# A Tour on Optimization Methods

## Unconstrained Minimization - Line Search Methods

Jayadev Naram

IIIT-H

October 17, 2020

# Basic Terminology

## Optimization Problem

Let $f : \mathbb{R}^n \to \mathbb{R}$ be any cost function, and let $f_i : \mathbb{R}^n \to \mathbb{R}$ and $g_i : \mathbb{R}^n \to \mathbb{R}$ be the constraint functions.

$$\begin{aligned} \min_{x \in \mathcal{S}} \quad & f(x) \\ \text{subject to} \quad & f_i(x) = 0, \ i = 1, \cdots, m \\ & g_i(x) \leq 0, \ i = 1, \cdots, n, \end{aligned}$$

where $\mathcal{S}$ is the intersection of domains of all the functions.

This talk will be focused on uncontrained optimization problems such as follows:

$$\min_{x \in \mathbb{R}^n} f(x),$$

where $f$ satisfies some special conditions.

# Basic Terminology

## Types of Minimizers

- A point $x^* \in \mathcal{S}$ is called the **global minimizer** if $f(x^*) \leq f(x)$ for all $x \in \mathcal{S}$.
- A point $x^* \in \mathcal{S}$ is called the **local minimizer** if there is a neighborhood $\mathcal{N} \subseteq \mathcal{S}$ of $x^*$ such that $f(x^*) \leq f(x)\ \forall x \in \mathcal{N}$.
- A point $x^* \in \mathcal{S}$ is called the **strict local minimizer**(also called as strong local minimizer) if there is a neighborhood $\mathcal{N} \subseteq \mathcal{S}$ of $x^*$ such that $f(x^*) < f(x)\ \forall x \in \mathcal{N}, x \neq x^*$.

# Q - Rate of Convergence

Let $\{x_k\}$ be a sequence in $\mathbb{R}^n$ that converges to $x^*$.

- The convergence is said to be **Q-linear** if there is a constant $r \in (0, 1)$ such that

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

- The convergence is said to be **Q-superlinear** if

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

- The convergence is said to be **Q-quadratic** if there is a positive constant $M$, not necessarily less than 1, such that

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

Example: $\{1 + (0.5)^k\}$, $\{1 + k^{-k}\}$, $\{1 + (0.5)^{2^k}\}$ respectively.

# Identifying Minimizers

## First-Order Necessary Condition

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

**Remark**: From the above condition notice that every local minimizer is a stationary point(i.e, point at which gradient of f vanishes).

## Second-Order Necessary Condition

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

## Second-Order Sufficient Condition

Suppose that $\nabla^2 f$ 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 $f$.

# Identifying Minimizers

Adding convexity to the picture yields the following results:

## Optimality Conditions for convex problems

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

## Goal of an Optimization Method

Given an initial point $x_0 \in \mathbb{R}^n$, iteratively find a sequence $\{x_n\}$ at each step such that as the limit tends to infinity, some of the above mentioned conditions of optimality are satisfied.

We now take a look at two broad classes of optimization methods.

# Classification of Optimization Methods

- Line Search Methods
- Trust-Region Methods

# Line Search Methods

**Algorithm** Prototype Line Search Methods

1: Pick an initial point $x_0$
2: **repeat**
3: Pick a direction $p_k$
4: Pick a step length $\alpha_k > 0$ in that direction
5: Update the solution. $x_{k+1} \leftarrow x_k + \alpha_k p_k$
6: **until** Convergence

**Note**: Steps 1,3,4 are non-trivial. Later we'll see why step 1 is non-trivial. Roughly speaking, we want to pick $p_k$ and $\alpha_k$ such that $f(x_{k+1}) < f(x_k)$ except when $x_k$ is optimal. Such methods are called **Descent Methods**. As $-\nabla f(x_k)$ is the steepest descent direction, we define $p_k$ to be a **descent direction** if $\nabla f(x_k)^T p_k < 0$.

# Step Length Selection

Given a descent direction $p_k$, the ideal step length would be

$$\alpha^* = \underset{\alpha>0}{\text{argmin}}\ f(x_k + \alpha_k p_k).$$

Generally, we prefer inexact methods, where we try out a sequence of step lengths until certain conditions are satisfied which guarantee optimality.

# Wolfe Conditions on step length $\alpha_k$

Define $\phi(\alpha_k) = f(x_k + \alpha_k p_k)$ and $l(\alpha_k) = f(x_k) + (c_1 \nabla f(x_k)^T p_k)\alpha_k$, which is a linear function in $\alpha_k$ with negative slope $(c_1 \nabla f(x_k)^T p_k)$.

- **Armijo condition or sufficient decrease condition**

$$f(x_k + \alpha_k p_k) \leq f(x_k) + c_1 \alpha_k \nabla f(x_k)^T p_k \ \textbf{(or)} \ \phi(\alpha_k) \leq l(\alpha_k)$$

- **Curvature condition**

$$\nabla f(x_k + \alpha_k p_k)^T p_k \geq c_2 \nabla f(x_k)^T p_k \ \textbf{(or)} \ \phi'(\alpha_k) \geq c_2 \phi'(0)$$

where $0 < c_1 < c_2 < 1$.

# Wolfe Conditions on step length $\alpha_k$

**Theorem**

*Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable. Let $p_k$ be a descent direction at $x_k$, and assume that $f$ is bounded below along the ray $\{x_k + \alpha p_k : \alpha > 0\}$. Then if $0 < c_1 < c_2 < 1$, there exists intervals of step lengths satisfying the Wolfe conditions.*

# Descent Direction Selection

Let the angle between the steepest descent direction $-\nabla f(x_k)$ and $p_k$ be $\theta_k$, then $cos\ \theta_k = \dfrac{-\nabla f(x_k)^T p_k}{\|\nabla f(x_k)\|\|p_k\|}$.

## Theorem

*Consider any iteration of the form $x_k + \alpha_k p_k$, where $p_k$ is a descent direction and $\alpha_k$ satisfies the Wolfe conditions. Suppose that $f$ is bounded below in $\mathbb{R}^n$ and that $f$ is continuously differentiable in an open set $\mathcal{N}$ containing the level set $\mathcal{L} \equiv \{x :\ f(x) \leq f(x_0)\}$, where $x_0$ is the initial point. Assume also that the gradient $\nabla f$ is Lipschitz continuous on $\mathcal{N}$, i.e, there exists a constant $L > 0$ such that $\|\nabla f(x) - \nabla f(\tilde{x})\| \leq L\|x - \tilde{x}\|$, for all $x, \tilde{x} \in \mathcal{N}$. Then,*

$$\sum_{k \geq 0} cos^2\theta_k \|\nabla f(x_k)\|^2 < \infty \ \textbf{(or)} \ \lim_{k \to \infty} cos^2\theta_k \|\nabla f(x_k)\|^2 = 0.$$

# Steepest Descent Method

**Theorem**

*Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable, and that the iterates generated by steepest descent method (where $p_k = -\nabla f(x_k)$) with the exact line searches converges Q-linearly to a point $x^*$ where the Hessian matrix $\nabla^2 f(x^*)$ is positive definite. Then*

$$f(x_{k+1}) - f(x^*) \leq \left(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}\right)^2 [f(x_k) - f(x^*)],$$

*where $\lambda_1 \leq \cdots \leq \lambda_n$ are the eigenvalues of $\nabla^2 f(x^*)$.*

# Newton's Method

The descent direction $p_k^N \equiv -\nabla^2 f(x_k)^{-1} \nabla f(x_k)$.

## Theorem

*Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is twice differentiable, and that the Hessian $\nabla^2 f(x)$ is Lipschitz continuous in a neighborhood of a solution $x^*$ at which the second-order sufficient conditions hold. Consider the iteration $x_{k+1} = x_k + p_k^N$. Then*

- *if the starting point $x_0$ is sufficiently close to $x^*$, the sequence of iterates converges to $x^*$.*
- *the rate of convergence of $\{x_k\}$ is Q-quadratic; and*
- *the sequence of gradient norms $\{\nabla f(x_k)\}$ converges Q-quadratically to zero.*

# References I

 J. Nocedal and S. Wright. *Numerical Optimization*.