DSA5103 Optimization Problem for Data Modelling

AY2024/25 Sem2 By Zhao Peiduo

Lecture 1

Nonlinear Programming A general **nonlinear programming problem (NLP)** is to minimize/maximize a function $f(x)$, subject to equality constraints $g_i(x) = 0$, $i \in [m]$, and inequality constraints $h_j(x) \le 0$, $j \in [p]$. Here, f, g_i, and h_j are functions of the variable $x = (x_1, \ldots, x_n)^T \in \mathbb{R}^n$. The term definitions are as follows:

- f: **Objective function**
- $g_i(x) = 0$: **Equality constraints**
- $h_j(x) \le 0$: **Inequality constraints**

It suffices to discuss minimization problems since minimizing $f(x)$ is equivalent to maximizing $-f(x)$.

Feasible Set

$$S = \{x \in \mathbb{R}^n \mid g_1(x) = 0, \ldots, g_m(x) = 0,\ h_1(x) \le 0, \ldots, h_p(x) \le 0\}.$$

A point in the feasible set is a **feasible solution** or **feasible point** wbere all constraints are satisfied; otherwise, it is an **infeasible solution** or **infeasible point**. When there is no constraint, $S = \mathbb{R}^n$, we say the NLP is **unconstrained**.

Local and Global Minimizer Let S be the feasible set. Define $B_\epsilon(y) = \{x \in \mathbb{R}^n \mid \|x - y\| < \epsilon\}$ to be the open ball with center y and radius ϵ. Here, $\|x\| = \sqrt{x_1^2 + \cdots + x_n^2}$.

1. A point $x^* \in S$ is said to be a **local minimizer** of f if there exists $\epsilon > 0$ such that
$$f(x^*) \le f(x) \quad \forall x \in S \cap B_\epsilon(x^*).$$
2. A point $x^* \in S$ is said to be a **global minimizer** of f if
$$f(x^*) \le f(x) \quad \forall x \in S.$$

Interior point Let $S \subseteq \mathbb{R}^n$ be a nonempty set. An point $x \in S$ is called an **interior point** of S if

$$\exists \epsilon > 0 \quad s.t. \quad B_\epsilon(x) \subseteq S.$$

Gradient Vector Let $S \subseteq \mathbb{R}^n$ be a nonempty set. Suppose $f : S \to \mathbb{R}$, and x is an interior point of S such that f is differentiable at x. Then the **gradient vector** of f at x is

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(x) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x) \end{bmatrix}.$$

Hessian Matrix Let $S \subseteq \mathbb{R}^n$ be a nonempty set. Suppose $f : S \to \mathbb{R}$, and x is an interior point of S such that f has second-order partial derivatives at x. Then the **Hessian** of f at x is the $n \times n$ matrix:

$$H_f(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2}(x) & \frac{\partial^2 f}{\partial x_1 \partial x_2}(x) & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(x) \\ \frac{\partial^2 f}{\partial x_2 \partial x_1}(x) & \frac{\partial^2 f}{\partial x_2^2}(x) & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n}(x) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1}(x) & \frac{\partial^2 f}{\partial x_n \partial x_2}(x) & \cdots & \frac{\partial^2 f}{\partial x_n^2}(x) \end{bmatrix}.$$

- The ij-entry of $H_f(x)$ is $\frac{\partial^2 f}{\partial x_i \partial x_j}(x)$.
- In general, $H_f(x)$ is not symmetric. However, if f has continuous second-order derivatives, then the Hessian matrix is symmetric since ∂x_i and ∂x_j are interchangeable.

Positive (Semi)Definite Let A be a real $n \times n$ matrix.

1. A is said to be **positive semidefinite** if $x^T Ax \ge 0$, $\forall x \in \mathbb{R}^n$.
2. A is said to be **positive definite** if $x^T Ax > 0$, $\forall x \ne 0$.
3. A is said to be **negative semidefinite** if $-A$ is positive (semi)definite.
4. A is said to be **negative definite** if $-A$ is positive definite.
5. A is said to be **indefinite** if A is neither positive nor negative semidefinite.

Eigenvalue Test Theorem Let A be a real symmetric $n \times n$ matrix.

1. A is **positive semidefinite** iff every eigenvalue of A is nonnegative.
2. A is **positive definite** iff every eigenvalue of A is positive.
3. A is **negative semidefinite** iff every eigenvalue of A is nonpositive.
4. A is **negative definite** iff every eigenvalue of A is negative.
5. A is **indefinite** iff it has both a positive eigenvalue and a negative eigenvalue.

Proof for: A is positive semidefinite iff every eigenvalue of A is nonnegative

(Forward) Suppose A is positive semidefinite, show that its eigenvalues are nonnegative. By definition, a Hermitian matrix A is positive semidefinite if for all nonzero vectors $x \in \mathbb{C}^n$:

$$x^* Ax \ge 0$$

Let λ be an eigenvalue of A with corresponding eigenvector x such tha $Ax = \lambda x$. Taking the inner product of both sides with x, we obtain:

$$x^* Av = v^*(\lambda x) = \lambda(x^* x)$$

Since $v^* v$ (the squared norm of v) is always positive for nonzero v, the above equation implies $\lambda \ge 0$

(Backward) Since A is Hermitian, it has an orthonormal basis of eigenvectors $\{q_1, q_2, \ldots, q_n\}$ with corresponding real eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$.

For any vector x, we can express it in terms of the eigenvectors as:

$$x = \sum_{i=1}^{n} c_i q_i$$

for some scalars c_i, and compute the quadratic form:

$$x^* Ax = \left(\sum_{i=1}^{n} c_i^* q_i^*\right) A \left(\sum_{j=1}^{n} c_j q_j\right)$$

Expanding the expression using the orthonormality of the eigenvectors:

$$x^* Ax = \sum_{i=1}^{n} \lambda_i |c_i|^2$$

Since we are given that all eigenvalues $\lambda_i \ge 0$, and the squared magnitudes $|c_i|^2$ are nonnegative, it follows that:

$$x^* Ax \ge 0 \quad \forall x \ne 0$$

Thus, A is positive semidefinite.

Necessary and Sufficient Conditions

Assume that $f : \mathbb{R}^n \to \mathbb{R}$ is nonlinear and differentiable. A point x^* is called a **stationary point** of f if $\nabla f(x^*) = 0$.

Necessary condition: Confine our search for global minimizers within the set of stationary points

If x^* is a local minimizer of f, then

1. x^* is a stationary point, i.e., $\nabla f(x^*) = 0$
2. The Hessian $H_f(x^*)$ is positive semidefinite

Sufficient condition: Verify that a point is indeed a local minimizer

If the following conditions hold, then x^* is a local minimizer of f.

1. x^* is a stationary point, i.e., $\nabla f(x^*) = 0$
2. The Hessian $H_f(x^*)$ is positive definite,

Convex set A set $D \in \mathbb{R}^n$ is said to be a **convex** set if for any two points x and y in D, the line segment joining x and y also lies in D. That is,

$$x, y \in D \Rightarrow \lambda x + (1 - \lambda)y \in D \quad \forall \lambda \in [0, 1].$$

Strictly convex function

Let $D \subseteq \mathbb{R}^n$ be a convex set. Consider a function $f : D \to \mathbb{R}$.

1. The function f is said to be **convex** if $f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda)f(y), \quad \forall x, y \in D,\ \lambda \in [0, 1]$.
2. The function f is said to be **strictly convex** if $f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda)f(y)$. for all distinct $x, y \in D$, $\lambda \in (0, 1)$.

For a convex f It holds that

1. any local minimizer is a global minimizer.
2. if f is strictly convex, then the global minimizer is unique.

Test for convexity of a differentiable function

Suppose that f has continuous second partial derivatives on an open convex set D in $\mathbb{R}^n$.

1. The function f is convex on D **iff** the Hessian matrix $H_f(x)$ is positive semidefinite at each $x \in D$.
2. If $H_f(x)$ is positive definite at each $x \in D$, then f is strictly convex on D.
3. If $H_f(\hat{x})$ is indefinite at some point $\hat{x} \in D$, then f is not a convex nor a concave function on D.

Eigenvalue Decomposition: The eigenvalue decomposition of $A \in \mathbb{S}^n$ is given by:

$$A = Q\Lambda Q^T = [Q_{\cdot 1} \quad \cdots \quad Q_{\cdot n}] \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix} [Q_{\cdot 1} \quad \cdots \quad Q_{\cdot n}]^T$$

where Q is an orthogonal matrix whose **columns** are eigenvectors of A, Λ is a diagonal matrix with eigenvalues of A on the diagonal.

Change of bases using eigenvectors Denote the ith column of orthogonal matrix Q as q_i. Change the bases to $\{q_1, q_2\}$. With new bases,

- For any vector x, $x = Q(Q^T x)$, so its representation becomes
$$\tilde{x} = Q^T x = \begin{bmatrix} \tilde{x}_1 \\ \tilde{x}_2 \end{bmatrix}$$
- Since $y = Ax = Q\Sigma Q^T x$, the representation of y is
$$\tilde{y} = \Sigma \tilde{x} = \begin{bmatrix} \lambda_1 \tilde{x}_1 \\ \lambda_2 \tilde{x}_2 \end{bmatrix}$$

Hence, the linear transformation results in a scaling of λ along the eigenvector associated with λ.

Statistical Properties Let $x_1, \dots, x_n \in \mathbb{R}^p$ be n observations of a random variable x.

- Mean vector: $\mu = \bar{x} = \frac{1}{n}\sum_{i=1}^n x_i \in \mathbb{R}^p$
- (Sample/Empirical) Covariance matrix: $\Sigma = \frac{1}{n-1}\sum_{i=1}^n (x_i - \mu)(x_i - \mu)^T \in \mathbb{R}^{p\times p}$ (Covariance matrices are symmetric and positive semidefinite)
- Standard deviation (for $p = 1$): $\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^n (x_i - \mu)^2}$

PCA

- PCA is often used to **reduce the dimensionality** of large data sets while preserving as much information as possible.
- PCA allows us to identify the **principal directions in which the data varies**.

Let $x_1, \dots, x_n \in \mathbb{R}^p$ be n observations of a random variable x and

$$X = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix}.$$

The mean vectors of x_i and $Q^T x_i$ (for $i = 1, \dots, n$) are, respectively,

$$\mu = \frac{1}{n}\sum_{i=1}^n x_i, \quad \hat{\mu} = \frac{1}{n}\sum_{i=1}^n Q^T x_i = Q^T \mu.$$

Consequently, the associated covariance matrices are, respectively,

$$\Sigma = \frac{1}{n-1}\sum_{i=1}^n (x_i - \mu)(x_i - \mu)^T,$$

$$\hat{\Sigma} = \frac{1}{n-1}\sum_{i=1}^n (Q^T x_i - Q^T \mu)(Q^T x_i - Q^T \mu)^T = Q^T \Sigma Q.$$

Optimization problem of PCA

$$\max_{Q\in\mathbb{R}^{p\times k},\, Q^T Q = I} \operatorname{trace}(Q^T \Sigma Q).$$

Let the eigenvalue decomposition of Σ be

$$\Sigma = \begin{bmatrix} q_1 & \cdots & q_p \end{bmatrix} \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_p \end{bmatrix} \begin{bmatrix} q_1 & \cdots & q_p \end{bmatrix}^T,$$

where

$$\lambda_1 \ge \cdots \ge \lambda_p \ge 0.$$

Then

$$Q = \begin{bmatrix} q_1 & \cdots & q_k \end{bmatrix}.$$

Standard PCA workflow

1. Make sure the data X are rows = observations and columns = variables.
2. Standardize the columns of X.
3. Run $[Q, X_{\text{new}}, d, \text{tsquared}, \text{explained}] = \text{pca}(X)$.
4. Using the variance% in "explained", choose k (usually 1, 2, or 3) components for visual analysis.
    - For example, if $d = (1.9087, 0.0913)$, explained= $(95.4, 4.6)$, one may choose $k = 1$ as the first principal component carries 95.4% of the information.
    - For example, if $d = (2.9108, 0.9212, 0.1474, 0.0206)$, explained= $(72.8, 23.0, 3.7, 0.5)$, one may choose $k = 2$ as the first two principal components carry 95.8% of the information.
5. Plot $X_{\text{new}}(:, 1), \dots, X_{\text{new}}(:, k)$ on a k-dimensional plot.