ECE 586: Vector Space Methods
Lecture 16 Flip Video: Derivatives in Banach Spaces

Henry D. Pfister
Duke University

Engineering and Optimization

The foundation of engineering is the ability to use math and physics to design and optimize complex systems.

Computers have now made this possible on an unprecedented scale.

Well-known applications include:

- Physical Modeling: fitting large physical models (e.g., weather) to huge amounts of collected data
- Machine Learning: optimizing non-linear functions (e.g., neural networks) to minimize classification loss on supervised samples
- In both cases, optimization benefits from computing derivatives

Derivatives

In vector analysis, derivatives provide local linear approximations:

- For $f\colon \mathbb{R}^n \to \mathbb{R}^m$ with Jacobian matrix $J(\underline{x}) \in \mathbb{R}^{m\times n}$, this gives

$$f(\underline{x}+\underline{h}) \approx f(\underline{x}) + J(\underline{x})\underline{h}, \qquad J(\underline{x}) \triangleq f'(\underline{x}) \triangleq \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(\underline{x}) & \frac{\partial f_1}{\partial x_2}(\underline{x}) & \cdots & \frac{\partial f_1}{\partial x_n}(\underline{x}) \\ \frac{\partial f_2}{\partial x_1}(\underline{x}) & \frac{\partial f_2}{\partial x_2}(\underline{x}) & \cdots & \frac{\partial f_2}{\partial x_n}(\underline{x}) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1}(\underline{x}) & \frac{\partial f_m}{\partial x_2}(\underline{x}) & \cdots & \frac{\partial f_m}{\partial x_n}(\underline{x}) \end{bmatrix}$$

- For $f\colon X \to Y$, it is best to view $f'(x)$ as a linear transform $T\colon X \to Y$ from the domain to codomain. This transform maps an infinitesimal input perturbation $\underline{h}$ to an infinitesimal output perturbation

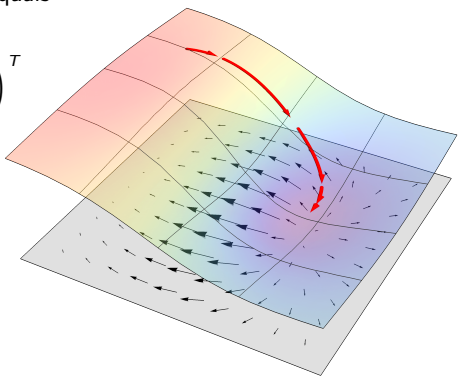
$$f'(\underline{x})(\underline{h}) = T\underline{h} = J(\underline{x})\underline{h}$$

- Notice that this definition of the derivative requires a linear structure (to define differences) and a topology (to define convergence) for X and Y

Gradient Descent in $\mathbb{R}^n$

- Consider a cost function $f\colon \mathbb{R}^n \to \mathbb{R}$
  - Can we adjust $\underline{x} \in \mathbb{R}^n$ to minimize the cost?
  - The gradient vector $\nabla f(\underline{x}) \triangleq f'(\underline{x})^T$ equals the direction of maximum increase

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



- Gradient descent
  - Discrete time (step size δ_n)

$$\underline{x}_{n+1} = \underline{x}_n - \delta_n \nabla f(\underline{x}_n)$$

  - Continuous time:

$$\frac{d}{dt}\underline{x}(t) = -\nabla f\big(\underline{x}(t)\big)$$

  - Standard method for training machine learning models like neural networks

Derivatives in Banach Spaces

In abstract math, derivatives are usually introduced using Banach spaces:

- For a function $f : X \to Y$, the concept of a derivative requires a linear structure (to define differences) and a topology (to define convergence) on both X and Y
- If $X = \mathbb{R}^n$ and $Y = \mathbb{R}^m$, then the derivative of f is a linear transform from X to Y represented by the Jacobian matrix $f'(\underline{x}) \in \mathbb{R}^{m \times n}$
- Thus, we generally assume that $f : X \to Y$ is a mapping from the Banach space $(X, \|\cdot\|_X)$ to the Banach space $(Y, \|\cdot\|_Y)$
- For directional derivatives, one needs even less structure and it suffices to let X be only a vector space.

What is Meant by Differentiable?

In abstract math, there are many related definitions that are slightly different. To distinguish between these, one often uses names that are less common in the engineering literature (e.g., Hamel vs. Schauder basis).

Definition (Differentiable)

Let $f : X \to Y$ be a mapping from a Banach space $(X, \|\cdot\|_X)$ to a Banach space $(Y, \|\cdot\|_Y)$. Then, f is Fréchet differentiable at $\underline{x}$ if there is a continuous linear transformation $T : X \to Y$ satisfying

$$\lim_{\underline{h} \to \underline{0}} \frac{\|f(\underline{x} + \underline{h}) - f(\underline{x}) - T(\underline{h})\|_Y}{\|\underline{h}\|_X} = 0,$$

where the limit is with respect to the implied Banach space mapping $X \to \mathbb{R}$. In this case, the Fréchet derivative $f'(\underline{x})$ equals T.

Properties of the Derivative

A useful property of the derivative is a characterization Lipschitz continuity.

Lemma

Let X, Y be Banach spaces and $f\colon X \to Y$ be a function. If the Fréchet derivative $f'(\underline{x})$ exists and satisfies $\|f'(\underline{x})\|_{\mathrm{op}} \leq L$ for all $\underline{x}$ in a convex set $A \subseteq X$, then f is Lipschitz continuous on A with Lipschitz constant L.

One can also prove a general chain rule for the Fréchet derivative.

Theorem

Let X, Y, Z be Banach spaces and let $f\colon X \to Y$ and $g\colon Y \to Z$ be functions. If f is Fréchet differentiable at $\underline{x}$ and g is Fréchet differentiable at $\underline{y} = f(\underline{x})$, then $(g \circ f)(\underline{x}) = g(f(\underline{x}))$ is Fréchet differentiable at $\underline{x}$ with derivative $g'(f(\underline{x})) \circ f'(\underline{x})$.

Directional Derivatives

Directional derivatives differ from standard derivatives in that the perturbation vector is provided as a argument. For $f\colon \mathbb{R}^n \to \mathbb{R}^m$, this means that the directional derivative is not a linear transform but just a vector in $\mathbb{R}^m$.

Definition (Directional Derivative)

Let $f\colon X \to Y$ map vector space X to a Banach space $(Y, \|\cdot\|)$. Then, if it exists, the Gâteaux differential of f at $\underline{x}$ in direction $\underline{h}$ is given by

$$\delta f(\underline{x}; \underline{h}) \triangleq \lim_{t \to 0} \frac{f(\underline{x} + t\underline{h}) - f(\underline{x})}{t}.$$

Example

Consider $X = Y = \mathbb{R}^2$ and $f(\underline{x}) = (x_1 x_2, x_1 + x_2^2)$. For $\underline{x} = (1,1)$, $\underline{h} = (1,2)$:

$$\delta f(\underline{x}, \underline{h}) = \frac{d}{dt}((1+t)(1+2t), (1+t) + (1+2t)^2)\Big|_{t=0} = (3,5).$$

Properties of the Directional Derivative

Lemma

Let $Y = \mathbb{R}$ and suppose that $\delta f(\underline{x}; \underline{h})$ exists and is negative for some f, $\underline{x}$, and $\underline{h}$. Then, there exists $t_0 > 0$ such that, for all $t \in (0, t_0)$, one has

$$f(\underline{x} + t\underline{h}) < f(\underline{x}).$$

Proof in live session.

Definition

Let $f : X \to Y$ be a mapping from a vector space X to a Banach space $(Y, \|\cdot\|)$. Then, f is Gâteaux differentiable at $\underline{x}$ if the Gâteaux differential $\delta f(\underline{x}; \underline{h})$ exists for all $\underline{h} \in X$ and is a continuous linear function of $\underline{h}$.

On the Gradient in Hilbert Space

For $f\colon \mathbb{R}^n \to \mathbb{R}^m$ with $m = 1$, the Jacobian is related to the gradient

$$\nabla f(\underline{x}) \triangleq f'(\underline{x})^T = \left[\begin{array}{cccc} \frac{\partial f}{\partial x_1}(\underline{x}) & \frac{\partial f}{\partial x_2}(\underline{x}) & \cdots & \frac{\partial f}{\partial x_n}(\underline{x}) \end{array} \right]^T.$$

It is worth noting that the orientation of the gradient vector (i.e., row versus column vector) is sometimes defined differently. This is because derivatives can be understood as linear transforms and either orientation can be used to define the correct linear transform.

Example

Let X be a Hilbert space over $\mathbb{R}$ and $f\colon X \to \mathbb{R}$ be a real functional. If the Fréchet derivative $f'(\underline{x})$ exists, then it is a continuous linear functional on X. Thus, the Riesz representation theorem guarantees that there is a vector $\underline{u} \in X$ such that $f'(\underline{x})(\underline{h}) = \langle \underline{h}, \underline{u} \rangle$ for all $\underline{h} \in X$. This vector is called the gradient $\nabla f(\underline{x})$ and it follows that

$$f'(\underline{x})(\underline{h}) = \langle \underline{h}, \nabla f(\underline{x}) \rangle \text{ for all } \underline{h} \in X.$$

Gradient Descent

- Gradient descent subtracts the gradient $\nabla f(\underline{x})$ from an element of X
- For a Banach space, the derivative is a linear functional mapping X to $\mathbb{R}$!
- How can one add a linear mapping to X?
- In Hilbert space, the Riesz representation theorem states every linear functional is represented by the inner product with a fixed vector
- Thus, the gradient $\nabla f(\underline{x}) \in X$ is defined as the representative vector

Definition (Gradient Descent)

Let $f : X \to \mathbb{R}$ be a mapping from a Hilbert space X to the standard Banach space of real numbers. Starting from $\underline{x}_1 \in X$, gradient descent defines the sequence

$$\underline{x}_{n+1} = \underline{x}_n - \delta_n \nabla f(\underline{x}_n),$$

where δ_n is the step size for the n-th step.

Bounds for a Lipschitz Gradient

Lemma

Let $f : X \to \mathbb{R}$ map the Hilbert space X to the real numbers. If $\nabla f(\underline{x})$ exists and satisfies $\|\nabla f(\underline{y}) - \nabla f(\underline{x})\| \leq L\|\underline{y} - \underline{x}\|$, then

$$\left|f(\underline{y}) - f(\underline{x}) - \langle \underline{y} - \underline{x}, \nabla f(\underline{x})\rangle\right| \leq \frac{1}{2}L\|\underline{y} - \underline{x}\|^2.$$

Proof.

Let $\underline{h} = \underline{y} - \underline{x}$ and $\phi(t) = f(\underline{x} + t\underline{h})$. Then, $\phi'(t) = \langle \underline{h}, \nabla f(\underline{x} + t\underline{h})\rangle$ and

$$\begin{aligned}
\left|f(\underline{y}) - f(\underline{x}) - \langle \underline{h}, \nabla f(\underline{x})\rangle\right| &= \left|\int_0^1 (\phi'(t) - \phi'(0))\,\mathrm{d}t\right| \\
&= \left|\int_0^1 \langle \underline{h}, \nabla f(\underline{x} + t\underline{h}) - \nabla f(\underline{x})\rangle\,\mathrm{d}t\right| \\
&\leq \left|\int_0^1 \|\underline{h}\|\,\|\nabla f(\underline{x} + t\underline{h}) - \nabla f(\underline{x})\|\,\mathrm{d}t\right| \\
&\leq \int_0^1 \|\underline{h}\|L\|t\underline{h}\|\,\mathrm{d}t = \frac{1}{2}L\|\underline{h}\|^2.
\end{aligned}$$

Next Steps

- To continue studying after this video –
  - Try the required reading: Course Notes EF 5.1
  - Also, look at the gradient descent problem in Assignment 6