Chapter 2: Fundamentals of Unconstrained Optimization

Outline

Outline

Vectors and Matrices

- A vector $x \in \mathbb{R}^n$: $x = (x_1, \ldots, x_n)^T$
- Inner product: given $x, y \in \mathbb{R}^n$, $x^T y = \sum_{i=1}^n x_i y_i$
- A matrix $A \in \mathbb{R}^{m \times n}$
- $A \in \mathbb{R}^{n \times n}$ is *positive semidefinite*, if $x^T A x \geq 0$ for any $x \in \mathbb{R}^n$
- $Q \in \mathbb{R}^{n \times n}$ is *orthogonal*, if $Q^T Q = Q Q^T = I$.
- Eigenvalue λ, eigenvector x: $Ax = \lambda x$

Vector Norms

- $x \in \mathbb{R}^n$,

$$l_1\text{-norm:} \quad \|x\|_1 = \sum_{i=1}^{n} |x_i|$$

$$l_2\text{-norm:} \quad \|x\|_2 = \left(\sum_{i=1}^{n} x_i^2 \right)^{1/2} = (x^T x)^{1/2}$$

$$l_\infty\text{-norm:} \quad \|x\|_\infty = \max_{i=1,\ldots,n} |x_i|$$

- $\|x\|_\infty \leq \|x\|_2 \leq \sqrt{n}\|x\|_\infty$ and $\|x\|_\infty \leq \|x\|_1 \leq n\|x\|_\infty$
- Cauchy-Schwarz inequality: $|x^T y| \leq \|x\|_2 \|y\|_2$

Dual Norm

- Dual norm of $\|\cdot\|$:

$$\|x\|_D = \max_{\|y\|=1} x^T y = \max_{y \neq 0} \frac{x^T y}{\|y\|}$$

- $|x^T y| \leq \|y\| \|x\|_D$
- $\|\cdot\|_1 \sim \|\cdot\|_\infty$
- $\|\cdot\|_2 \sim \|\cdot\|_2$

Matrix Norms

- Given $A \in \mathbb{R}^{m\times n}$, define $\|A\| = \sup_{x\neq 0} \frac{\|Ax\|}{\|x\|}$,

$$\|A\|_1 = \max_{j=1,\dots,n} \sum_{i=1}^{m} |A_{ij}|,$$

$$\|A\|_2 = \text{largest eigenvalue of } (A^T A)^{1/2},$$

$$\|A\|_\infty = \max_{i=1,\dots,m} \sum_{j=1}^{n} |A_{ij}|$$

- Frobenius norm:

$$\|A\|_F = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij}^2 \right)^{1/2}$$

- Condition number: $\kappa(A) = \|A\|\|A^{-1}\|$

Subspaces

- Given $\mathcal{S} \subset \mathbb{R}^n$, it is called a subspace if for any $x, y \in \mathcal{S}$,

$$\alpha x + \beta y \in \mathcal{S}, \text{ for all } \alpha, \beta \in \mathbb{R}.$$

- Given $a_i \in \mathbb{R}^n$, $i = 1, \ldots, m$, are the following sets

$$\mathcal{S} = \{w \in \mathbb{R}^n | a_i^T w = 0, i = 1, \ldots, m\}$$

and

$$\mathcal{S} = \{w \in \mathbb{R}^n | a_i^T w \geq 0, i = 1, \ldots, m\}$$

subspaces?

- *Null space*: given $A \in \mathbb{R}^{m \times n}$, $\text{Null}(A) = \{w \in \mathbb{R}^n | Aw = 0\}$
- *Range space*: $\text{Range}(A) = \{w \in \mathbb{R}^m | w = Av \text{ for some vector } v \in \mathbb{R}^n\}$
- $\text{Null}(A) \bigoplus \text{Range}(A^T) = \mathbb{R}^n$

Continuity

- Let $f : \mathcal{D} \subseteq \mathbb{R}^n \to \mathbb{R}^m$. For some $x_0 \in \text{cl}\mathcal{D}$, we write

$$\lim_{x \to x_0} f(x) = f_0, \tag{1.1}$$

  if for all $\epsilon > 0$, there is a value $\delta > 0$ such that

$$\|x - x_0\| < \delta \text{ and } x \in \mathcal{D} \Rightarrow \|f(x) - f_0\| < \epsilon.$$

- We say f is continuous at x_0 if $x_0 \in \mathcal{D}$ and (1.1) holds with $f_0 = f(x_0)$. We say f is continuous on $\mathcal{D}$ if it is continuous for all $x_0 \in \mathcal{D}$.
- We say f is Lipschitz continuous on some set $\mathcal{N} \subset \mathcal{D}$ if there is a constant $L > 0$ such that

$$\|f(x_1) - f(x_0)\| \leq L\|x_1 - x_0\|, \quad \text{for all } x_0, x_1 \in \mathcal{N}.$$

  (L is called the *Lipschitz constant*.)

Derivatives

- Let $\phi : \mathbb{R} \to \mathbb{R}$. The first derivative $\phi'(\alpha) = \frac{d\phi}{d\alpha} := \lim_{\epsilon \to 0} \frac{\phi(\alpha+\epsilon)-\phi(\alpha)}{\epsilon}$.
- Frechet differentiability: $f : \mathbb{R}^n \to \mathbb{R}$, is differentiable at x if there exists $g \in \mathbb{R}^n$ such that

$$\lim_{y \to 0} \frac{f(x+y) - f(x) - g^T y}{\|y\|} = 0$$

- *Gradient* of f:

$$g(x) = \nabla f(x) = \left(\frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_n} \right)^T \in \mathbb{R}^n$$

  where $\frac{\partial f}{\partial x_i} = \lim_{\epsilon \to 0} \frac{f(x+\epsilon e_i)-f(x)}{\epsilon}$

- *Hessian* of f:

$$H(x) = \nabla^2 f(x) = \left[\frac{\partial^2 f}{\partial x_i \partial x_j} \right] \in \mathbb{R}^{n \times n}$$

- Notations: $g(x) = \nabla f(x)$, $H(x) = \nabla^2 f(x)$

Derivatives

- Chain rule: $\alpha, \beta \in \mathbb{R}$ and $\alpha = \alpha(\beta)$. Then

$$\frac{d\phi(\alpha(\beta))}{d\beta} = \phi'(\alpha)\alpha'(\beta)$$

- Chain rule: $x, t \in \mathbb{R}^n$ and $x = x(t)$. Define $h(t) = f(x(t))$, then

$$\nabla h(t) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} \nabla x_i(t) = \nabla x(t) \nabla f(x(t)).$$

- Directional derivative: $D(f(x) : p) = \lim_{\epsilon \to 0} \frac{f(x+\epsilon p) - f(x)}{\epsilon} = \nabla f(x)^T p$

Convergence Rate

- Let $\{x_k\}$ be a sequence in $\mathbb{R}^n$ that converges to x^*.
- The convergence is *Q-linear* if there exists a constant $\gamma \in (0,1)$ such that

$$\frac{\|x_{k+1} - x_*\|}{\|x_k - x^*\|} \leq r, \quad \text{for all } k \text{ sufficiently large.}$$

- The convergence is *Q-superlinear* if

$$\lim_{k\to\infty} \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} = 0.$$

- The convergence is *Q-quadratic* if there exists a constant M such that

$$\frac{\|x_{k+1} - x_*\|}{\|x_k - x^*\|^2} \leq M, \quad \text{for all } k \text{ sufficiently large.}$$

- The convergence is *R-linear* if there is sequenc of nonnegative scalars $\{\nu_k\}$ such that

$$\|x_k - x^*\| \le \nu_k \text{ for all } k, \text{ and } \{\nu_k\} \text{ converges } Q\text{-linearly to zero.}$$

- The sequence $\{x_k - x^*\}$ is said to be dominated by $\{\nu_k\}$.
- We say $\{x_k\}$ converges *R-superlinearly* to x^* if $\{\|x_k - x^*\|\}$ is dominated by a sequence of scalars converging Q-superlinearly to zero.
- We say $\{x_k\}$ converges *R-quadratically* to x^* if $\{\|x_k - x^*\|\}$ is dominated by a sequence of scalars converging Q-quadratically to zero.

Outline

Mathematical Formulation for Unconstrained Optimization

- Unconstrained optimization problem:

$$\min_{x \in \mathbb{R}^n} \quad f(x) \tag{2.1}$$

  where $f : \mathbb{R}^n \to \mathbb{R}$ is a smooth function.

- Often the information about f does not come cheaply, so we usually prefer algorithms that do not call for this information unnecessarily.

Solution Definition

- A point x^* is a *global minimizer* if

$$f(x^*) \leq f(x) \quad \text{for all } x \in \mathbb{R}^n.$$

- A point x^* is a *local minimizer* if there is a neighborhood $\mathcal{N}$ of x^* such that

$$f(x^*) \leq f(x) \quad \text{for all } x \in \mathcal{N}.$$

- A point x^* is a *strict local minimizer* if there is a neighborhood $\mathcal{N}$ of x^* such that

$$f(x^*) < f(x) \quad \text{for all } x \in \mathcal{N} \text{ with } x \neq x^*.$$

- A point x^* is an *isolated local minimizer* if there is a neighborhood $\mathcal{N}$ of x^* such that

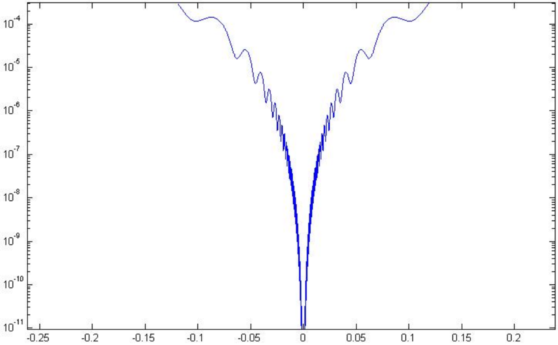
$$x^* \text{ is the only local minimizer in } \mathcal{N}.$$

A Counter Example

All isolated local minimizer are strict. But strict minimizer are not always isolated.
For example, for function

$$f(x) = x^4 \cos\left(\frac{1}{x}\right) + 2x^4, \qquad f(0) = 0.$$

$x = 0$ is a strict local minimizer. However, there are strict local minimizers at many nearby points x_j, and we can label these points so that $x_j \to 0$ as $j \to \infty$.



Recognizing a Local Minimum

- From the definitions given above, it might seem that the only way to find out whether a point x^* is a local minimum is to examine all the points in its immediate vicinity, to make sure that none of them has a smaller function value.
- When the function f is *smooth*, however, there are more efficient and practical ways to identify local minima.
- In particular, if f is twice continuously differentiable, we may be able to tell that x^* is a local minimizer (and possibly a strict local minimizer) by examining just the gradient $\nabla f(x^*)$ and the Hessian $\nabla^2 f(x^*)$.
- The mathematical tool used to study minimizers of smooth functions is Taylor's theorem.

Recognizing a Local Minimum

Theorem 1 (Taylor's Theorem)

Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ *is a continuously differentiable and that* $p \in \Re^n$*. Then we have that*

$$f(x+p) = f(x) + \nabla f(x+tp)^T p, \tag{2.2}$$

for some $t \in (0,1)$*. Moreover, if* f *is twice continuously differentiable, we have that*

$$\nabla f(x+p) = \nabla f(x) + \int_0^1 \nabla^2 f(x+tp) p dt, \tag{2.3}$$

and that

$$f(x+p) = f(x) + \nabla f(x)^T p + \frac{1}{2} p^T \nabla^2 f(x+tp) p, \tag{2.4}$$

for some $t \in (0,1)$*.*

Recognizing a Local Minimum - Necessary Conditions

Theorem 2 (First-Order Necessary Conditions)

If x^* is a local minimizer and f is a continuously differentiable in an open neighborhood of x^*, then $\nabla f(x^*) = 0$.

Proof sketch. By contradiction. Assume that $\nabla f(x^*) \neq 0$. Define $p = -\nabla f(x^*)$. Because ∇f is continuous near x^*, there exists $T > 0$ such that

$$p^T \nabla f(x^* + tp) < 0, \quad \forall t \in [0, T].$$

Then for any $\bar{t} \in (0, T]$, by Taylor's theorem we have for some $t \in [0, T]$,

$$f(x^* + \bar{t}p) = f(x^*) + \bar{t}p^T \nabla f(x^* + tp) < f(x^*).$$

Theorem 3 (Second-Order Necessary Conditions)

If x^* is a local minimizer and f and $\nabla^2 f(x)$ exists and is a continuous in an open neighborhood of x^*, then $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ is positive semidefinite.

Proof sketch. By contradiction to prove second part. Assume that $\nabla^2 f(x^*)$ is not positive semidefinite. Then choose p such that $p^T \nabla^2 f(x^*) p < 0$. Then $\exists\, T > 0$ such that

$$p^T \nabla^2 f(x^* + tp) p < 0 \quad \forall t \in [0, T].$$

Therefore, for any $\bar{t} \in (0, T]$, we have for some $t \in (0, \bar{t})$ that

$$f(x^* + \bar{t}p) = f(x^*) + \bar{t}p^T \nabla f(x^*) + \frac{1}{2}\bar{t}^2 p^T \nabla^2 f(x^* + tp) p < f(x^*).$$

Remark

- *Necessary conditions* for optimality are derived by assuming that x^* is a local minimizer and then proving the facts about $\nabla f(x^*)$ and $\nabla^2 f(x^*)$;
- We call x^* a *stationary point* if $\nabla f(x^*) = 0$. According to the above theorem, any local minimizer must be a stationary point.

We now describe *sufficient conditions*, which are conditions on the derivatives of f at the point x^* that guarantee that x^* is a local minimizer.

Theorem 4 (Second-Order Sufficient Conditions)

Suppose that $\nabla^2 f(x)$ is continuous in an open neighborhood of x^* and that $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ is positive definite. Then x^* is a strict local minimizer of f.

Proof sketch. Choose a radius $r > 0$ such that $\nabla^2 f(x)$ is positive definite in the set $D = \{x | \|x - x^*\| < r\}$. For any nonzero p with $\|p\| < r$, there exists $t \in (0, 1)$ such that

$$f(x^* + p) = f(x^*) + p^T \nabla f(x^*) + \frac{1}{2} p^T \nabla^2 f(x^* + tp) p > f(x^*).$$

Recognizing a Local Minimum - Sufficient Conditions

Remark

- The second-order sufficient conditions of above theorem guarantee something stronger than the necessary conditions discussed earlier; namely, that the minimizer is a *strict* local minimizer.
- Note too that the second-order sufficient conditions are not necessary: A point x^* may be a strict local minimizer, and yet may fail to satisfy the sufficient conditions.

A simple example:

$$f(x) = x^4,$$

for which the point $x^* = 0$ is a strict minimizer at which the Hessian matrix vanishes (and is therefore not positive definite).

Recognizing a Local Minimum - Convex Functions

When the objective function is convex and global minimizer are simple to characterize.

Theorem 5

When f is convex, any local minimizer x^* is a global minimizer of f. If in addition f is differential, then any stationary point x^* is a global minimizer of f.

Proof sketch. By contradiction to prove the first part. Suppose that x^* is a local but not a global minimizer. Then $\exists\ z$ with $f(z) < f(x^*)$. Consider the line segment:

$$x = \lambda z + (1-\lambda)x^*, \quad \lambda \in (0,1].$$

By convexity, we have

$$f(x) \leq \lambda f(z) + (1-\lambda) f(x^*) < f(x^*).$$

For the second part, suppose that the stationary point x^* is not a global minimizer. Then

$$\begin{aligned}\nabla f(x^*)^T(z - x^*) &= \lim_{\lambda \downarrow 0} \frac{f(x^* + \lambda(z - x^*)) - f(x^*)}{\lambda} \\ &\leq \lim_{\lambda \downarrow 0} \frac{\lambda f(z) + (1-\lambda)f(x^*) - f(x^*)}{\lambda} = f(z) - f(x^*) < 0.\end{aligned}$$

Recognizing a Local Minimum

These results, which are based on elementary calculus, provide the foundations for unconstrained optimization algorithms. In one way or another, all algorithms seek a point where $\nabla f(\cdot)$ vanishes, namely *stationary point*.

Outline

Overview of Algorithms

Choose a starting point, denote by x_0.

- The user with knowledge about the application and the data set may be in a good position to choose x_0 to be a reasonable estimate of the solution.
- Otherwise, the starting point must be chosen by the algorithm, either by a systematic approach or in some arbitrary manner.

Beginning at x_0, optimization algorithms generate a sequence of iterates

$$x_1, x_2, \ldots$$

that terminate when either no more progress can be made or when it seems that a solution point has been approximated with sufficient accuracy, often measured in $\|\nabla f(x_k)\|$.

Two Strategies: Line Search and Trust Region

- How to move from x_k to the next ?
- Often use information about the function f at x_k, and possibly also information from earlier iterates $x_0, x_1, \cdots, x_{k-1}$.
  - Monotone algorithms: Find a new iterate x_{k+1} with $f(x_{k+1}) < f(x_k)$;
  - Nonmonotone algorithms: Find a new iterate x_{k+1} with $f(x_{k+1}) < f(x_{k-m})$.
- Two strategies for moving from the current point x_k to a new iterate x_{k+1}: *Line Search* and *Trust Region*.

Line Search Strategy

- First choose a direction p_k
- Search along this direction from the current iterate x_k for a new iterate with a lower function value. Simply,

$$x_k \to x_k + \alpha_k p_k$$

  with $f(x_k + \alpha_k p_k) < f(x_k)$.
- At the new point, a new search direction and step length are computed, and the process is repeated.

Line Search Strategy

- How to choose α_k?
- After finding the search direction p_k, the distance to move along p_k can be determined by finding a step length α_k through solving

$$\min_{\alpha>0} \quad f(x_k + \alpha p_k). \tag{3.1}$$

- There are generally two ways to find the step length:
  - By solving (3.1) exactly, we would derive the maximum benefit from the direction p_k, but an exact minimization may be expensive and is usually unnecessary.
  - Instead, the line search algorithm generates a limited number of trial step lengths until it finds one that loosely approximates the minimum of (3.1).

Trust Region Strategy

- The information gathered about f is used to construct a *model function* m_k whose behavior near the current point x_k is similar to that of the actual objective function f.
- Because the model m_k may not be a good approximation of f when x is far from x_k, we restrict the search for a minimizer of m_k to some *trust region* around x_k.

In other words, we find the candidate step p by approximately solving the following sub-problem:

$$\min_{p\in\mathbb{R}^n} m_k(x_k + p), \text{where } x_k + p \text{ lies inside the trust region.} \tag{3.2}$$

If the candidate solution does not produce a sufficient decrease in f, we conclude that the trust region is too large and shrink it to re-solve (3.2).

Trust Region Strategy

- Usually, the trust region subproblem is in the form

$$\min_{p} \quad m_k(x_k + p) = f_k + p^T \nabla f_k + \frac{1}{2} p^T B_k p,$$
$$\text{s.t.} \quad \|p\|_2 \leq \Delta_k,$$

- Here m_k in (3.2) is a quadratic function, which is an approximation to $f(x_k + p)$. Notice from Taylor's theorem that

$$f(x + p) = f(x) + \nabla f(x)^T p + \frac{1}{2} p^T \nabla^2 f(x + tp) p,$$

- The matrix B_k is either the Hessian $\nabla^2 f_k$ or some approximation to it.
- The scalar $\Delta_k > 0$ is called the *trust-region radius*. Elliptical and box-shaped trust regions may also be used.

Comparison between Two Strategies

In a sense, the line search and trust-region approaches differ in the order in which they choose the *direction* and *distance* of the move to the next iterate.

- Line search starts by fixing the direction p_k and then identifying an appropriate distance, namely the step length α_k.
- In trust region, we seek a direction and step that attain the best improvement possible subject to $\|p\| \leq \Delta_k$. If this step proves to be unsatisfactory, we reduce the distance measure Δ_k and try again.

Thanks for your attention!