Optimization Problems

Optimización.

Gustavo A. Bula

Universidad Nacional de Colombia

February 19, 2024

Table of contents

Mathematical Optimization

(mathematical) optimization problem

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq b_i, \quad i = 1, \ldots, m \end{array}$$

- $x = (x_1, \ldots, x_n)$: optimization variables
- $f_0 : \mathbf{R}^n \to \mathbf{R}$: objective function
- $f_i : \mathbf{R}^n \to \mathbf{R}$, $i = 1, \ldots, m$: constraint functions

optimal solution $x^\star$ has smallest value of f_0 among all vectors that satisfy the constraints

Solving Optimizations Problems

general optimization problem

- very difficult to solve
- methods involve some compromise, *e.g.*, very long computation time, or not always finding the solution

exceptions: certain problem classes can be solved efficiently and reliably

- least-squares problems
- linear programming problems
- convex optimization problems

Convex Optimization Problem

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq b_i, \quad i = 1, \ldots, m \end{array}$$

- objective and constraint functions are convex:

$$f_i(\alpha x + \beta y) \leq \alpha f_i(x) + \beta f_i(y)$$

  if $\alpha + \beta = 1$, $\alpha \geq 0$, $\beta \geq 0$

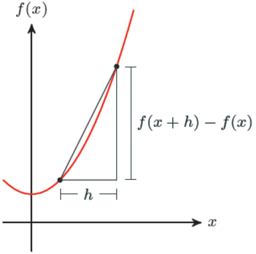
- includes least-squares problems and linear programs as special cases

Derivative in One Dimension

Definition

Let X be an open subset of $\mathbb{R}$ and let $f : X \to \mathbb{R}$. Then f is differentiable at $x \in X$ with derivative $\frac{d}{dx}f(x)$ if the following limit exists:

$$\frac{d}{dx}f(x) = \lim_{h\to 0}\frac{1}{h}\left(f(x+h) - f(x)\right)$$



Example for $f(x) = 2x + 3x^2$

$$\begin{aligned}
\tfrac{d}{dx}f(x) &= \lim_{h\to 0}\tfrac{1}{h}\left(2(x+h) + 3(x+h)^2 - 2x - 3x^2\right) \\
&= \lim_{h\to 0}\tfrac{1}{h}\left(2(x+h) + 3(x^2 + 2xh + h^2) - 2x - 3x^2\right) \\
&= \lim_{h\to 0}\tfrac{1}{h}\left(2h + 3(2xh + h^2)\right) \\
&= \lim_{h\to 0}\left(2 + 6x + 3h\right) = 2 + 6x
\end{aligned}$$

Gradient and Partial Derivative

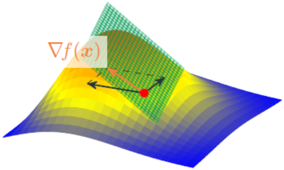
A function of several variables can be written as $f(x_1, x_2)$, etc. Often times, we abbreviate multiple arguments in a single vector as $f(\boldsymbol{x})$.

Let a function $f : \mathbb{R}^n \to \mathbb{R}$. The ***gradient of f*** is the ***column vector of partial derivatives***

$$\nabla f(\boldsymbol{x}) := \begin{pmatrix} \frac{\partial f(\boldsymbol{x})}{\partial x_1} \\ \vdots \\ \frac{\partial f(\boldsymbol{x})}{\partial x_n} \end{pmatrix}$$

Suppose now a function $g(\boldsymbol{x}, \boldsymbol{y})$ with signature $g : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$. Its ***derivative with respect to just $\boldsymbol{x}$*** is written as $\nabla_{\boldsymbol{x}} g(\boldsymbol{x}, \boldsymbol{y})$.

Gradient and Tangent Plane / 1st Degree Taylor Expansion



$$\tau^1_{\boldsymbol{x}}(\boldsymbol{y}) = f(\boldsymbol{x}) + (\boldsymbol{y} - \boldsymbol{x})^\top \nabla f(\boldsymbol{x})$$

The Hessian Matrix

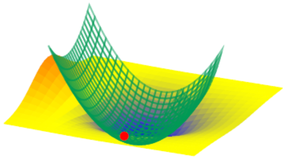
Let $f : \mathbb{R}^n \to \mathbb{R}$ twice differentiable. Its ***second (partial) derivatives*** make up the ***Hessian Matrix*** $\nabla^2 f(\boldsymbol{x})$:

$$\nabla^2 f(\boldsymbol{x}) := \begin{pmatrix} \frac{\partial^2 f(\boldsymbol{x})}{\partial x_1 \partial x_1} & \cdots & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f(\boldsymbol{x})}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_n \partial x_n} \end{pmatrix}$$

- The ***order of differentiation does not matter*** if the function has continuous second (higher-order) partial derivatives (Schwarz's Theorem)
- Then the ***Hessian is symmetric***

$$\nabla^2 f(\boldsymbol{x}) = [\nabla^2 f(\boldsymbol{x})]^\top$$

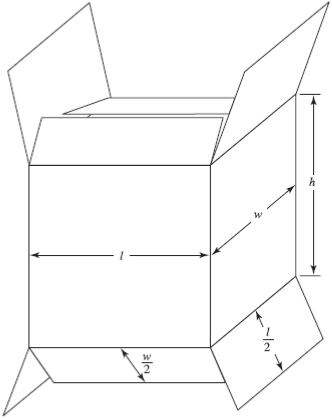
2nd Degree Taylor Expansion



$$\tau_{\boldsymbol{x}}^2(\boldsymbol{y}) = f(\boldsymbol{x}) + (\boldsymbol{y}-\boldsymbol{x})^\top \nabla f(\boldsymbol{x}) + \frac{1}{2}(\boldsymbol{y}-\boldsymbol{x})^\top [\nabla^2 f(\boldsymbol{x})] (\boldsymbol{y}-\boldsymbol{x})$$

Example

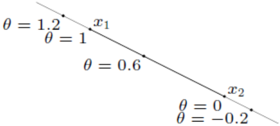
A cardboard box (see Figure) is to be designed to have a volume of $1.67 ft^3$. Determine the optimal values of l, w, and h so as to minimize the amount of cardboard material. (Hint: use the volume constraint equation to eliminate one of the variables.)



Affine Set

line through x_1, x_2: all points

$$x = \theta x_1 + (1-\theta)x_2 \qquad (\theta \in \mathbf{R})$$



affine set: contains the line through any two distinct points in the set

example: solution set of linear equations $\{x \mid Ax = b\}$

(conversely, every affine set can be expressed as solution set of system of linear equations)

Convex Set

line segment between x_1 and x_2: all points

$$x = \theta x_1 + (1 - \theta) x_2$$

with $0 \leq \theta \leq 1$

convex set: contains line segment between any two points in the set

$$x_1, x_2 \in C, \quad 0 \leq \theta \leq 1 \quad \Longrightarrow \quad \theta x_1 + (1 - \theta) x_2 \in C$$

examples (one convex, two nonconvex sets)

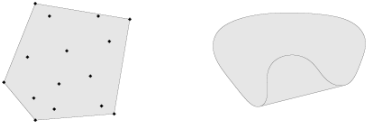
Convex Combination and Convex Hull

convex combination of $x_1, \ldots, x_k$: any point x of the form

$$x = \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_k x_k$$

with $\theta_1 + \cdots + \theta_k = 1$, $\theta_i \geq 0$

convex hull $\mathbf{conv}\, S$: set of all convex combinations of points in S



Convexity of Functions

$f : \mathbf{R}^n \to \mathbf{R}$ is convex if $\mathbf{dom}\, f$ is a convex set and

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y)$$

for all $x, y \in \mathbf{dom}\, f$, $0 \leq \theta \leq 1$

- f is concave if $-f$ is convex
- f is strictly convex if $\mathbf{dom}\, f$ is convex and

$$f(\theta x + (1-\theta)y) < \theta f(x) + (1-\theta) f(y)$$

for $x, y \in \mathbf{dom}\, f$, $x \neq y$, $0 < \theta < 1$

First Order Conditions

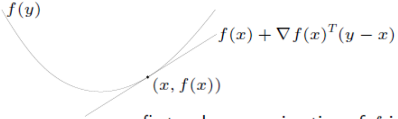
f is **differentiable** if $\mathbf{dom}\, f$ is open and the gradient

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

exists at each $x \in \mathbf{dom}\, f$

1st-order condition: differentiable f with convex domain is convex iff

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) \quad \text{for all } x, y \in \mathbf{dom}\, f$$



first-order approximation of f is global underestimator

Second Order Conditions

f is **twice differentiable** if $\mathbf{dom}\, f$ is open and the Hessian $\nabla^2 f(x) \in \mathbf{S}^n$,

$$\nabla^2 f(x)_{ij} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}, \quad i, j = 1, \ldots, n,$$

exists at each $x \in \mathbf{dom}\, f$

2nd-order conditions: for twice differentiable f with convex domain

- f is convex if and only if

$$\nabla^2 f(x) \succeq 0 \quad \text{for all } x \in \mathbf{dom}\, f$$

- if $\nabla^2 f(x) \succ 0$ for all $x \in \mathbf{dom}\, f$, then f is strictly convex

Examples

quadratic function: $f(x) = (1/2)x^T P x + q^T x + r$ (with $P \in \mathbf{S}^n$)

$$\nabla f(x) = Px + q, \qquad \nabla^2 f(x) = P$$

convex if $P \succeq 0$

least-squares objective: $f(x) = \|Ax - b\|_2^2$

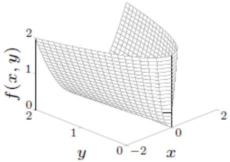
$$\nabla f(x) = 2A^T(Ax - b), \qquad \nabla^2 f(x) = 2A^T A$$

convex (for any A)

quadratic-over-linear: $f(x, y) = x^2/y$

$$\nabla^2 f(x, y) = \frac{2}{y^3} \begin{bmatrix} y \\ -x \end{bmatrix} \begin{bmatrix} y \\ -x \end{bmatrix}^T \succeq 0$$

convex for $y > 0$



Epigraph and Sublevel Set

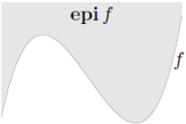
α-**sublevel set** of $f : \mathbf{R}^n \to \mathbf{R}$:

$$C_\alpha = \{x \in \mathbf{dom}\, f \mid f(x) \le \alpha\}$$

sublevel sets of convex functions are convex (converse is false)

epigraph of $f : \mathbf{R}^n \to \mathbf{R}$:

$$\mathbf{epi}\, f = \{(x, t) \in \mathbf{R}^{n+1} \mid x \in \mathbf{dom}\, f,\ f(x) \le t\}$$



f is convex if and only if $\mathbf{epi}\, f$ is a convex set

Jensen's Inequality

basic inequality: if f is convex, then for $0 \leq \theta \leq 1$,

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y)$$

extension: if f is convex, then

$$f(\mathbf{E}\, z) \leq \mathbf{E}\, f(z)$$

for any random variable z

basic inequality is special case with discrete distribution

$$\mathbf{prob}(z = x) = \theta, \qquad \mathbf{prob}(z = y) = 1 - \theta$$

Convexity of Functions

practical methods for establishing convexity of a function

1. verify definition (often simplified by restricting to a line)
2. for twice differentiable functions, show $\nabla^2 f(x) \succeq 0$
3. show that f is obtained from simple convex functions by operations that preserve convexity
   - nonnegative weighted sum
   - composition with affine function
   - pointwise maximum and supremum
   - composition
   - minimization
   - perspective

Direction of Steepest Descent

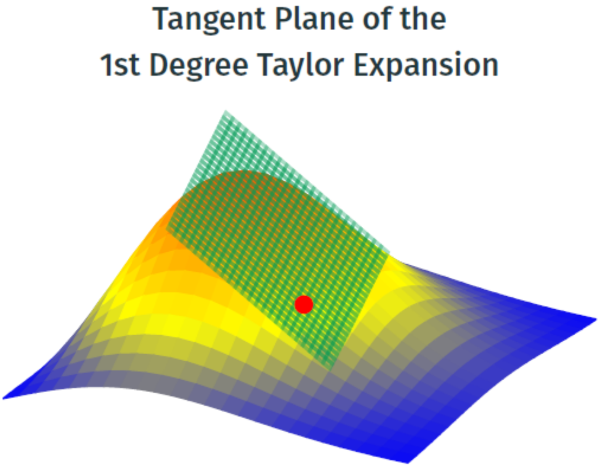
- The 1st degree Taylor Expansion linearizes the function $f : \mathbb{R}^n \to \mathbb{R}$ at the selected point x:

$$\tau_x^1(y) = f(x) + (y - x)^T \nabla f(x)$$

- In which direction can we step downwards as far as possible on the tangent plane of the 1st Degree Taylor Expansion (for a step of length 1)?

$$-\frac{\nabla f(x)}{\| \nabla f(x) \|_2} = \underset{h \in \mathbb{R}^n, \|h\|_2 = 1}{\arg\min} \; h^T \nabla f(x)$$

- The steepest descent is in the direction of the negative gradient $-\nabla f(x)$



$$\tau_x^1(y) = f(x) + (y - x)^T \nabla f(x)$$

Proof:The Cauchy-Schwarz Inequality

The Cauchy-Schwarz Inequality tells us that

$$|h^T \nabla f(x)| = \| h^T \nabla f(x) \|_2 \leq \| h \|_2 \| \nabla f(x) \|_2$$

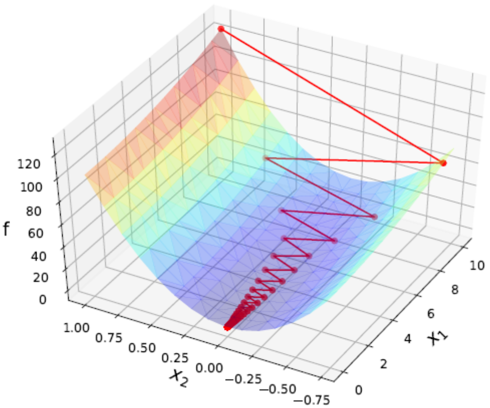
The right-hand-side is fixed since $\| h \|_2 = 1$. The left-hand-side is maximized when equality is achieved. Equality is achieved for

$$h^* = \frac{\nabla f(x)}{\| \nabla f(x) \|_2}$$

Take $-h^*$ to minimize.

Gradient Descent

- Iterative method starting at an initial point $x^{(0)}$
- Step to the next point $x^{(k+1)}$ in the direction of the negative gradient
$$x^{(k+1)} = x^{(k)} - \nabla f(x^{(k)})$$
- Repeat until $\| \nabla f(x^{(k)}) \| < \epsilon$ for a chosen ϵ
- But: No convergence is guaranteed. For convergence, an additional line search is required



Gradient Descent for

$$f(\boldsymbol{x}) = \frac{1}{2}(x_1)^2 + 5(x_2)^2$$

Line Search

- Take the descent step direction $d = -\nabla f(x)$
- Select the step length α as $\min_{\alpha \geq 0} f(x + \alpha d)$
- In practice, α is selected with heuristics

Backtracking Line Search [Armijo66]

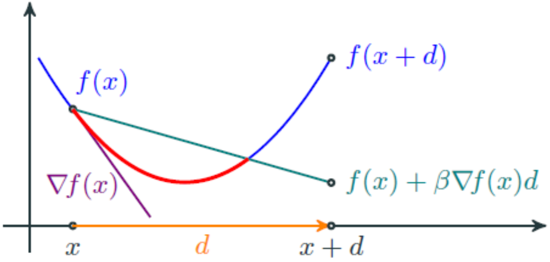
- Heuristic line search (not exact), but simple and efficient.
- Start with a step direction d
- Iteratively reduce the step length factor α. A common rule is $\alpha \leftarrow p\alpha$ with $p = 0.5$.
- Stop when minimum "descent steepness" is reached (Armijo condition)

  $$f(x + \alpha d) \leq f(x) + \beta[\nabla f(x)]^T(\alpha d)$$

  choose $\beta \in (0, 1)$. A common choice is $\beta = 10^{-4}$
  $\beta \approx 0$: Any descent is valid (must be only minimallly downwards) $\rightarrow$ Big nearly horizontal steps allowed
  $\beta \approx 1$: The descent needs to be nearly as step as $-\nabla f(x) \rightarrow$ Steps sizes can become very small

Backtracking Line Search [Armijo66]

1. procedure LINESEARCH (f, x, d, p, β)
2. $\quad \alpha \leftarrow 1$
3. $\quad$ while $f(x + \alpha d) > f(x) + \beta[\nabla f(x)]^T(\alpha d)$do
4. $\qquad \alpha \leftarrow p\alpha$
5. $\quad$ end while
6. $\quad$ return α
7. end procedure


$f(x+d)$
$f(x)$
$\nabla f(x)$
$f(x) + \beta \nabla f(x) d$
x
d
$x+d$

Steepest Descent Algorithm

1. Select a starting design point $\mathbf{x}_0$ and parameters ε_G, ε_A, ε_R. Set iteration index $k = 0$.
2. Compute $\nabla f(\mathbf{x}_k)$. Stop if $||\nabla f(\mathbf{x}_k)|| \leq \varepsilon_G$. Otherwise, define a normalized direction vector $d_k = -\nabla f(\mathbf{x}_k)/||\nabla f(\mathbf{x}_k)||$
3. Obtain α_k from exact or approximate line search techniques $f(\alpha) = f(\mathbf{x}_k + \alpha d_k), \alpha > 0$.Update $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha d_k$
4. Evaluate $f(\mathbf{x}_{k+1})$. Stop if

$$|f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k)| \leq \varepsilon_A + \varepsilon_R |f(\mathbf{x}_k)|$$

   is satisfied for two successive iterations. Otherwise set $k = k + 1$, $\mathbf{x}_k = \mathbf{x}_{k+1}$ and go to step 2.

Model Fitting: Empirical Risk Minimization

1. Select a model class for the independent variable prediction: These are parametric models. They define a fixed number of model parameters.
2. Fit the model parameters.
   - Finding "good" model parameters is called model fitting.
   - Empirical risk minimization is commonly used for model fitting.

Model Fitting: Empirical Risk Minimization

- **Given a dataset** with the data observations $\{(y_1, X_1), (y_2, X_2), \ldots, (y_n, X_n)\}$
- **A model class is assumed**. $y_i = X_i\beta$
- **A lost function is selected**. A quadratic prediction error for each observation (y_i, X_i) is used. The loss function returns the overall prediction error for a combination of model parameters $\hat{y}_i = X_i\theta$

$$lost\ function(\theta) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- Minimize the loss function to find the "best" model parameter:

$$\theta^* = \arg\min_{\theta} lost\ function(\theta)$$

Least-squares

$$minimize \parallel Ax - b \parallel_2^2 = \sum i = 1^k (a_i^T x - b_i)^2$$

solving least-squares problems

- analytical solution: $x^* = (A^T A)^{-1} A^T b$
- reliable and efficient algorithms and software
- computation time proportional to $n^2 k (A \in \mathbb{R}^{k \times n}, k \geq n)$; less if structured
- a mature technology

using least-squares

- least-squares problems are easy to recognize
- a few standard techniques increase flexibility (e.g., including weights, adding regularization terms)

Least-squares

least-squares objective: $f(x) = minimize \parallel Ax - b \parallel_2^2$

$$\nabla f(x) = 2A^T(Ax - b)$$

$$\nabla^2 f(x) = 2A^T A$$

convex (for any A)