All second derivatives are positive
2. All points in a tangent plane (or hyperplane) have function values less than or equal to the objective function value.
3. The function is convex.

Item 1:

$$d^2y/dx^2 = f''(x^*) > 0$$

The Hessian Matrix $\mathbf{H}(\mathbf{x})$ is positive definite,
i.e., for all R^n, $\mathbf{x}^T\mathbf{H}\mathbf{x} \geq 0$:

Second Derivatives



Second-Order Sufficiency Conditions

TFAE (for determining minima)

1. All second derivatives are positive
2. All points in a tangent plane (or hyperplane) have function values less than or equal to the objective function value.
3. The objective function is convex.

Item 2:

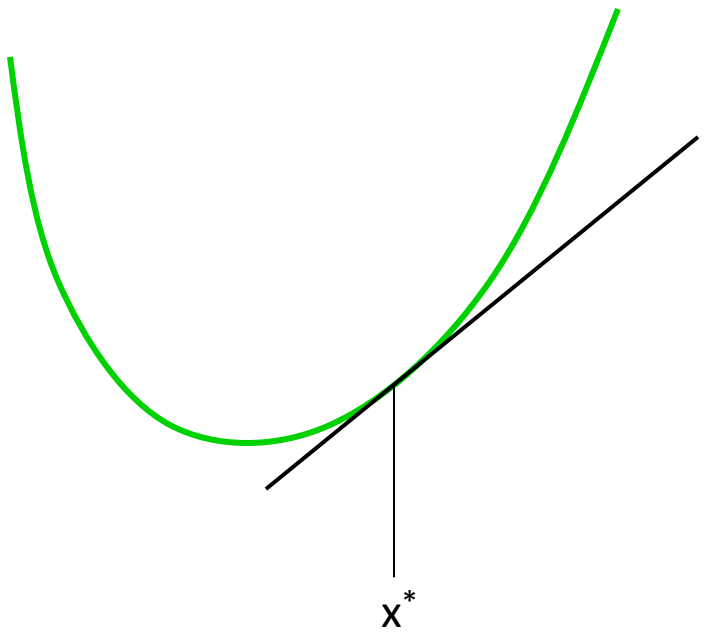
2. $\forall \mathbf{x}, \mathbf{x}^* \in R^n,$

$$f(\mathbf{x}) \geq f(\mathbf{x}^*) + \nabla f(\mathbf{x}^*)(\mathbf{x} - \mathbf{x}^*).$$

Item 2: Tangent Plane

2. $\forall \mathbf{x}, \mathbf{x}^* \in R^n$,

$$f(\mathbf{x}) \geq f(\mathbf{x}^*) + \nabla f(\mathbf{x}^*)(\mathbf{x} - \mathbf{x}^*).$$



$$f(\mathbf{x}^*) + \nabla f(\mathbf{x}^*)(\mathbf{x} - \mathbf{x}^*)$$

$= f(x^*) + m(x - x^*)$
$= f(x^*) + mx - mx^*$
$= mx + [f(x^*) - mx^*]$
$= mx + b$

Second-Order Sufficiency Conditions

TFAE (for determining minima)

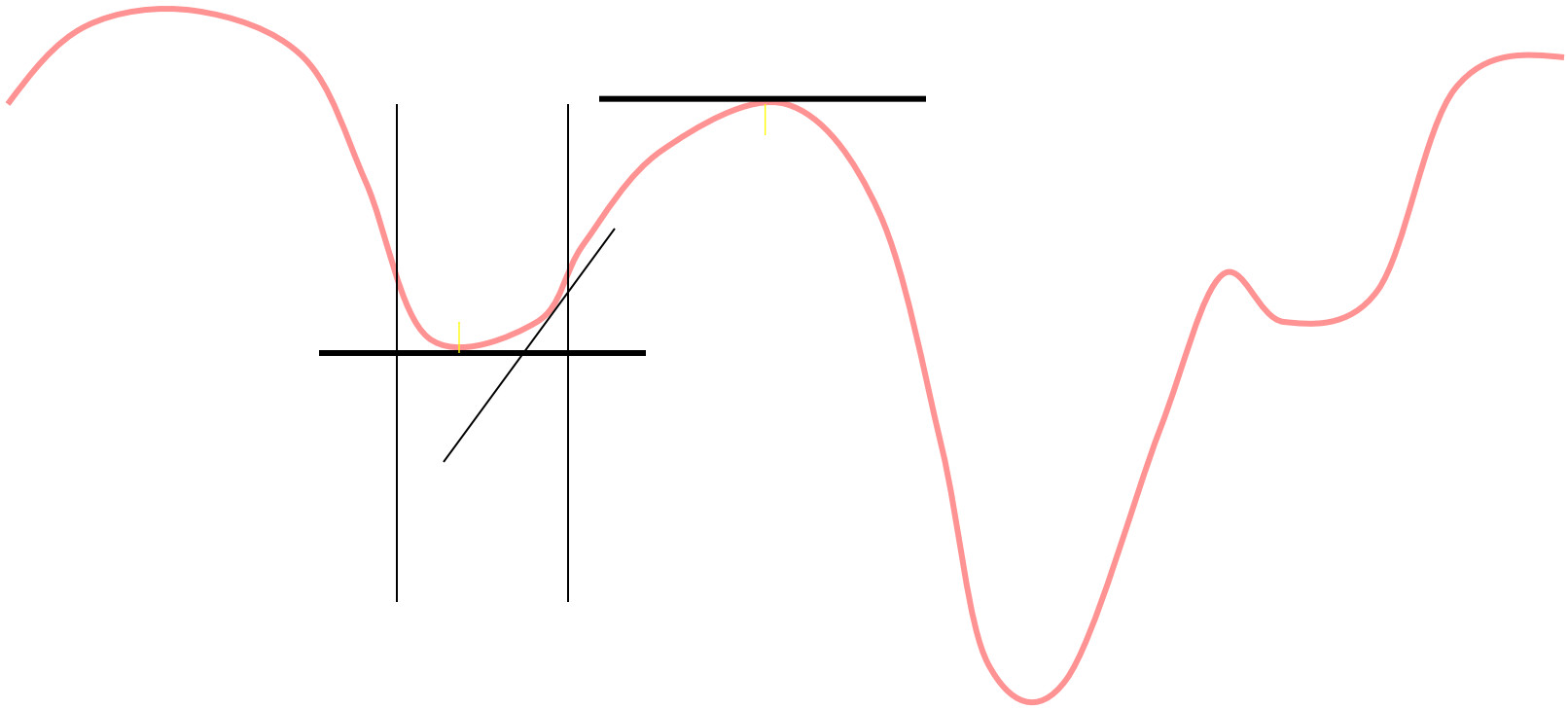
1. All second derivatives are positive
2. All points in a tangent plane (or hyperplane) have function values less than or equal to the objective function value.
3. The objective function is convex.

Item 3:

3. $\forall \mathbf{x}, \mathbf{x}^* \in R^n$ and $0 \leq \lambda \leq 1$,

$$\lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{x}^*) \geq f(\lambda \mathbf{x} + (1-\lambda)\mathbf{x}^*).$$

Calculus Based Optimization



Example

Suppose we want to minimize the function:

$$f(x_1, x_2) = x_1 x_2 - 6x_1 - 3x_2 + 18$$

Taking all partial derivatives, we see that:

$$\frac{\partial f}{\partial x_1} = x_2 - 6 = 0$$

$$\frac{\partial f}{\partial x_2} = x_1 - 3 = 0$$

$$x_1 = 3; x_2 = 6$$

Example

To show that it is a minimum, it is *sufficient* to show that one of the three second-order sufficiency conditions holds. Using the condition in item 2, we see that the equation for the tangent plane at say point (4,3) is:

Remembering that

$$\frac{\partial f}{\partial x_1} = x_2 - 6 = 0$$

$$\frac{\partial f}{\partial x_2} = x_1 - 3 = 0$$

$$\begin{aligned} f_P(x_1, x_2) &= f(4,3) + \nabla f(4,3) \begin{pmatrix} x_1 - 4 \\ x_2 - 3 \end{pmatrix} \\ &= -3 + (-3, 1) \begin{pmatrix} x_1 - 4 \\ x_2 - 3 \end{pmatrix} \\ &= -3 - 3(x_1 - 4) + 1(x_2 - 3) \\ &= -3x_1 + x_2 + 6 \end{aligned}$$

Example

Now consider any point (x_1, x_2) and compare the value of f_P with the objective function value f. For example, at point (0,0) the value of f_p = 6. For the objective function, $f(0,0)$ = 18.
Since the relationship between the objective function and tangent plane holds as item 2 above, it *suggests* that the second-order conditions hold. To establish this however requires that we prove this relation holds for *all* points (x_1, x_2).
Can you prove that they do?

Remembering that $f(x_1, x_2) = x_1 x_2 - 6x_1 - 3x_2 + 18$

$$
\begin{aligned}
f_P(x_1, x_2) &= f(4,3) + \nabla f(4,3) \begin{pmatrix} x_1 - 4 \\ x_2 - 3 \end{pmatrix} \\
&= -3 + (-3, 1) \begin{pmatrix} x_1 - 4 \\ x_2 - 3 \end{pmatrix} \\
&= -3 - 3(x_1 - 4) + 1(x_2 - 3) \\
&= -3x_1 + x_2 + 6
\end{aligned}
$$

Taylor's Theorem

Single variable case

$$
\begin{aligned}
f(x) &= a_0 + a_1(x - x_0) + a_2(x - x_0)^2 \ldots \\
&= f(x_0) + (x - x_0) f'(x_0) + \frac{1}{2!}(x - x_0)^2 f''(x_0) \ldots
\end{aligned}
$$

Multi - variable case

$$
\begin{aligned}
f(\mathbf{x}) &= a_0 + a_1(\mathbf{x} - \mathbf{x}_0) + a_2(\mathbf{x} - \mathbf{x}_0)^2 \ldots \\
&= f(\mathbf{x}_0) + (\mathbf{x} - \mathbf{x}_0) \nabla f(\mathbf{x}_0) + \frac{1}{2!}(\mathbf{x} - \mathbf{x}_0)^{\mathrm{T}} \mathbf{H} (\mathbf{x} - \mathbf{x}_0) \ldots
\end{aligned}
$$

Mathematical Review

So far we've reviewed:

- Basic Vector/Matrix operations: inner, outer products
- Linear Independence
- Differential calculus
- Partial Differentiation,
- Directional Derivatives $\nabla f(x,y,z)\cdot\mathbf{d} = \left\|\nabla f(x,y,z)\right\| \times \left\|\mathbf{d}\right\| \times \cos\theta$
- Gradient vector $\nabla f(\mathbf{x}) = (\partial f(\mathbf{x})/\partial x_1,\ \partial f(\mathbf{x})/\partial x_2\ ,\ \ldots\ ,\ \partial f(\mathbf{x})/\partial x_n)$
- First order necessary conditions, Second order sufficiency conditions

In the next sub-modules we will review:

- Metric Spaces:
  - Distance and Magnitude of vectors, matrices
  - Definitional requirements of "norms"
    - **Positivity, homogeneity, Triangle Inequality**

Introduction to Neural Networks

Johns Hopkins University
Engineering for Professionals Program
605-447.71/625-438.71
Dr. Mark Fleischer

Module 2.4: Mathematical Review-Metric Spaces

This Sub-Module Covers …

- Some mathematical review of Metric Spaces and will set the stage for using:
  - directional derivatives in conjunction with calculus based optimization to define the
  - Method of Steepest Descent (MOSD)---used to train Perceptrons.
- Also provides additional insights into dynamical systems.

What is 'Distance' and 'Magnitude'?

- Need capability to rank and/compare objects using a simple criterion.
- Want a flexible yet abstract notion of distance or length or magnitude.
- The following principles provide a very abstract, general way of defining the essential properties of a 'length'.

Length

Desirable properties of a “length”(magnitude or norm):

1. Positivity: $||\mathbf{y}|| \geq 0$ for all y and $||\mathbf{y}|| = 0$ if and only if $\mathbf{y} = \mathbf{0}$.
2. Homogeneity: $||c\mathbf{y}|| = |c|\ ||\mathbf{y}||$ for all scalars c and vectors $\mathbf{y}$.
3. The Triangle Inequality: $||\mathbf{x} + \mathbf{y}|| \leq ||\mathbf{x}|| + ||\mathbf{y}||$ for all vectors $\mathbf{x}$ and $\mathbf{y}$.

Examples of a Norm

$$\left|\vec{a}\right| \equiv \left\|\vec{a}\right\| = \left[\sum_{i=1}^{n} a_i^2\right]^{1/2}$$

This is called the Euclidean Norm.

$$\left|\vec{a}\right|_p \equiv \left\|\vec{a}\right\|_p = \left[\sum_{i=1}^{n} a_i^p\right]^{1/p}$$

This is the *p*-norm.

If $\mathbf{y} = \sum_{1=1}^{n} \alpha_i \mathbf{v}_i$ for some set of *n* linearly independent vectors $\mathbf{v}_i$. Then

$\left\|\mathbf{y}\right\|_{\mathbf{V}} \equiv \sum_{i=1}^{n} \left|\alpha_i\right|$ is a norm with respect to the matrix $(\mathbf{v}_1 | \mathbf{v}_2 | \cdots |^{\mathbf{v}}_{\mathrm{n}})$.

Metric Spaces

We've already alluded to some issues for measuring mathematical things.
We need tools that allow us to *rank* mathematical objects and *rank relationships* between and among mathematical objects.

1. $\rho(x, y) = \rho(y, x) \geq 0$. Positivity
2. $\rho(x, x) = 0$; $\rho(x, y) = 0 \Leftrightarrow x = y$. Homogeneity
3. $\rho(x, z) + \rho(z, y) \geq \rho(x, y)$. Triangle Inequality

where ρ is a function defined on $M \times M$, *hence* (M, ρ) is called a Metric Space

The metric function ρ can take on many forms!

We define ρ based on what is necessary and convenient!

Various Forms of Metrics

$$\rho_\infty(x,y) = \operatorname{Max}_i |x_i - y_i|$$

$$|x_i - y_i| \le |x_i - z_i| + |z_i - y_i|$$

$$\rho(f(x), g(x)) = \operatorname{Max}_{x \in [a,b]} |f(x) - g(x)|$$

$$\rho(\vec{a},\vec{b}) = \left\|\vec{a} - \vec{b}\right\| = \left[\sum_{i=1}^{n} (a_i - b_i)^2\right]^{1/2} = \left(\vec{a} - \vec{b}, \vec{a} - \vec{b}\right)^{1/2}$$

Orthogonality and Angles

We can posit that for any two mathematical objects, there corresponds a number between -1 and 1 that conveys how these objects are related.

$$\cos\theta = \frac{\langle \mathbf{a}, \mathbf{b} \rangle}{\|\mathbf{a}\|\|\mathbf{b}\|}$$

We can say that this corresponds to the angle between two vectors.

Convergent Sequences

Limit Points: A sequence $\{x_n\}$ is convergent and has a limit point

$$x^* \in M, \text{ if } \rho(x_n, x^*) \to 0 \text{ as } n \to \infty,$$

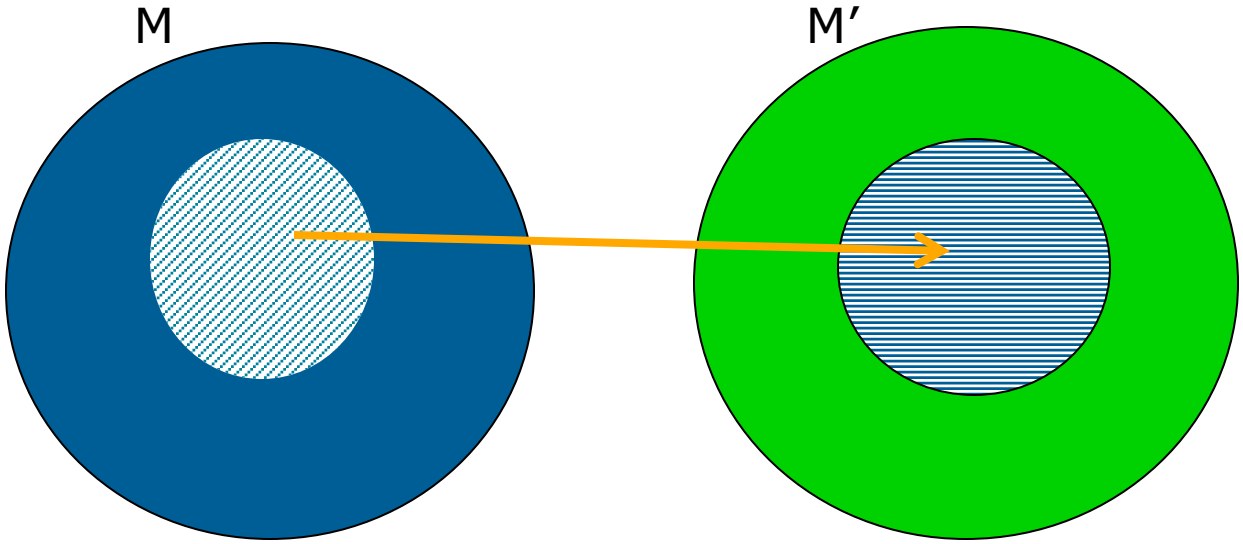
$$x_n \to x^*$$

Mappings in Metric Spaces

- Define two sets and two Metric Spaces:

$$D(f) = \{x : x \in M \wedge f(x) \text{ is defined}\}$$

$$R(f) = \{x' : x' \in M' \wedge \exists\, x \in D(f) \ni x' = f(x)\}$$



Boundedness

$$\exists k \geq 0, \ni \forall x, y \in D(f)$$

$$\rho'\big(f(x), f(y)\big) \leq k\rho(x, y)$$

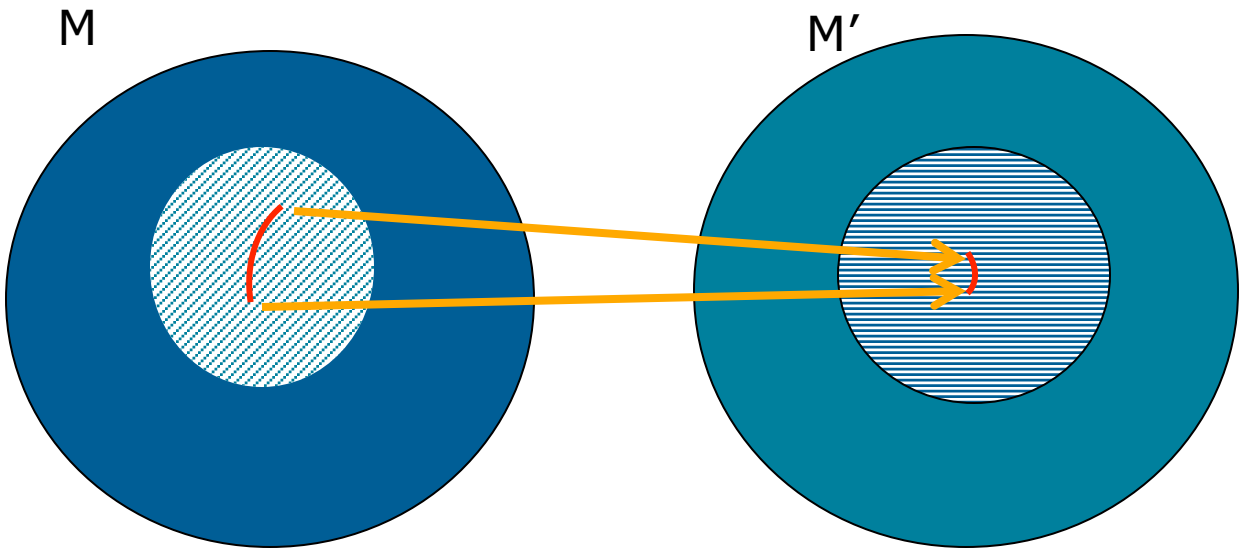
Metric value associated with the range.

Metric value associated with the domain.

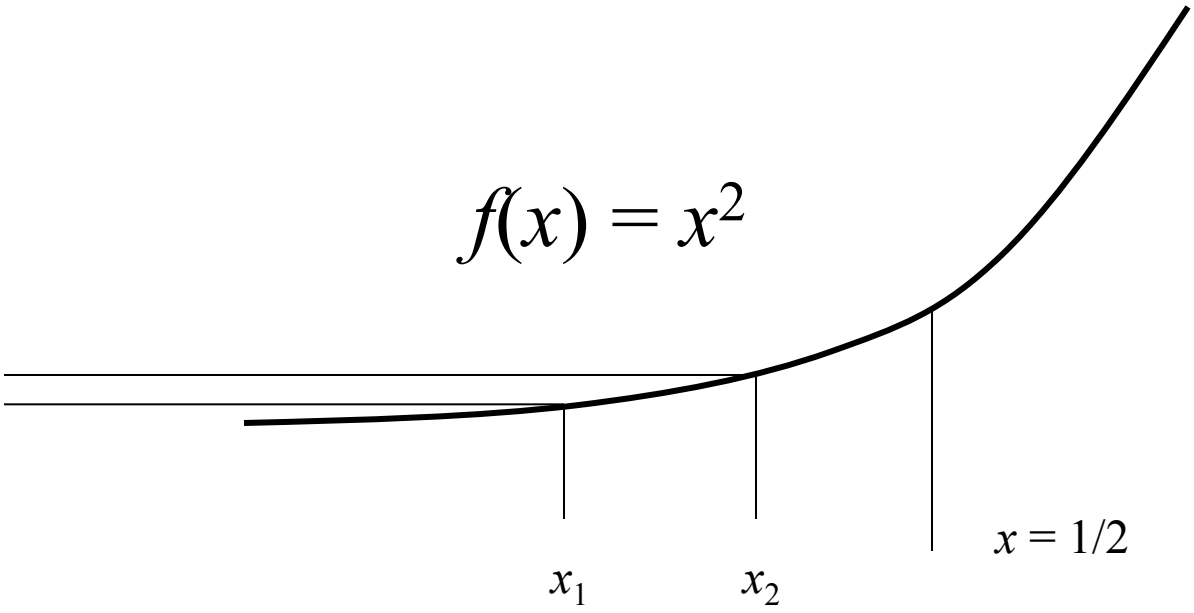
Contraction Mapping

A Mapping is a contraction if it is bounded and

$$\exists k \geq 0 \ \wedge \ 0 \leq k < 1, \ni \forall x, y \in D(f)$$
$$\rho'\big(f(x), f(y)\big) \leq k\rho(x, y)$$



An Example



Define $\rho(x_1, x_2) = |x_1 - x_2|$

A Useful Fact

$\Delta y = f'(\xi)\Delta x$ where ξ is some "intermediate" value.

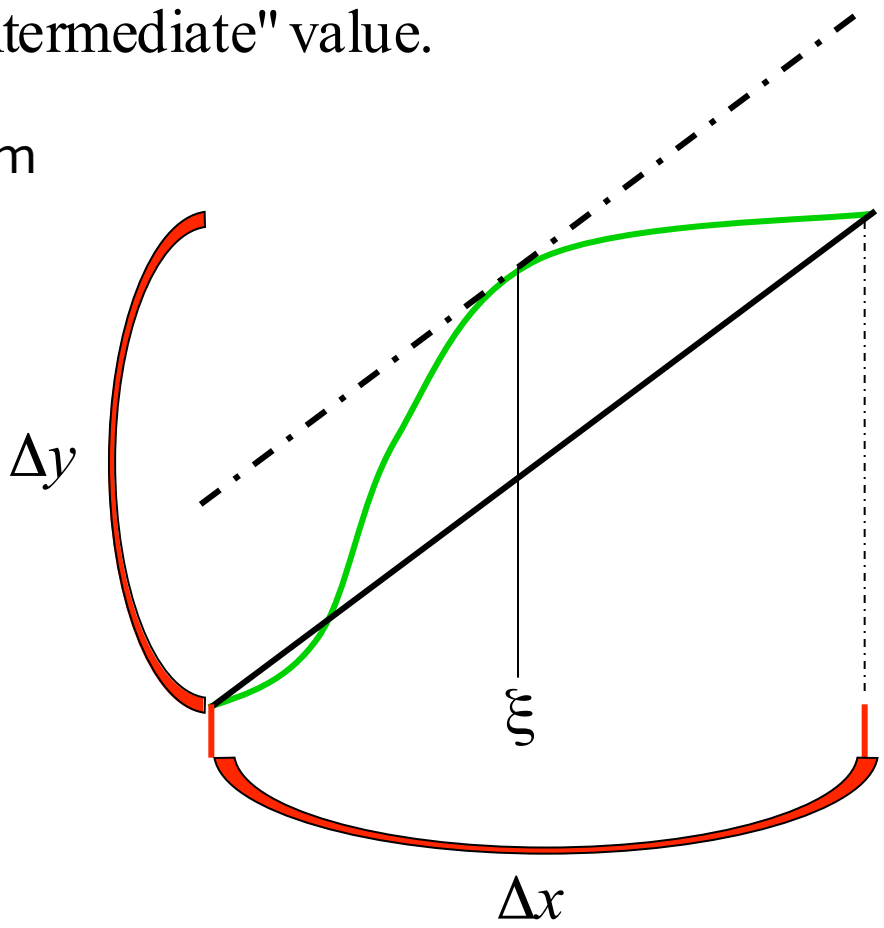
The Mean Value Theorem

$$|\Delta y| = |f(x_1) - f(x_2)|$$

$$= |f'(\xi)\Delta x|$$

$$\leq |f'(\xi)||\Delta x|$$

$$= f'(\xi)|x_1 - x_2|$$



Putting It All Together

$$|\Delta y| = |f(x_1) - f(x_2)| = \rho\big(f(x_1), f(x_2)\big)$$

Therefore,

$$\rho\big(f(x_1), f(x_2)\big) \le f'(\xi)\rho(x_1, x_2)$$

$$\text{and since } 0 \le x_1, x_2 < \tfrac{1}{2}, \text{ and } f'(x) = 2x$$

$$\text{then it follows that } f'(\xi) < 1$$

$$\forall x, y \in D(f) \text{ and } 0 \le k < 1$$

$$\rho'\big(f(x), f(y)\big) \le k\rho(x, y)$$

$$\text{is a Contraction Mapping}$$

A Numerical Example

$$\rho\left(\frac{1}{4},\frac{1}{2}\right)=\frac{1}{4}$$

$$\begin{aligned}\rho\left(f\left(\tfrac{1}{4}\right),f\left(\tfrac{1}{2}\right)\right)&=\rho\left(\left(\tfrac{1}{4}\right)^2,\left(\tfrac{1}{2}\right)^2\right)\\&=\rho\left(\tfrac{1}{16},\tfrac{1}{4}\right)\\&=\tfrac{3}{16}\end{aligned}$$

Note that $\frac{3}{16}<\frac{1}{4}$

How Would You Use This in The Context of Fixed Points?

$$\forall x_1, x_2 \in D(f) \text{ and } 0 \le k < 1$$
$$\rho'\left(f(x_1), f(x_2)\right) \le k\rho(x_1, x_2)$$

$$\forall x_1, x_2 \in D(f) \text{ and } 0 \le k < 1$$
$$\rho'\left(x^*, f(x_2)\right) \le k\rho(x^*, x_2)$$

because by definition, a fixed point x^* is such that

$$f\left(x^*\right) = x^*$$

Dynamical Systems

- Neural Networks can be fashioned into dynamical systems.
- Feeding outputs back as inputs and cycling them ala fixed point exercise.
- How does such a system behave?
- Stay tuned.

A SIMPLE PROOF OF THE CHAIN RULE

1 Let's Start at the Very Beginning

Let's look at the basic definition of the derivative where we'll emphasize certain, specific patterns that will help us navigate through a more complicated-looking derivative. First, the derivative a function $g(x)$ should be quite familiar:

$$\frac{dg}{dx} = \lim_{h \to 0} \frac{g(x+h) - g(x)}{h} \tag{1}$$

which tells us what the relative change is of the function value g compared to changes in the value of x. In other words, we want to know how fast g changes if we slightly change the value of its argument by a small amount h. Notice that the increment h is thus *added to the argument of the function* which is why we refer to this as the derivative of g *with respect to* x — because x is the variable we are *perturbing* with h. The value of h simply represents the difference between $x + h$ and x —the "run" and $g(x+h)-g(x)$ is the "rise" and their ratio corresponds to the slope and in the limit corresponds to the derivative. When we add the increment h directly to the argument of the function, as is the case here, we can write the derivative as $dg(x)/dx$ or simply dg/dx — again, the derivative of g *with respect to* x which tells us the proportional change in g when we change x by adding h to x. Also notice that the increment h in the numerator of (1) is the very same increment that's in the denominator of (1). This is necessary so that we can correctly compute the rise over the run. If the h in the numerator was different than the one in the denominator, we wouldn't actually be calculating the slope would we.

2 What About a Function of a Function?

Now let's look at a compound function: $f(g(x))$. In this case, as before, the argument or variable of the function g is x, but the argument or variable of the function f is g! So changing x changes g and changes in g change the value of f. So if we want to define the derivative of f *with respect to* x, we have to add the increment h to the variable x and so by the definition of the derivative and using the same general form as in (1),

$$\frac{df}{dx} = \lim_{h \to 0} \frac{f(g(x+h)) - f(g(x))}{h} \tag{2}$$

So this is the derivative of f with respect to x based on the definition of a derivative. The problem with (2) however is that it's hard to actually use and work with. This is due to several issues. For one thing, the argument of the function f is the value of g and *not* the value x. Remember, that in (1) we added the increment h directly to the argument of that function. In this case, we're not exactly doing that. Instead of adding the increment h to the argument of f, we're adding h to the argument of g!

It might be helpful therefore if we could somehow modify the numerator of (2) so that the increment h is added to the argument of f as this is the function we're taking the derivative of and since the argument of f is g, we want to add the increment h to g and not to x. Thus, we need to somehow change the expression thusly:

$$f(g(x+h)) \longrightarrow f(g(x)+h)$$

or, ignoring the argument x since it is the argument of g (think of g as the argument or variable for f), we want to somehow establish a conversion

$$f(g(x+h)) \longrightarrow f(g+h)$$

in the numerator of (2). In this way, we would be adding the increment h directly to the argument of f (which is g) and not to the argument of the function g (which is x)! Do you see the difference? So the question is: is there a way to change $f(g(x+h))$ to look something more like $f(g(x)+h)$ or $f(g+h)$ in a legitimate way? To do this requires that we find some way of relating or converting $g(x+h)$ to $g(x)+h$ or something similar to it. Do you see a way to connect these two expressions? (*hint:* look at equation (1)).

From (1) we see quite directly (after doing just a little algebra) that

$$g(x+h) = g(x) + h\frac{dg}{dx} \tag{3}$$

Now we're getting somewhere. Note also that in (3) as $h \to 0$ so does hdg/dx which means the increment goes to 0 just like it's supposed to in the definition of a derivative like in (1). So let's simplify some notation and write $h\frac{dg}{dx}$ as $\hat{h}$ so that (3) looks like

$$g(x+h) = g(x) + \hat{h} \tag{4}$$

which looks much cleaner and simpler. Now, we can substitute (4) for $g(x+h)$ in equation (2) which now looks like

$$\frac{df}{dx} = \lim_{h\to 0} \frac{f(g(x)+\hat{h}) - f(g(x))}{h} \tag{5}$$

or, ignoring the argument x of g to emphasize that it is g that is the argument of f, (5) can be rewritten as

$$\frac{df}{dx} = \lim_{h \to 0} \frac{f(g + \hat{h}) - f(g)}{h} \tag{6}$$

Now we see very clearly that g is the variable or argument of f and that we've added a type of increment $\hat{h}$ directly to the argument of f just like in (1). Thus, the right-hand side of (6) is beginning to look a lot like df/dg!

The only problem in (6) is that the increment $\hat{h}$ in the numerator of (6) is not the same as the increment h in the denominator of (6). These two increments must be the same for there to be a proper evaluation of the slope and to correspond to the definition of a derivative as in (1). How can we get these two increments to be the same?

Well, if we multiply the the quantity in (6) by $h/\hat{h}$ then the h's would cancel as in

$$\frac{df}{dx} = \lim_{h \to 0} \left[\frac{f(g + \hat{h}) - f(g)}{\not{h}} \right] \cdot \frac{\not{h}}{\hat{h}}$$

and we'd end up with an $\hat{h}$ in both the numerator and denominator and that really would be df/dg! But we can't just willy-nilly multiply the the right-hand side of (6) by $h/\hat{h}$ just to make it look like we want it to look — doing that could change its value. The only multiplication that we can do to (6) *without changing its value* is to multiply (somehow) by a factor that is *always* 1! So instead of just multiplying (6) by $h/\hat{h}$, we're going to multiply by

$$\frac{h}{\hat{h}} \cdot \frac{\hat{h}}{h}$$

which is just another way of writing 1. Thus, (6) becomes

$$\frac{df}{dx} = \lim_{h \to 0} \left[\frac{f(g + \hat{h}) - f(g)}{h} \right] \cdot \frac{h}{\hat{h}} \cdot \frac{\hat{h}}{h} \tag{7}$$

and the value is still the same. Notice that we can now cancel the h's thusly:

$$\frac{df}{dx} = \lim_{h \to 0} \left[\frac{f(g + \hat{h}) - f(g)}{\not{h}} \right] \cdot \frac{\not{h}}{\hat{h}} \cdot \frac{\hat{h}}{h} \tag{8}$$

and then moving the $\hat{h}$ in the denominator into the denominator of the bracketed quantity we get

$$\frac{df}{dx} = \lim_{h \to 0} \left[\frac{f(g + \hat{h}) - f(g)}{\hat{h}} \right] \cdot \frac{\hat{h}}{h} \tag{9}$$

and now we see that the quantity in brackets is actually $\frac{df}{dg}$ because g is the argument of f to which the increment $\hat{h}$ has been added, the increments in the numerator and denominator of the bracketed quantity are the same, and as $h \to 0$ so does $\hat{h}$. Thus, everything in the brackets corresponds to the definition of a derivative of the function f with respect to g!

So what about the fraction $\hat{h}/h$ on the right of the brackets in (9)? Well, recall from (3) and (4) that $\hat{h} = h\frac{dg}{dx}$ therefore

$$\frac{\hat{h}}{h} = \frac{h(dg/dx)}{h} = \frac{dg}{dx} \tag{10}$$

because the h's cancel and so (9) becomes

$$\frac{df}{dx} = \lim_{h \to 0} \left[\frac{f(g + \hat{h}) - f(g)}{\hat{h}} \right] \cdot \frac{dg}{dx} \tag{11}$$

or

$$\frac{df}{dx} = \frac{df}{dg} \cdot \frac{dg}{dx} \tag{12}$$

3 Final Thoughts

In our quest to understand the chain rule, we first had to use symbols and abstractions to clearly state what we meant by a derivative of some function with respect to some variable. Such a quantity represents the rate of change of a function's value as its argument is changed and we were able to write (1) to symbolize this concept. Then using good 'ole algebra and our knowledge of limits, we manipulated the symbols, all the while adhering to fundamental rules of mathematics to show a *chain* of reasoning and logic to arrive at the simple rule that you will use a lot — the chain rule — a multiplication of rates of change where one variable affects another. Remember, a chain is only as strong as its weakest link, so learn this rule well!

◇ ◇ ◇