Fundamentos de Optimización

Alan Reyes-Figueroa

Métodos Numéricos II (Aula 20) 03.octubre.2024

Optimización

Comenzamos ahora el tema de optimización numérica. Más precisamente, vamos a trabajar, optimización continua no restricta.

Problema de Optimización:
Resolvemos el problema

$$\min_{x \in \Omega} f(x), \tag{1}$$

donde $f : \mathbb{R}^n \to \mathbb{R}$ es la **función objetivo**.
Además, $\mathbf{x} = (x_1, \ldots, x_n)^T \in \mathbb{R}^n$ es un vector de variables independientes. Estas variables usualmente se llaman las **variables de decisión**.
EL conjunto $\Omega \subseteq \mathbb{R}^n$ se llama *conjunto factible* o **conjunto de restricciones**. En el caso de optimización no restricta, Ω es el dominio de la función f, (por lo general $\Omega = \mathbb{R}^n$).

Casi siempre, requerimos que f posea alguna propiedad de interés. Por ejemplo, f es diferenciable, f es convexa, etc. Por lo general, en la **optimización continua** se diseñan métodos y algoritmos para optimizar funciones diferenciables f diferenciables (aunque esto no es un requisito indispensable).

Ejemplos:

$$\min_{\mathbf{x}} \frac{\mathbf{x}^T A \mathbf{x}}{\mathbf{x}^T \mathbf{x}}, \quad \text{donde } A \in \mathbb{R}^{n\times n} \text{ es simétrica.}$$

(cociente de Rayleigh)

$$\min_{\mathbf{x}\in\mathbb{R}^n} \sum_{i=1}^{n} (x_i - y_i)^2 + \lambda \sum_{i=1}^{n-1} (x_{i+1} - x_i)^2.$$

(mínimos cuadrados con regularización de Tychonoff)

$$\min_{\mathbf{x}\in\mathbb{R}^n} \sum_{i=1}^{n-1} \left[(x_{i+1} - x_i^2)^2 + (1 - x_i)^2 \right].$$

(función de Rosenbrock).

Tipos de Extremos

Definición

Suponga que $f : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}$ *es una función con valaores reales, definida sobre* Ω*. Un punto* $\mathbf{x}^* \in \Omega$ *es un* ***mínimo local*** *o* ***minimizador local*** *de* f *si existe* $\varepsilon > 0$ *tal que*

$$f(\mathbf{x}) \geq f(\mathbf{x}^*), \qquad \textit{para todo } \mathbf{x} \in \Omega - \{\mathbf{x}^*\} \textit{ con } ||\mathbf{x} - \mathbf{x}^*|| < \varepsilon.$$

Definición

Suponga que $f : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}$ *es una función con valaores reales, definida sobre* Ω*. Un punto* $\mathbf{x}^* \in \Omega$ *es un* ***mínimo global*** *o* ***minimizador global*** *de* f *sobre* Ω *si*

$$f(\mathbf{x}) \geq f(\mathbf{x}^*), \qquad \textit{para todo } \mathbf{x} \in \Omega - \{\mathbf{x}^*\}.$$

Obs! Reemplazando $\geq$ con $>$ en las definiciones anteriores obtenemos el concepto de un **mínimo local estricto** y de un **mínimo global estricto**, respectivamente.

Notación

A lo largo de este tema, vamos a utilizar algunas notaciones comunes:

- Todos los vectores se consideran vectores columna. Esto es

$$\mathbf{x} = (x_1, x_2, \ldots, x_n)^T \in \mathbb{R}^n \quad \textit{significa} \quad \mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}.$$

- $\mathbf{1} = (1, 1, \ldots, 1)^T \in \mathbb{R}^n$, $\mathbf{0} = (0, 0, \ldots, 0)^T \in \mathbb{R}^n$, $I = \begin{pmatrix} 1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{pmatrix} \in \mathbb{R}^{n \times n}$.
- Una función $f : \mathbb{R}^n \to \mathbb{R}^m$ se representará como $f = (f_1, f_2, \ldots, f_m)^T \in \mathbb{R}^m$. Así

$$f(\mathbf{x}) = \begin{pmatrix} f_1(\mathbf{x}) \\ f_2(\mathbf{x}) \\ \vdots \\ f_m(\mathbf{x}) \end{pmatrix}.$$

Lema (Preservación del signo)

Sea $f : \mathbb{R}^n \to \mathbb{R}$ continua en $\mathbf{a}$, y tal que $f(\mathbf{a}) \neq 0$. Entonces, existe un $\delta > 0$ tal que para todo punto $\mathbf{x} \in \mathbb{D}_\delta(\mathbf{a})$, $f(\mathbf{x})$ tiene el mismo signo que $f(\mathbf{a})$.

Prueba: Sin pérdida, suponga que $f(\mathbf{a}) > 0$. Tome $\varepsilon > 0$. Usando la continuidad de f, existe $\delta > 0$ tal que

$$\begin{aligned} ||\mathbf{x} - \mathbf{a}|| < \delta \quad &\Longrightarrow \quad |f(\mathbf{x}) - f(\mathbf{a})| < \varepsilon \\ &\Longrightarrow \quad -\varepsilon < f(\mathbf{x}) - f(\mathbf{a}) < \varepsilon \\ &\Longrightarrow \quad f(\mathbf{a}) - \varepsilon < f(\mathbf{x}) < f(\mathbf{a}) + \varepsilon. \end{aligned}$$

Tomando, $0 < \varepsilon < f(\mathbf{a})$, por ejemplo $\varepsilon = \frac{f(\mathbf{a})}{2}$, se obtiene que

$$||\mathbf{x} - \mathbf{a}|| < \delta \quad \Longrightarrow \quad f(\mathbf{x}) > f(\mathbf{a}) - \varepsilon > 0,$$

de modo que $f(\mathbf{x}) > 0$ tiene el mismo signo que $f(\mathbf{a})$ en el disco $\mathbb{D}_\delta(\mathbf{a})$. $\square$

En muchos métodos de optimización, se requiere información sobre la primera o segunda derivada de f.

Recordemos que si $f:\mathbb{R}^n \to \mathbb{R}$, es de clase C^1, entonces f tiene primeras derivadas continuas. La **derivada** o **Jacobiana** de f en el punto $\mathbf{p} \in \mathbb{R}^n$, es el mapa lineal $Df:\mathbb{R}^n \to \mathbb{R}$ dado por

$$Df(\mathbf{p}) = Jf(\mathbf{p}) = \begin{pmatrix} \frac{\partial f}{\partial x_1}(\mathbf{p}) & \frac{\partial f}{\partial x_2}(\mathbf{p}) & \cdots & \frac{\partial f}{\partial x_n}(\mathbf{p}) \end{pmatrix}.$$

Si f es de clase C^2, f tiene segundas derivadas parciales continuas. El **Hessiano** de f es

$$D^2f(\mathbf{p}) = Hf(\mathbf{p}) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2}(\mathbf{p}) & \frac{\partial^2 f}{\partial x_1 \partial x_2}(\mathbf{p}) & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(\mathbf{p}) \\ \frac{\partial^2 f}{\partial x_2 \partial x_1}(\mathbf{p}) & \frac{\partial^2 f}{\partial x_2^2}(\mathbf{p}) & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n}(\mathbf{p}) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1}(\mathbf{p}) & \frac{\partial^2 f}{\partial x_n \partial x_2}(\mathbf{p}) & \cdots & \frac{\partial^2 f}{\partial x_n^2}(\mathbf{p}) \end{pmatrix}.$$

Obs! Si $f \in C^2$, vale la igualdad de las segundas parciales mixtas, y $D^2f(\mathbf{p})$ es simétrica.

Derivadas Vectoriales

En el caso general de $f : \mathbb{R}^n \to \mathbb{R}^m$, con $f = (f_1, \ldots, f_m)^T \in \mathbb{R}^m$, la derivada de f es

$$Df(\mathbf{p}) = Jf(\mathbf{p}) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(\mathbf{p}) & \frac{\partial f_1}{\partial x_2}(\mathbf{p}) & \ldots & \frac{\partial f_1}{\partial x_n}(\mathbf{p}) \\ \frac{\partial f_2}{\partial x_1}(\mathbf{p}) & \frac{\partial f_2}{\partial x_2}(\mathbf{p}) & \ldots & \frac{\partial f_2}{\partial x_n}(\mathbf{p}) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1}(\mathbf{p}) & \frac{\partial f_m}{\partial x_2}(\mathbf{p}) & \ldots & \frac{\partial f_m}{\partial x_n}(\mathbf{p}) \end{pmatrix}.$$

Valen las propiedades ya conocidas del cálculo. En particular, vale la pena recordar la

Regla de la Cadena:
Si $f : \mathbb{R}^m \to \mathbb{R}^k$ y $g : \mathbb{R}^n \to \mathbb{R}^m$, osn funciones diferenciables en $g(\mathbf{p}) \in \mathbb{R}^m$ y $\mathbf{p} \in \mathbb{R}^m$, respectivamente, entonces $f \circ g : \mathbb{R}^n \to \mathbb{R}^k$ es diferenciable en $\mathbf{p}$ y

$$D(f \circ g)(\mathbf{p}) = Df\big(g(\mathbf{p})\big) \cdot Dg(\mathbf{p}).$$

Derivadas Vectoriales

Revisamos algunas propiedades útiles de derivadas vectoriales y matriciales.

Sea $\mathbf{x} = (x_1, \ldots, x_n)^T \in \mathbb{R}^n$. Si $f : \mathbb{R}^n \to \mathbb{R}$ es diferenciable, recordemos que su gradiente es el vector

$$\nabla_{\mathbf{x}} f(\mathbf{p}) = \Big(\frac{\partial f}{\partial x_1}(\mathbf{p}), \frac{\partial f}{\partial x_2}(\mathbf{p}), \ldots, \frac{\partial f}{\partial x_n}(\mathbf{p}) \Big)^T \in \mathbb{R}^n.$$

En ocasiones, representaremos $\nabla_{\mathbf{x}} f(\mathbf{p})$ como la aplicación lineal $Df(\mathbf{p}) : \mathbb{R}^n \to \mathbb{R}$.

La derivada direccional de f en la dirección del vector unitario $\mathbf{u} \in \mathbb{R}^n$ es

$$\tfrac{\partial f}{\partial \mathbf{u}}(\mathbf{p}) = \nabla_{\mathbf{x}} f(\mathbf{p}) \cdot \mathbf{u}.$$

Si $f : \mathbb{R}^n \to \mathbb{R}^m$ es diferenciable, recordemos que su derivada o gradiente es la aplicación lineal $\nabla_{\mathbf{x}} f : \mathbb{R}^n \to \mathbb{R}^m$

$$\nabla_{\mathbf{x}} f(\mathbf{p}) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(\mathbf{p}) & \frac{\partial f_1}{\partial x_2}(\mathbf{p}) & \ldots & \frac{\partial f_1}{\partial x_n}(\mathbf{p}) \\ \frac{\partial f_2}{\partial x_1}(\mathbf{p}) & \frac{\partial f_2}{\partial x_2}(\mathbf{p}) & \ldots & \frac{\partial f_2}{\partial x_n}(\mathbf{p}) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1}(\mathbf{p}) & \frac{\partial f_m}{\partial x_2}(\mathbf{p}) & \ldots & \frac{\partial f_m}{\partial x_n}(\mathbf{p}) \end{pmatrix} \in \mathbb{R}^{m \times n}.$$

Derivadas Vectoriales

Observe que, en este caso, la matriz derivada de f puede verse como

$$\nabla_{\mathbf{x}} f(\mathbf{p}) = \begin{pmatrix} \nabla_{\mathbf{x}} f_1(\mathbf{p})^T \\ \nabla_{\mathbf{x}} f_2(\mathbf{p})^T \\ \vdots \\ \nabla_{\mathbf{x}} f_m(\mathbf{p})^T \end{pmatrix},$$

o bien

$$\nabla_{\mathbf{x}} f(\mathbf{p}) = \begin{pmatrix} \nabla_{x_1} f(\mathbf{p}) & \nabla_{x_2} f(\mathbf{p}) & \dots & \nabla_{x_n} f(\mathbf{p}) \end{pmatrix}.$$

En particular, si $\nabla_{\mathbf{x}} = (\nabla_{x_1}, \dots, \nabla_{x_n})^T$ y $f = (f_1, f_2, \dots, f_m)^T \in \mathbb{R}^m$, tenemos que

$$\nabla_{\mathbf{x}} f(\mathbf{p}) = \begin{pmatrix} \nabla_{x_1} \\ \dots \\ \nabla_{x_n} \end{pmatrix} \begin{pmatrix} f_1(\mathbf{p}) & \dots & f_m(\mathbf{p}) \end{pmatrix}.$$

Derivadas Vectoriales

Ejemplo: $f : \mathbb{R}^n \to \mathbb{R}^m$ dada por $f(\mathbf{x}) = A\mathbf{x}$, con $A = (a_{ij}) \in \mathbb{R}^{m\times n}$.
Tenemos

$$f(\mathbf{x}) = \begin{pmatrix} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n \\ \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n \end{pmatrix} \implies \nabla_{\mathbf{x}} f(\mathbf{x}) = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{pmatrix}.$$

En consecuencia, $\nabla_{\mathbf{x}}(A\mathbf{x}) = A$.

Ejemplo: $f : \mathbb{R}^n \to \mathbb{R}^m$ dada por $f(\mathbf{x}) = \mathbf{x}^T A$, con $A = (a_{ij}) \in \mathbb{R}^{n\times m}$.
Tenemos

$$f(\mathbf{x}) = \begin{pmatrix} a_{11}x_1 + a_{21}x_2 + \ldots + a_{n1}x_n \\ a_{12}x_1 + a_{22}x_2 + \ldots + a_{n2}x_n \\ \vdots \\ a_{1m}x_1 + a_{2m}x_2 + \ldots + a_{nm}x_n \end{pmatrix} \implies \nabla_{\mathbf{x}} f(\mathbf{x}) = \begin{pmatrix} a_{11} & a_{21} & \ldots & a_{n1} \\ a_{12} & a_{22} & \ldots & a_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1m} & a_{2m} & \ldots & a_{nm} \end{pmatrix}.$$

En consecuencia, $\nabla_{\mathbf{x}}(\mathbf{x}^T A) = A^T$.

Derivadas Vectoriales

En general,

$$\begin{aligned}\nabla_{\mathbf{x}}\left(A\mathbf{x}+\mathbf{b}\right) &= A,\\ \nabla_{\mathbf{x}}\left(\mathbf{x}^TA+\mathbf{b}^T\right) &= A^T = \left(\nabla_{\mathbf{x}}\left(A^T\mathbf{x}+\mathbf{b}\right)^T\right) = \left(\nabla_{\mathbf{x}}\left(A^T\mathbf{x}+\mathbf{b}\right)\right)^T.\end{aligned}$$

Otra forma de verlo. Recordemos que el producto matriz-vector, puede verse como el producto punto $A\mathbf{x} = \langle A^T, \mathbf{x}\rangle$.

Tenemos la siguientes propiedades:

$$\begin{aligned}\nabla_{\mathbf{x}}\langle A^T, \mathbf{x}\rangle &= A, & \nabla_{\mathbf{x}}\langle A, \mathbf{x}\rangle &= A^T,\\ \nabla_{\mathbf{x}}\langle \mathbf{x}, A^T\rangle &= A, & \nabla_{\mathbf{x}}\langle \mathbf{x}, A\rangle &= A^T.\end{aligned}$$

Otras propiedades:

$$\begin{aligned}\nabla_{\mathbf{x}}\,\mathbf{b} &= \mathbf{0}, & \nabla_{\mathbf{x}}\,\mathbf{y}^T\mathbf{x} &= \mathbf{y}^T,\\ \nabla_{\mathbf{x}}\,\mathbf{x} &= I, & \nabla_{\mathbf{x}}\,c\mathbf{x} &= cI.\end{aligned}$$

Ejemplo: Sea $f : \mathbb{R}^n \to \mathbb{R}$ dada por $f(\mathbf{x}) = \mathbf{x}^T A\mathbf{x}$, con $A \in \mathbb{R}^{n\times n}$.
El gradiente es

$$\begin{aligned}
\nabla_{\mathbf{x}} f(\mathbf{x}) &= \nabla_{\mathbf{x}} \langle \mathbf{x}, A\mathbf{x}\rangle = \langle \nabla_{\mathbf{x}} \mathbf{x}, A\mathbf{x}\rangle + \langle \mathbf{x}, \nabla_{\mathbf{x}} A\mathbf{x}\rangle = \langle I, A\mathbf{x}\rangle + \langle \mathbf{x}, A\rangle = \langle I, A\mathbf{x}\rangle + \langle A, \mathbf{x}\rangle \\
&= A\mathbf{x} + A^T\mathbf{x} = (A + A^T)\mathbf{x}.
\end{aligned}$$

En el caso en que A es simétrica, obtenemos

$$\nabla_{\mathbf{x}} (\mathbf{x}^T A\mathbf{x}) = 2A\mathbf{x}.$$

Otras propiedades similares:

$$\begin{aligned}
\nabla_{\mathbf{x}} ||\mathbf{x}||^2 &= 2\mathbf{x}, \\
\nabla_{\mathbf{x}} \frac{\mathbf{x}^T A\mathbf{x}}{\mathbf{x}^T\mathbf{x}} &= \frac{||\mathbf{x}||^2 (A + A^T)\mathbf{x} - 2(\mathbf{x}^T A\mathbf{x})\mathbf{x}}{||\mathbf{x}||^4}.
\end{aligned}$$

Derivadas Matriciales

Ahora mencionamos algunas derivadas matriciales. Sea $X \in \mathbb{R}^{n\times p}$ una matriz con entradas $X = (x_{ij})$, y $f : \mathbb{R}^{n\times p} \to \mathbb{R}$ es una función diferenciable. Entonces definimos

$$\nabla_X f(X) = \left(\frac{\partial f}{\partial x_{ij}}\right) = \begin{pmatrix} \frac{\partial f}{\partial x_{11}} & \frac{\partial f}{\partial x_{12}} & \cdots & \frac{\partial f}{\partial x_{1p}} \\ \frac{\partial f}{\partial x_{21}} & \frac{\partial f}{\partial x_{22}} & \cdots & \frac{\partial f}{\partial x_{2p}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial x_{n1}} & \frac{\partial f}{\partial x_{n2}} & \cdots & \frac{\partial f}{\partial x_{np}} \end{pmatrix}.$$

Ejemplo: Sea $f : \mathbb{R}^{n\times n} \to \mathbb{R}$, dada por $f(X) = \operatorname{tr} X$. En este caso

$$f(X) = \sum_{j=1}^{n} x_{jj} \quad\Longrightarrow\quad \nabla_X f(X) = \begin{pmatrix} 1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{pmatrix} = I.$$

Derivadas Matriciales

Ejemplo: Sea $f : \mathbb{R}^{n\times n} \to \mathbb{R}$, dada por $f(X) = \text{tr}(AX)$, $A \in \mathbb{R}^{n\times n}$.

En este caso, las entradas de AX son de la forma $(AX)_{ij} = \sum_{k=1}^{n} a_{ik}x_{kj}$.

Calculando la traza, obtenemos

$$f(X) = \text{tr}(AX) = \sum_{j=1}^{n}\sum_{k=1}^{n}(AX)_{jj} = \sum_{j=1}^{n}\sum_{k=1}^{n} a_{jk}x_{kj}.$$

De ahí que

$$\nabla_X f(X) = \Big(\nabla_{x_{ij}} \sum_{j=1}^{n}\sum_{k=1}^{n} a_{jk}x_{kj}\Big) = \begin{pmatrix} a_{11} & a_{21} & \dots & a_{n1} \\ a_{12} & a_{22} & \dots & a_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2n} & \dots & a_{nn} \end{pmatrix} = A^T.$$

Portanto $\nabla_X \text{tr}(AX) = A^T$.
Similarmente, $\nabla_X \text{tr}(XA) = A^T$.

Derivadas Matriciales

Otras propiedades útiles:

- $\nabla_X\, \mathbf{a}^T X \mathbf{b} = \mathbf{a}\mathbf{b}^T$.
- $\nabla_X \operatorname{tr}(AXB) = \nabla_X \operatorname{tr}(BAX) = A^T B^T$.
- $\nabla_X \operatorname{tr}(X^T A X) = (A + A^T)X$.
- $\nabla_X \operatorname{tr}(X^{-1}) = -(X^{-T})^2$.
- $\nabla_X \operatorname{tr}(AX^{-1}) = \nabla_X \operatorname{tr}(X^{-1}A) = -(X^T)^{-1} A^T (X^{-1})^T$.
- $\nabla_X \det(X) = |X|(X^{-1})^T$.

Ver una lista más completa de propiedades de derivadas vectoriales y matriciales en `https://en.wikipedia.org/wiki/Matrix_calculus`