IOE 519: Introduction to Nonlinear Programming

October 6, 2024

1 Calculus and analysis review

Almost all books on nonlinear programming have an appendix reviewing the relevant notions.

Most of these should be familiar to you from a course in analysis. Most of material in this course is based in some form on these concepts, therefore, to succeed in this course you should be not just familiar, but comfortable working with these concepts.

A few words on symbols and wording in mathematical statements

Three symbols used frequently in these notes are

- $\forall$ — read as "for every," or "for any," or "for all."
  - E.g., the expression "for any x in $\mathbf{R}^n$ such that x is nonnegative..." is abbreviated as "$\forall x \in \mathbf{R}^n$ such that $x \geq 0$..."
- $\exists$ — read as "there exists a..." or, less formally, "one can find a..."
  - E.g., the expression "for any $x \in S$, there exists (or one can find) a value $\delta > 0$ such that $\|x\| \leq \delta$" is abbreviated as "$\forall x \in S$, $\exists \delta > 0$ such that $\|x\| \leq \delta$."
  - Note that the above is meant to indicate that the value of δ can depend on the specific x; however, if the statement read "one can find (or there exists) a value $\delta > 0$ such that $\|x\| \leq \delta$ for any $x \in S$" (abbreviated as "$\exists \delta > 0$: $\|x\| \leq \delta$ $\forall x \in S$"), that would mean that *the same* value of δ has to "work" for every $x \in S$.
- $\Leftrightarrow$ — read as "if and only if."

Vectors and Norms

- $\mathbf{R}^n$: set of n-dimensional real vectors $(x_1, \ldots, x_n)^T$ ("x^T" — transpose)
- Definition: *norm* $\|\cdot\|$ on $\mathbf{R}^n$: a mapping of $\mathbf{R}^n$ into $\mathbf{R}$ such that:
  1. $\|x\| \geq 0$ $\forall x \in \mathbf{R}^n$; $\|x\| = 0 \Leftrightarrow x = 0$.
  2. $\|cx\| = |c| \cdot \|x\|$ $\forall c \in \mathbf{R}, x \in \mathbf{R}^n$.
  3. $\|x + y\| \leq \|x\| + \|y\|$ $\forall x, y \in \mathbf{R}^n$.
- Euclidean norm: $\|\cdot\|_2$: $\|x\|_2 = \sqrt{x^T x} = \left(\sum_{i=1}^n x_i^2\right)^{1/2}$.
  - Schwartz inequality: $|x^T y| \leq \|x\|_2 \cdot \|y\|_2$; equality holds if and only if $x = \alpha y$.

- Other norm examples: $\|x\|_1 = \sum_{i=1}^n |x_i|$; $\|x\|_\infty = \max_{i=1,\dots,n} |x_i|$, etc.
- All norms in $\mathbf{R}^n$ are *equivalent*, i.e., for any $\|\cdot\|^1$ and $\|\cdot\|^2$ $\exists \alpha_1, \alpha_2 > 0$ s.t. $\alpha_1\|x\|^1 \le \|x\|^2 \le \alpha_2\|x\|^1$ $\forall x \in \mathbf{R}^n$.
- ϵ*–Neighborhood*: $N_\epsilon(x) = B(x,\epsilon) = \{y \ : \ \|y - x\| \le \epsilon\}$ (sometimes — strict inequality).

Sequences and Limits.

Sequences in R

- Notation: a *sequence*: $\{x_k \ : \ k = 1, 2, \dots\} \subset \mathbf{R}$, $\{x_k\}$ for short.
- Definition: $\{x_k\} \subset \mathbf{R}$ *converges* to $x \in \mathbf{R}$ ($x_k \to x$, $\lim_{k\to\infty} x_k = x$) if

$$\forall \epsilon > 0 \ \exists K : \ |x_k - x| \le \epsilon \text{ (equiv. } x_k \in B(x,\epsilon)) \ \forall k \ge K.$$

  $x_k \to \infty$ $(-\infty)$ if

$$\forall A \ \exists K : \ x_k \ge A \ (x_k \le A) \ \forall k \ge K.$$

- Definition: $\{x_k\}$ is *bounded above (below)*: $\exists A : x_k \le A$ $(x_k \ge A)$ $\forall k$.
- Definition: $\{x_k\}$ is *bounded*: $\{|x_k|\}$ is bounded; equivalently, $\{x_k\}$ bounded above and below.
- Definition: $\{x_k\}$ is *nonincreasing (nondecreasing)*: $x_{k+1} \le x_k$ $(x_{k+1} \ge x_k)$ $\forall k$;
  monotone: nondecreasing or nonincreasing.
- Proposition: Every monotone sequence in $\mathbf{R}$ has a limit (possibly infinite). If it is also bounded, the limit is finite.

Sequences in $\mathbf{R}^n$

- Definition: $\{x_k\} \subset \mathbf{R}^n$ *converges to* $x \in \boldsymbol{R}^n$ *(is bounded)* if $\{x_k^i\}$ (the sequence of ith coordinates of x_k's) converges to the x^i (is bounded) $\forall i$.
- Propositions:
  - $x_k \to x \Leftrightarrow \|x_k - x\| \to 0$
  - $\{x_k\}$ is bounded$\Leftrightarrow \{\|x_k\|\}$ is bounded
- Note: $\|x_n\| \to \|x\|$ does not imply that $x_n \to x$!! (Unless $x = 0$).

Limit Points

- Definition: x is a limit point of $\{x_k\}$ if there exists an infinite subsequence of $\{x_k\}$ that converges to x.
- Definition: x is a limit point of $A \subseteq \mathbf{R}^n$ if there exists an infinite sequence $\{x_k\} \subset A$ that converges to x.
- To see the difference between limits and limit points, consider the sequence

$$\{(1,0),(0,1),(-1,0),(0,-1),(1,0),(0,1),(-1,0),(0,-1),\dots\}$$

- Proposition: let $\{x_k\} \subset \mathbf{R}^n$
  - If $\{x_k\}$ is bounded, then $\{x_k\}$ converges if and only if it has a unique limit point

  - If $\{x_k\}$ is bounded, it has at least one limit point

Infimum and Supremum

- Let $A \subset \mathbf{R}$.
  Supremum of A ($\sup A$): smallest $y \in \mathbf{R}$ that satisfies $x \leq y \ \forall x \in A$.
  Infimum of A ($\inf A$): largest $y \in \mathbf{R}$ that satisfies $x \geq y \ \forall x \in A$.
- Not the same as *maximum* and *minimum*, which are the largest and smallest elements of the set A! Consider, for example, $A = (0, 1)$.

Closed and Open Sets

- Definition: a set $A \subseteq \mathbf{R}^n$ is closed if it contains all its limit points. In other words, for any sequence $\{x_k\} \subset A$ that has a limit x, $x \in A$.
- Definition: a set $A \subseteq \mathbf{R}^n$ is open if its complement, $\mathbf{R}^n \backslash A$, is closed
- Definition: a point $x \in A$ is *interior* if there is a neighborhood of x contained in A
- Proposition
  1. A set is open $\Leftrightarrow$ All of its elements are interior points.
  2. Union of *finitely many* closed sets is closed.
  3. Intersection of closed sets is closed.
  4. Union of open sets is open.
  5. Intersection of *finitely many* open sets is open.
  6. Every subspace of $\mathbf{R}^n$ is closed.
- Examples: neighborhoods of x:
  $\{y \ : \ \|y - x\| \leq \epsilon\}$ — closed
  $\{y \ : \ \|y - x\| < \epsilon\}$ — open
- Some sets are neither: $(0, 1]$.

Functions and Continuity

- $A \subseteq \mathbf{R}^m$, $f : A \to \mathbf{R}$ – a function.
- Definition: f is *continuous* at $\bar{x}$ if

$$\forall \epsilon > 0 \ \exists \delta > 0 : \ x \in A, \ \|x - \bar{x}\| < \delta \Rightarrow |f(x) - f(\bar{x})| < \epsilon.$$

- Proposition: f is continuous at $\bar{x} \Leftrightarrow$ for any $\{x_n\} \subset A : \ x_n \to \bar{x}$ we have $f(x_n) \to f(\bar{x})$. (In other words, $\lim f(x_n) = f(\lim x_n)$.)
- Proposition:
  - Sums, products and inverses of continuous functions are continuous (in the last case, provided the function is never zero).
  - Composition of two continuous functions is continuous.
  - Any vector norm is a continuous function.

Differentiation

Real-valued functions: Let $f : X \to \mathbf{R}$, where $X \subset \mathbf{R}^n$ is open.

- Definition: f is *differentiable* at $\bar{x} \in X$ if there exists a vector $\nabla f(\bar{x})$ (the *gradient* of f at $\bar{x}$) and a function $\alpha_{\bar{x}}(y) : X \to \mathbf{R}$ satisfying $\lim_{y\to 0} \alpha_{\bar{x}}(y) = 0$, such that for each $x \in X$

$$f(x) = f(\bar{x}) + \nabla f(\bar{x})^T(x - \bar{x}) + \|x - \bar{x}\|\alpha_{\bar{x}}(x - \bar{x}).$$

  f is *differentiable on* X if f is differentiable $\forall \bar{x} \in X$. The gradient vector is a vector of partial derivatives:

$$\nabla f(\bar{x}) = \left(\frac{\partial f(\bar{x})}{\partial x_1}, \ldots, \frac{\partial f(\bar{x})}{\partial x_n}\right)^T.$$

  The *directional derivative* of f at $\bar{x}$ in the direction d is

$$\lim_{\lambda\to 0} \frac{f(\bar{x} + \lambda d) - f(\bar{x})}{\lambda} = \nabla f(\bar{x})^T d$$

- Definition: the function f is *twice differentiable* at $\bar{x} \in X$ if there exists a vector $\nabla f(\bar{x})$ and an $n \times n$ symmetric matrix $H(\bar{x})$ (the *Hessian* of f at $\bar{x}$) such that for each $x \in X$

$$f(x) = f(\bar{x}) + \nabla f(\bar{x})^T(x - \bar{x}) + \frac{1}{2}(x - \bar{x})^T H(\bar{x})(x - \bar{x}) + \|x - \bar{x}\|^2\alpha_{\bar{x}}(x - \bar{x}),$$

  and $\lim_{y\to 0} \alpha_{\bar{x}}(y) = 0$. f is *twice differentiable on* X if f is twice differentiable $\forall \bar{x} \in X$. The Hessian, which we often denote by $H(x)$ for short, is a matrix of second partial derivatives:

$$[H(x)]_{ij} = \frac{\partial^2 f(\bar{x})}{\partial x_i \partial x_j},$$

  and for functions with continuous second derivatives, it will always be symmetric:

$$\frac{\partial^2 f(\bar{x})}{\partial x_i \partial x_j} = \frac{\partial^2 f(\bar{x})}{\partial x_j \partial x_i}$$

- Example:

$$f(x) = 3x_1^2x_2^3 + x_2^2x_3^3$$

$$\nabla f(x) = \begin{pmatrix} 6x_1x_2^3 \\ 9x_1^2x_2^2 + 2x_2x_3^3 \\ 3x_2^2x_3^2 \end{pmatrix}$$

$$H(x) = \begin{bmatrix} 6x_2^3 & 18x_1x_2^2 & 0 \\ 18x_1x_2^2 & 18x_1^2x_2 + 2x_3^3 & 6x_2x_3^2 \\ 0 & 6x_2x_3^2 & 6x_2^2x_3 \end{bmatrix}$$

- See additional handout to verify your understanding and derive the gradient and Hessian of linear and quadratic functions.

Vector-valued functions: Let $f : X \to \mathbf{R}^m$, where $X \subset \mathbf{R}^n$ is open.

- $$f(x) = f(x_1, \ldots, x_n) = \begin{pmatrix} f_1(x_1, \ldots, x_n) \\ f_2(x_1, \ldots, x_n) \\ \vdots \\ f_m(x_1, \ldots, x_n) \end{pmatrix},$$

  where each of the functions f_i is a real-valued function.

- Definition: the *Jacobian* of f at point $\bar{x}$ is the matrix whose jth row is the gradient of f_j at $\bar{x}$, transposed. More specifically, the Jacobian of f at $\bar{x}$ is defined as $\nabla f(\bar{x})^T$, where $\nabla f(\bar{x})$ is the matrix with entries:
  $$[\nabla f(\bar{x})]_{ij} = \frac{\partial f_j(\bar{x})}{\partial x_i}.$$
  Notice that the jth column of $\nabla f(\bar{x})$ is the gradient of f_j at $\bar{x}$ (what happens when $m = 1$?)

- Example:
  $$f(x) = \begin{pmatrix} \sin x_1 + \cos x_2 \\ e^{3x_1 + x_2^2} \\ 4x_1^3 + 7x_1x_2^2 \end{pmatrix}.$$
  Then
  $$\nabla f(x)^T = \begin{pmatrix} \cos x_1 & -\sin x_2 \\ 3e^{3x_1 + x_2^2} & 2x_2e^{3x_1 + x_2^2} \\ 12x_1^2 + 7x_2^2 & 14x_1x_2 \end{pmatrix}.$$

Other well-known results from calculus and analysis will be introduced throughout the course as needed.

2 Examples of nonlinear programming problems formulations

2.1 Forms and components of a mathematical programming problems

A *mathematical programming problem* or, simply, a *mathematical program* is a mathematical formulation of an optimization problem.

Unconstrained Problem:

$$\begin{array}{lll} \text{(P)} & \text{minimize}_x & f(x) \\ & \text{subject to} & x \in X, \end{array}$$

where $x = (x_1, \ldots, x_n)^T \in \mathbf{R}^n$, $f(x) : \mathbf{R}^n \to \mathbf{R}$, and X is an open set (usually, but not always, $X = \mathbf{R}^n$).

Constrained Problem:

$$\begin{array}{lll} \text{(P)} & \text{mininimize}_x & f(x) \\ & \text{subject to} & g_i(x) \leq 0 \quad i = 1, \ldots, m \\ & & h_i(x) = 0 \quad i = 1, \ldots, l \\ & & x \in X, \end{array}$$

where $g_1(x), \ldots, g_m(x), h_1(x), \ldots, h_l(x) : \mathbf{R}^n \to \mathbf{R}$.

Let $g(x) = (g_1(x), \ldots, g_m(x))^T : \mathbf{R}^n \to \mathbf{R}^m$, $h(x) = (h_1(x), \ldots, h_l(x))^T : \mathbf{R}^n \to \mathbf{R}^l$. Then (P) can be written as

$$\begin{array}{lll} \text{(P)} & \text{minimize}_x & f(x) \\ & \text{subject to} & g(x) \preceq 0 \\ & & h(x) = 0 \\ & & x \in X. \end{array} \tag{1}$$

Some terminology: Function $f(x)$ in (**??**) is the *objective function*. Restrictions "$h_i(x) = 0$" are referred to as *equality constraints*, while "$g_i(x) \leq 0$" are *inequality constraints*. Notice that we do not use constraints in the form "$g_i(x) < 0$"!

A point x is *feasible* for (P) if it satisfies all the constraints. (For an unconstrained problem, $x \in X$.) The set of all feasible points forms the *feasible region*, or *feasible set* (let us denote it by S). The goal of an optimization problem in minimization form, as above, is to find a feasible point $\bar{x}$ such that $f(\bar{x}) \leq f(x)$ for any other feasible point x.

2.2 Markowitz portfolio optimization model

Suppose one has the opportunity to invest in n assets. Their future returns are represented by random variables, $R_1, \ldots, R_n$, whose expected values and covariances, $\mathrm{E}[R_i]$, $i = 1, \ldots, n$ and $\mathrm{Cov}(R_i, R_j)$, $i, j = 1, \ldots, n$, respectively, can be estimated based on historical data and, possibly, other considerations. At least one of these assets is a risk-free asset.

Suppose x_i, $i = 1, \ldots, n$, are the fractions of your wealth allocated to each of the assets (that is, $x \geq 0$ and $\sum_{i=1}^n x_i = 1$). The return of the resulting portfolio is a random variable $\sum_{i=1}^n x_i R_i$ with mean $\sum_{i=1}^n x_i \mathrm{E}[R_i]$ and variance $\sum_{i=1}^n \sum_{j=1}^n x_i x_j \mathrm{Cov}(R_i, R_j)$. A portfolio is usually chosen

to optimize some measure of a tradeoff between the expected return and the risk, such as

$$\begin{array}{ll} \text{maximize} & \sum_{i=1}^n x_i \mathrm{E}[R_i] - \mu \sum_{i=1}^n \sum_{j=1}^n x_i x_j \mathrm{Cov}(R_i, R_j) \\ \text{subject to} & \sum_{i=1}^n x_i = 1 \\ & x \geq 0, \end{array}$$

where $\mu > 0$ is a (fixed) parameter reflecting the investor's preferences in the above tradeoff. Since it is hard to assess anybody's value of μ, the above problem can (and should) be solved for a variety of values of μ, thus generating a variety of portfolios on the *efficient frontier*.

2.3 Least squares problem (parameter estimation)

Applications in model constructions, statistics (e.g., linear regression), neural networks, etc.

We consider a linear measurement model, i.e., we stipulate that an (output) quantity of interest $y \in \mathbf{R}$ can be expressed as a linear function $y \approx a^T x$ of input $a \in \mathbf{R}^n$ and model parameters $x \in \mathbf{R}^n$. Our goal is to find the vector of parameters x which provide the "best fit" for the available set of input-output pairs (a_i, y_i), $i = 1, \ldots, m$. If "fit" is measured by sum of squared errors between estimated and measured outputs, solution to the following optimization problem

$$\begin{array}{lll} \text{minimize} & \sum_{i=1}^m (v_i)^2 & = \text{minimize}_{x \in \mathbf{R}^n} \sum_{i=1}^m (y_i - a_i^T x)^2 \quad = \text{minimize}_{x \in \mathbf{R}^n} \|Ax - y\|_2^2, \\ x \in \mathbf{R}^n & & \\ \text{subject to} & v_i = y_i - a_i^T x, \quad i = 1, \ldots, m & \end{array}$$

provides the best fit. Here, A is the matrix with rows a_i^T.

2.4 Maximum likelihood estimation

Consider a family of probability distributions $p_\theta(\cdot)$ on $\mathbf{R}$, parameterized by vector θ. E.g., we could be considering the family of exponential distributions, which is parameterized by a single parameter $\theta = \lambda > 0$ and has the form

$$p_\lambda(t) = \begin{cases} \lambda e^{-\lambda t}, & t \geq 0 \\ 0, & t < 0. \end{cases}$$

Another example of a parametric family of probability distributions is the Normal distribution, parameterized by $\theta = (\mu, \sigma)$, where μ is the mean, and σ — the standard deviation, of the distribution.

When considered as a function of θ for a particular observation of a random variable $y \in \mathbf{R}$, the function $p_\theta(y)$ is called the *likelihood function*. It is more convenient to work with its logarithm, which is called the *log-likelihood function*:

$$l(\theta) = \log p_\theta(y).$$

Consider the problem of estimating the value of the parameter vector θ based on observing one sample point y from the distribution. One possible method, *maximum likelihood (ML) estimation*, is to estimate θ as

$$\hat{\theta} = \operatorname*{argmax}_\theta p_\theta(y) = \operatorname*{argmax}_\theta l(\theta),$$

i.e., to choose as the estimate the value of the parameter that maximizes the likelihood (or the log-likelihood) function for the observed value of y.

If there is prior information available about θ, we can add constraint $\theta \in C \subseteq \mathbf{R}^n$ explicitly, or impose it implicitly, by redefining $p_\theta(y) = 0$ for $\theta \notin C$ (note that in that case $l(\theta) = -\infty$ for $\theta \notin C$).

For m iid sample points $(y_1, \ldots, y_m)$, the log-likelihood function is

$$l(\theta) = \log\left(\prod_{i=1}^{m} p_\theta(y_i)\right) = \sum_{i=1}^{m} \log p_\theta(y_i).$$

The ML estimation is thus an optimization problem:

$$\text{maximize } l(\theta) \text{ subject to } \theta \in C.$$

For example, returning to the linear measurement model $y = a^T x + v$, let us now assume that the error v is iid random noise with known density $p(\cdot)$. If there are m measurement/output pairs (a_i, y_i) available, then the likelihood function is

$$p_x(y) = \prod_{i=1}^{m} p(y_i - a_i^T x),$$

and the log-likelihood function is

$$l(x) = \sum_{i=1}^{m} \log p(y_i - a_i^T x).$$

For example, suppose the noise is Gaussian (or Normal) with mean 0 and standard deviation σ. Then $p(z) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{z^2}{2\sigma^2}}$ and the log-likelihood function is

$$l(x) = -\frac{1}{2}\log(2\pi\sigma) - \frac{1}{2\sigma^2}\|Ax - y\|_2^2.$$

Therefore, the ML estimate of x is $\arg\min_x \|Ax - y\|_2^2$, the solution of the least squares approximation problem (note that the analysis is the same whether σ is known or not).

2.5 Current in a resistive electric network

Consider a linear resistive electric network with node set N and arc set A. Let v_i be the voltage of node i and let x_{ij} be the current on arc (i, j). Kirchhoff's current law says that for each node i, the total incoming current is equal to the total outgoing current:

$$\sum_{j:(j,i)\in A} x_{ji} = \sum_{j:(i,j)\in A} x_{ij}.$$

Ohm's law says that the current x_{ij} and the voltage drop $v_i - v_j$ along each arc (i, j) are related by

$$v_i - v_j = R_{ij}x_{ij} - t_{ij}.$$

where $R_{ij} \geq 0$ is a resistance parameter, and $t_{ij} \geq 0$ is another parameter that is nonzero when there is voltage source along the arc (i, j) (t_{ij} is positive if the voltage source pushes current in the direction from i to j).

Given the arc resistance and arc voltage parameters R_{ij} and t_{ij} for all $(i,j) \in A$, the current in the system is distributed so that to minimize the "energy loss" in the system, while satisfying Kirchhoff's current law. This can be modeled as the following nonlinear programming problem:

$$\begin{array}{ll} \text{minimize} & \sum_{(i,j)\in A} \left(\frac{1}{2} R_{ij} x_{ij}^2 - t_{ij} x_{ij} \right) \\ \text{subject to} & \sum_{j:(i,j)\in A} x_{ij} = \sum_{j:(j,i)\in A} x_{ji} \quad \forall i \in N \end{array}$$

It can be shown by studying the *optimality conditions* for this problem that the optimal solution of the above problem satisfies Ohm's law. Moreover, if a vector of currents values $x^\star$ and a vector of node voltage values $v^\star$ together satisfy Kirchhoff's and Ohm's laws, then the vector $x^\star$ is an optimal solution of the above optimization problem.