Variational Methods I

Luca Ballan

Half of the course…

	# variables	domain
▪ classic optimization	finite	dense $\mathbb{R}$
▪ discrete optimization	finite	discrete
▪ optimization on ∞-dimensional spaces	∞	dense $\mathbb{R}$
▪ optimization on manifolds	finite	dense $\mathbb{R}$ but highly constraint

Why Optimization?

- Why optimization is important for applied sciences?
- Where did the concept of "algorithm" go?

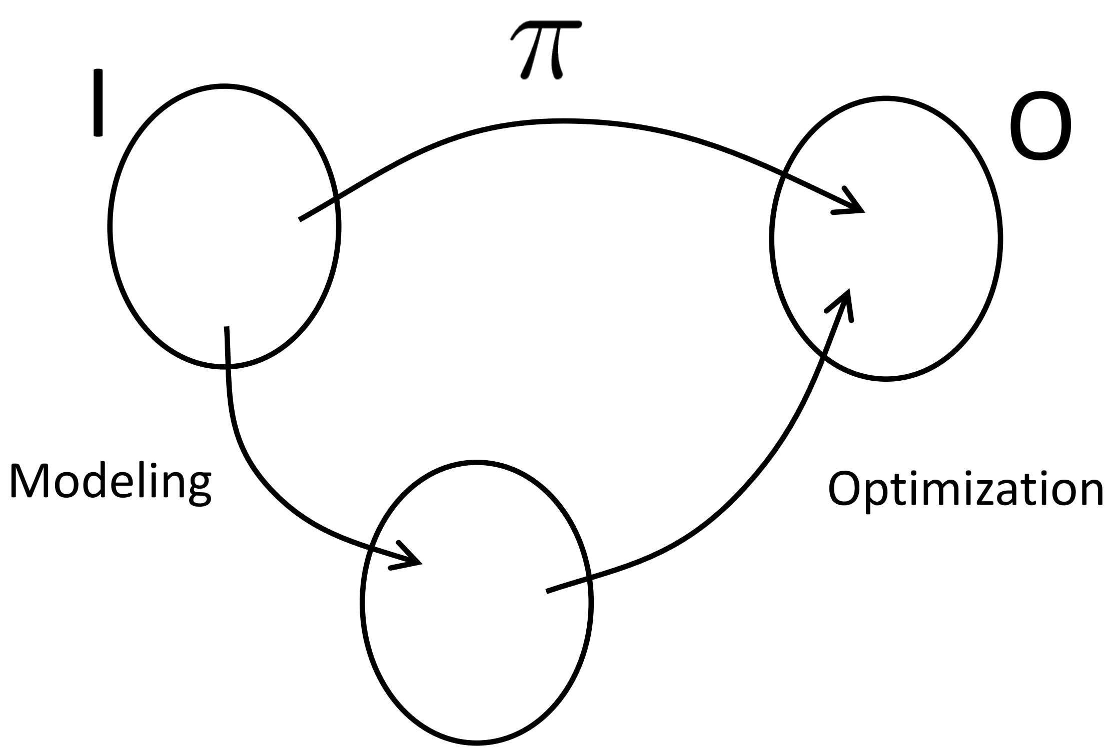
- Formulating a real-world problem as an optimization problem is just a **problem solving paradigm**. It is not the best method, it is just one of the possible ones.

- What is the formal definition of a problem?

Why Optimization?

- Why optimization is important for applied sciences?
- Where did the concept of “algorithm” go?

- Formulating a real-world problem as an optimization problem is just a **problem solving paradigm**. It is not the best method, it is just one of the possible ones.

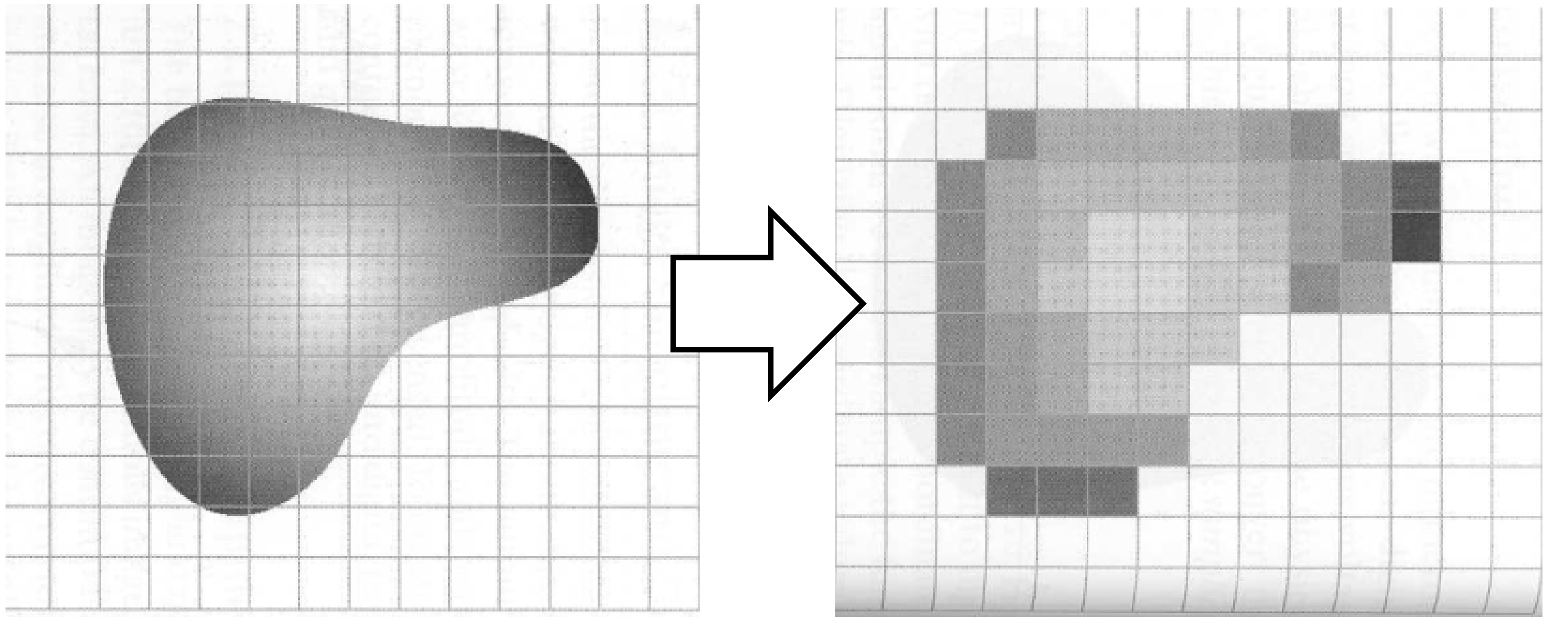


- Optimization problems are well studied.
- Reducing the problem to something that somebody else has already solved.
- If the optimization can be solved perfectly, can we say that we will be able to solve the problem perfectly?

Why Infinite-dimensional Spaces?

- Digital images and videos are discrete (numerical signal).
    - Discrete in color or brightness space (q**uantization**)
    - Discrete in the physical space (**space sampling**)
    - (videos) Discrete in time (**time sampling**)

- We are used to consider them discrete, because of the limitations of our processing units



"Infinite-Dimensional" Representation

Finite-Dimensional Representation

Why Infinite-dimensional Spaces?

- We are used to discrete representations, because:
  - digital objects are discrete, and their processing in a computer will ultimately require a discretization
  - No numerical approximations in modeling the transition from discrete to continuous
  - For various problems there exist efficient algorithms from discrete optimization

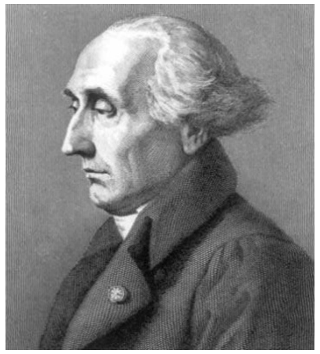
- But continuous representations have some advantages:
  - **The world is continuous** ergo **the images should be treated as continuous functions**
  - **There exists a huge mathematical theory for continuous functions** (functional analysis, differential geometry, partial differential equations, etc...)
  - Certain properties (e.g. rotational invariance) are easier to model in a continuous way
  - Finally, continuous models correspond to the limit of infinitely fine discretization

*

Calculus of Variations

- **Calculus of variations** is a classical topic in mathematics and in physics: in fact, in mechanics, it forms the basis for the **least action principle** which says that the motion of a particle lies on a stationary (minimum) point of a functional (the action).

- e.g., the Fermat's principle (the principle of least time): the path taken between two points by a ray of light is the path that can be traversed in the least time (the geodesic of the space)


Reference book: **Gelfand Fomin, Calculus of Variations, Prentice Hall, 1963**

Simple Optimization

- Given $\mathbb{R}^k$ (**Vector space** of dimension k over the **field** $\mathbb{R}$)
- Given $L \in C^1(\mathbb{R}^k, \mathbb{R})$ (**Loss functional** of class C^1 over $\mathbb{R}^k$)
- Find x^* such that

$$x^* = \arg\min_{x \in \mathbb{R}^k} L(x)$$

(unconstrained minimization problem)

- How do we solve for it?

There are many different ways to solve it!!

*

A possible solution

$$x^* = \arg\min_{x \in \mathbb{R}^k} L(x)$$

- Compute the gradient of L $\longrightarrow$ $\nabla L : \mathbb{R}^k \to \mathbb{R}^k$
- Compute the set of stationary points

$$S_L = \{x \in \mathbb{R}^k \mid \nabla L(x) = 0_k\}$$

- we know that, the point we are looking for, x^*, belongs to S_L **

- therefore, we can solve for this new problem

$$x^* = \arg\min_{x \in S_L} L(x)$$

it is guarantee that this coincides with the solution of the original problem $x^* = \arg\min_{x \in \mathbb{R}^k} L(x)$

If the cardinality of S_L is finite

⇩

- This optimization is relatively easy.
- It becomes an optimization over a discrete domain.
- One can use brute force, or some heuristics, or …)

A possible solution

- Given $$x^* = \arg\min_{x \in \mathbb{R}^k} L(x)$$
- Compute $\nabla L : \mathbb{R}^k \to \mathbb{R}^k$
- Compute $S_L = \{x \in \mathbb{R}^k \mid \nabla L(x) = 0_k\}$
- Solve $$x^* = \arg\min_{x \in S_L} L(x)$$

Real Scenario:

- maybe we cannot have analytical expression of ∇L
- maybe because we neither have an analytical expression of L

- To find S_L we need to solve for an equation

$$\nabla L(x) = 0_k$$

- which can be very difficult to solve!

- if S_L is not finite, this optimization is still not an easy problem
- we still have to discriminate between minimum, saddle, and maximum, local and global
- if the problem is convex and x^* is a minimum -> we know that x^* is the global minimum

Another solution

$$x^* = \arg\min_{x \in \mathbb{R}^k} L(x)$$

- That was what we are used to do in mathematical analysis
- But if we have a computer, we might prefer an “iterative” approach
- **Descent techniques**: “one starts from one point in $\mathbb{R}^k$ and you go straight down the hill until he/she hits a local minima”

*

- **Gradient descent** is a particular descent technique which always uses as descent direction the opposite of the gradient
(i.e. the direction in which the functional decreases most)

Another solution

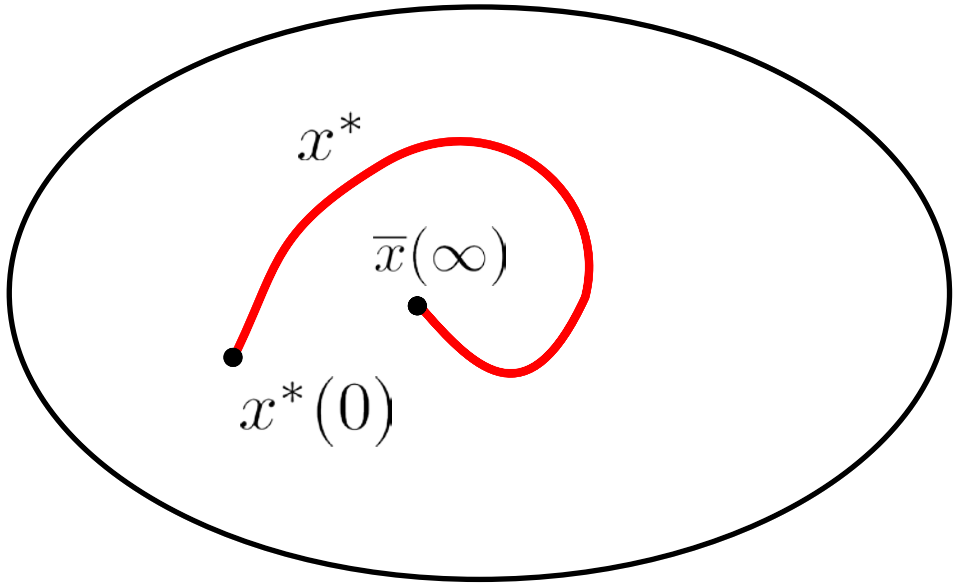
$$x^* = \arg\min_{x \in \mathbb{R}^k} L(x)$$

- This corresponds to adding a time dimension to our solution x^*

- i.e., x^* becomes a curve in our solution space $\mathbb{R}^k$

$$x^*(t) \in \mathbb{R}^k$$

- which start from an initial solution $x^*(0) = x_0$
- and it evolves according to a PDE

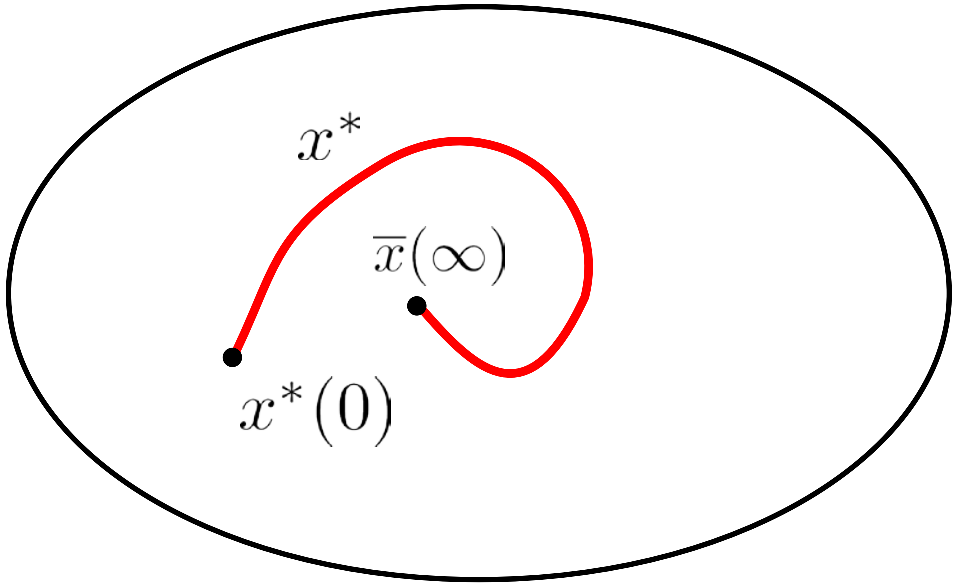


Another solution

$$x^* = \arg\min_{x \in \mathbb{R}^k} L(x)$$

- This corresponds to adding a time dimension to our solution x^*
- i.e., x^* becomes a curve in our solution space $\mathbb{R}^k$

$$x^*(t) \in \mathbb{R}^k$$



- which start from an initial solution $x^*(0) = x_0$
- and it evolves according to a PDE

$$\begin{cases} x^*(0) = x_0 \\ \dfrac{\partial x^*}{\partial t}(t) = -\nabla L(x^*(t)) \end{cases} \qquad \text{(PDE)}$$

Dinamic System

$$x^* = \lim_{t \to \infty} x^*(t)$$

We hope that this dynamical system converges to the solution of our problem (in a finite time better)

A More Complex Optimization Problem

- Given $\mathbb{F}(\mathbb{R}^k, \mathbb{R}^m)$ =
  - the set of all the functions of type $\mathbb{R}^k \to \mathbb{R}^m$
  - **vector space** (infinite dimensional) over the **field** $\mathbb{R}$

- Given $L \in C^1(\mathbb{F}(\mathbb{R}^k, \mathbb{R}^m) , \mathbb{R})$ (**Loss functional** of class C^1 over $\mathbb{F}(\mathbb{R}^k, \mathbb{R}^m)$)

- Find u^* such that

$$u^* = \underset{u \in \mathbb{F}(\mathbb{R}^k, \mathbb{R}^m)}{\arg\min} \; L(u)$$

(**unconstrained minimization problem** with functions as domain)

- How do we solve for it?

A More Complex Optimization Problem

$$L \in C^1(\mathbb{F}(\mathbb{R}^k, \mathbb{R}^m) \ , \mathbb{R})$$

- How does the gradient of a functional of functions look like?
- Can it look like

$$\nabla L = \left(\frac{\partial L}{\partial x_1}, \ldots, \frac{\partial L}{\partial x_k} \right) \ \textbf{?}$$

- What are $(x_1, \ldots, x_k)$ in this case?

- How does it look like the derivative? $\dfrac{\partial L}{\partial ?}$

- what do I need to place here?

- a lot of confusion!!!

Intuitively

$$L : \mathbb{R}^k \to \mathbb{R} \quad \Rightarrow \quad \nabla L = \left(\frac{\partial L}{\partial x_1}, \ldots, \frac{\partial L}{\partial x_k} \right)$$

the domain of optimization has dimension k

the gradient length is k

$$L : \mathbb{F}(\mathbb{R}^k, \mathbb{R}^m) \to \mathbb{R} \quad \Rightarrow \quad \nabla L$$

the domain of optimization is infinite dimensional

the gradient length is infinite,
so maybe, the gradient should be
a function

Calculus of Variations

- Calculus of variation extends the concept of gradient and derivative to all the functional defined on a generic topological vector space (either finite or infinite dimensional)

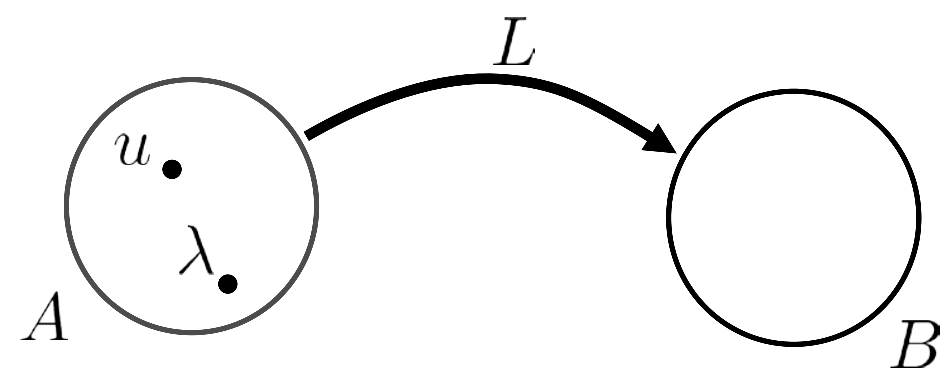
- We start describing this topic with the definition of the concept of **Directional Derivative**

Directional Derivative (of a function)

- Given $A,\ B$

Vector spaces over a same **field** (e.g. $\mathbb{R}$) with **topology**

- Given $L : A \to B$
- and $u, \lambda \in A$



- There exists an object

$$\frac{\partial L}{\partial \lambda}(u) = \lim_{\epsilon \to 0} \frac{L(u + \epsilon \cdot \lambda) - L(u)}{\epsilon}$$

$\lambda, u \in A$; $\lim_{\epsilon \to 0} \frac{L(u + \epsilon \cdot \lambda) - L(u)}{\epsilon} \in B$

Directional Derivative of L with direction λ evaluated in u

- The directional derivate is a function

$$\frac{\partial L}{\partial \lambda} : A \to B$$

*

Gradient (of a functional)

- If $B = \mathbb{R}$ (vector space with topology)
- If A is a general topological vector space with an **inner product**

$$\langle \cdot, \cdot \rangle_A : A \times A \to \mathbb{R}$$

- Given $L : A \to \mathbb{R}$ (**functional**), there might exists an object

$$\nabla L : A \to A$$

Gradient of L

informally, indicating the "direction" of maximal increase of L

- For each point $u \in A$, the gradient of L is formally defined as the unique element of A such that

$$\underbrace{\frac{\partial L}{\partial \lambda}(u)}_{\in \mathbb{R}} = \underbrace{\langle \underbrace{\nabla L(u)}_{\in A}, \underbrace{\lambda}_{\in A} \rangle_A}_{\in \mathbb{R}}$$

Properties

$$\frac{\partial L}{\partial \lambda}(u) = \langle \nabla L(u), \lambda \rangle_A$$

- If we restrict to all the directions with norm 1, the directional derivative has its maximum where the direction of λ is parallel to the gradient $\nabla L(u)$

$$\|\lambda\|_A = \sqrt{\langle \lambda, \lambda \rangle_A} = 1$$

Norm inherited by the inner product

- **Class C1**: If the gradient $\nabla L(u)$ exists and it is continuous for all $u \in A$, then L is said to be of class $C^1(A, \mathbb{R})$

- If $L \in C^1(A, \mathbb{R})$, $\frac{\partial L}{\partial \lambda}(u)$ is linear in both L and λ

Stationary points of L

- If $L \in C^1(A, \mathbb{R})$, the set of stationary point is defined as

$$S_L = \{u \in A \mid \nabla L(u) = 0_A\}$$

- L is locally flat
- there is no direction of maximum variation

*

$$= \left\{u \in A \mid \frac{\partial L}{\partial \lambda}(u) = 0_{\mathbb{R}}, \forall \lambda \in A\right\}$$

- if u is a minimum for L then any small variation of u (along any direction λ) would cause no effect on the value returned by L

*

Summary

- Given A a topological vector space with an inner product
- and given a functional of type $L : A \to \mathbb{R}$ sufficiently regular, i.e. $\in C^1(A,\mathbb{R})$

- it is defined an object called **directional derivative**

$$\frac{\partial L}{\partial \lambda} : A \to \mathbb{R}$$

(which, for every couple of elements in A, returns a value in $\mathbb{R}$)

- it is defined an object called **gradient**

$$\nabla L : A \to A$$

(which, for every elements of A, returns another element of A)

- it is defined the set of **stationary points** as

$$S_L = \{u \in A \mid \nabla L(u) = 0_A\}$$

$$= \left\{u \in A \mid \frac{\partial L}{\partial \lambda}(u) = 0_{\mathbb{R}}, \forall \lambda \in A\right\}$$

Let's pick $A = \mathbb{F}(\mathbb{R}^k, \mathbb{R}^m)$

- $A = \mathbb{F}(\mathbb{R}^k, \mathbb{R}^m)$ = the set of all the functions of type $\mathbb{R}^k \to \mathbb{R}^m$
- $(\mathbb{F}(\mathbb{R}^k, \mathbb{R}^m), +, \cdot_e)$ is a **vector space** over the field $\mathbb{R}$ (infinite dimensional)
- It admits an **inner product**, defined as

$$\langle f, g \rangle = \int_{\mathbb{R}^k} \langle f(x), g(x) \rangle_{\mathbb{R}^m} \, dx$$

- Therefore, it admits a **norm** $||f|| = \sqrt{\langle f, f \rangle}$ (inherited from the inner product)
- and it admits a **metric** $d(f, g) = ||f - g||$ (inherited from the norm)
- Therefore, it is a **topological vector space** ⇒ everything defined before should be defined also for this space!

PS: this space is called the **Lebesgue space of order 2** (L^2). It has the structure of an **Hilbert space** and it is a very important for the theory of the Fourier Transform and the theory of probability.

The functional $L : A = \mathbb{F}(\mathbb{R}^k, \mathbb{R}^m) \to \mathbb{R}$

- Given

$$L : \mathbb{F}(\mathbb{R}^k, \mathbb{R}^m) \to \mathbb{R}$$

- the **directional derivative** is defined (as before)

$$\frac{\partial L}{\partial \lambda}(u) = \lim_{\epsilon \to 0} \frac{L(u + \epsilon \cdot \lambda) - L(u)}{\epsilon}$$

(Gâteaux derivative)

where

$$u \in \mathbb{F}(\mathbb{R}^k, \mathbb{R}^m)$$
$$\lambda \in \mathbb{F}(\mathbb{R}^k, \mathbb{R}^m)$$

*

- the **gradient** is defined (as before)

$$\nabla L : \mathbb{F}(\mathbb{R}^k, \mathbb{R}^m) \to \mathbb{F}(\mathbb{R}^k, \mathbb{R}^m)$$

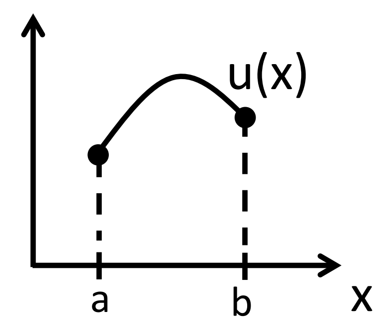
*

$$\frac{\partial L}{\partial \lambda}(u) = \langle \nabla L(u), \lambda \rangle$$

- **The gradient is unlikely to exist**

A simple Loss Functional

- Given $L : C^2([a,b],\mathbb{R}) \to \mathbb{R}$



$$L(u) = \int_a^b \psi(x, u(x), \dot{u}(x))dx$$

$$\frac{\partial L}{\partial \lambda}(u) = \int_a^b \left(\frac{\partial \psi}{\partial u}(x, u(x), \dot{u}(x)) - \frac{\partial}{\partial x}\frac{\partial \psi}{\partial \dot{u}}(x, u(x), \dot{u}(x)) \right) \lambda(x)dx + \left[\frac{\partial \psi}{\partial \dot{u}}(x, u(x), \dot{u}(x))\lambda(x) \right]_a^b$$

$$\nabla L(u)(x) = \frac{\partial \psi}{\partial u}(x, u(x), \dot{u}(x)) - \frac{\partial}{\partial x}\frac{\partial \psi}{\partial \dot{u}}(x, u(x), \dot{u}(x))$$