Module 03 — Part 02
Introduction to System Optimization

Dr. Ahmad F. Taha

CE 4240/5240 — Intro to Infrastructure Systems Engineering

Email: ahmad.taha@vanderbilt.edu

Webpage: http://lab.vanderbilt.edu/taha

VANDERBILT UNIVERSITY

September 26, 2024

Module 03 — Part 02 Outline

1. Optimization problems in standard form
2. Optimality conditions
3. Duality and KKT Conditions
4. Convergence

Part III — Optimization problem in standard form

Optimization problem in standard form

$$\begin{aligned} \text{minimize} \quad & f_0(x) \\ \text{subject to} \quad & f_i(x) \leq 0, \quad i = 1, \ldots, m \quad (\textbf{inequality constraints}) \\ & h_i(x) = 0, \quad i = 1, \ldots, p \quad (\textbf{equality constraints}) \end{aligned}$$

- $x = [x_1, x_2, \ldots, x_n]^\top$ is a vector of **unknowns**, also called **variables**
- $f_0(x) : \mathbb{R}^n \to \mathbb{R}$ is the **objective function**, also called **cost function**
- $f_i(x) : \mathbb{R}^n \to \mathbb{R}$ $(i = 1, \ldots, m)$ and $h_i(x) : \mathbb{R}^n \to \mathbb{R}$ are the **constraint functions**
- The set of points that satisfy the constraints is called the **feasible set**

$$C = \{x \in \mathbb{R}^n \mid f_i(x) \leq 0 \ (i = 1, \ldots, m) \text{ and } h_i(x) = 0 \ (i = 1, \ldots, p)\}$$

- If there are no constraints ($m = p = 0$), the problem is called **unconstrained**

Implicit and explicit constraints

Explicit constraints: $f_i(x) \leq 0$ $(i = 1, \ldots, m)$ and $h_i(x) = 0$ $(i = 1, \ldots, p)$
Implicit constraints: $x \in D$, where D is the *domain* of the problem

$$D = \mathrm{dom}(f_1) \cap \ldots \cap \mathrm{dom}(f_m) \cap \mathrm{dom}(h_1) \cap \ldots \cap \mathrm{dom}(h_p)$$

For example, the problem

$$\begin{aligned} \text{minimize} \quad & -\log x_1 - \log x_2 \\ \text{subject to} \quad & x_1 + x_2 - 1 \leq 0 \end{aligned}$$

has implicit constraint

$$x \in D = \{(x_1, x_2) \in \mathbb{R}^2 \,|\, x_1 > 0, x_2 > 0\}$$

Feasibility and optimal value

The **optimal value** of the problem is

$$f^* = \min\{f_0(x) \mid f_i(x) \leq 0 \ (i = 1, \ldots, m) \text{ and } h_i(x) = 0 \ (i = 1, \ldots, p)\}$$

- If one can find at least one x satisfying the constraints (that is, $f_i(x) \leq 0$ for all $i = 1, \ldots, m$, and $h_i(x) = 0$, for all $i = 1, \ldots, p$), the problem is called **feasible**.
- If no such x can be found, then the problem is called **infeasible**. In this case, the feasible set is empty and we can write $f^* = \infty$.
- If the problem is feasible, there are two possibilities for f^*.
  1. $f^* = -\infty$. This means that we can stay in the feasible set and move to a direction that $f_0(x)$ becomes smaller without bound. We say the problem is **unbounded**.
  2. f^* is a number.

Optimal points

$$C = \{x \in \mathbb{R}^n \mid f_i(x) \leq 0 \ (i = 1, \ldots, m) \text{ and } h_i(x) = 0 \ (i = 1, \ldots, p)\}$$
$$f^* = \inf\{f_0(x) \mid f_i(x) \leq 0 \ (i = 1, \ldots, m) \text{ and } h_i(x) = 0 \ (i = 1, \ldots, p)\}$$

- Even if f^* is a number, we don't know if there is an x^* that gives $f_0(x^*) = f^*$
- Consider minimizing $f_0(x) = e^{-x}$; we have $f^* = 0$, but there is no x^*
- If there an $x^* \in C$ that gives $f_0(x^*) = f^*$, we say that x^* is an **optimal point** and that the optimal value is *achieved*
- There may be more than one optimal point
- The set of all optimal points is called the **optimal set** and is

$$X_{\text{opt}} = \{x \in \mathbb{R}^n \mid f_i(x) \leq 0 \ (i = 1, \ldots, m) \text{ and } h_i(x) = 0 \ (i = 1, \ldots, p), f_0(x) = f^*\}$$

Feasibility problems

$$\begin{aligned} \text{minimize} \quad & 0 \\ \text{subject to} \quad & f_i(x) \leq 0, \quad i = 1, \ldots, m \\ & h_i(x) = 0, \quad i = 1, \ldots, p \end{aligned}$$

The problem is really to find an x that satisfies the system of inequalities and equalities given by the constraints.

Convex optimization problem in standard form

$$\begin{aligned} \text{minimize} \quad & f_0(x) \\ \text{subject to} \quad & f_i(x) \leq 0, \quad i = 1, \ldots, m \\ & a_i^\top x - b_i = 0, \quad i = 1, \ldots, p \end{aligned}$$

- $f_0, f_1, \ldots, f_m$: *convex* functions
- Linear equality constraints
- The feasible set of a convex optimization problem is a convex set
- The equality constraints can be organized in matrix form $Ax = b$ where

$$A = \begin{bmatrix} a_1^\top \\ a_2^\top \\ \ldots \\ a_p^\top \end{bmatrix}, \quad b = \begin{bmatrix} b_1 \\ b_2 \\ \ldots \\ b_p \end{bmatrix}$$

Maximization problems

$$\begin{aligned} \text{maximize} \quad & g_0(x) \\ \text{subject to} \quad & f_i(x) \leq 0, \quad i = 1, \ldots, m \\ & h_i(x) = 0, \quad i = 1, \ldots, p \end{aligned}$$

The problem is equivalent to minimizing $-g_0(x)$ over the feasible set.
The problem will be convex if

- $g_0(x)$ is a concave function
- $f_1, \ldots, f_m$ are convex functions
- $h_i(x) = a_i^\top x - b_i$

Likewise, a constraint $g(x) \geq 0$ can be equivalently written as $-g(x) \leq 0$.

Optimality properties of convex problems

$$\begin{aligned} \text{minimize} \quad & f_0(x) \\ \text{subject to} \quad & x \in C \end{aligned}$$

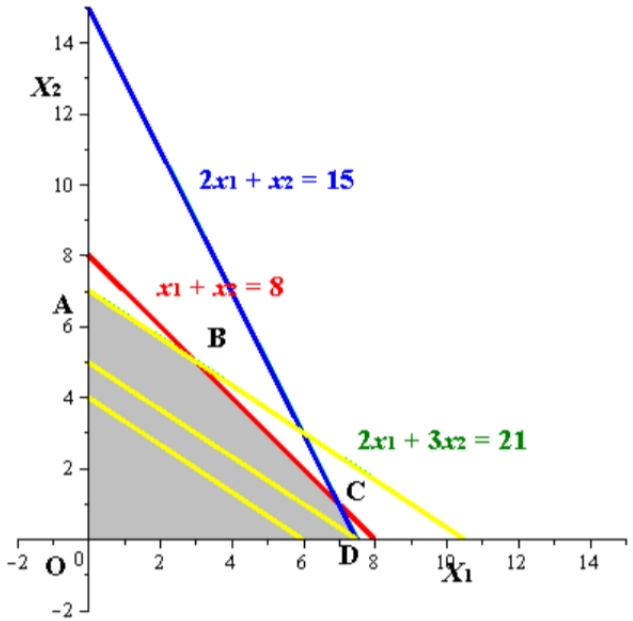
Suppose $f_0(x)$ is convex and C is a convex set. Then the following hold:

1. Any locally optimal point is also globally optimal.
2. If in addition $f_0(x)$ is strictly convex, there can be at most one global minimum.

Can a convex optimization problem have more than one solution?

Solve this problem:

$$\begin{aligned} \max \quad & 4x_1 + 6x_2 \\ \text{subject to} \quad & x_1 + x_2 \leq 8 \\ & 2x_1 + x_2 \leq 15 \\ & 2x_1 + 3x_2 \leq 21 \\ & x_{1,2} \geq 0 \end{aligned}$$



- Why does this happen? Well: lack of *strict* convexity...

Optimality criterion for unconstrained optimization

$$\begin{aligned} \text{minimize} \quad & f_0(x) \\ \text{subject to} \quad & x \in \mathbb{R}^n \end{aligned}$$

Suppose f_0 is differentiable.

- If x^* is a local minimum of f_0, then

$$\nabla f_0(x^*) = 0$$

- If in addition f_0 is convex, then the previous condition is sufficient for optimality

This is a system of n equations with n unknowns. When are these equations linear?

Restriction and relaxation

Original problem with optimal value f^*

$$\begin{aligned} f^* = \text{minimize} \quad & f_0(x) \\ \text{subject to} \quad & x \in C \end{aligned}$$

New problem, with optimal value $\tilde{f}^*$

$$\begin{aligned} \tilde{f}^* = \text{minimize} \quad & f_0(x) \\ \text{subject to} \quad & x \in \tilde{C} \end{aligned}$$

New problem is

- *Relaxation* of original if $\tilde{C} \supset C$, in which case, $\tilde{f}^* \leq f^*$
- *Restriction* of original if $\tilde{C} \subset C$, in which case, $\tilde{f}^* \geq f^*$

Example: If $f_0(x)$ is convex, C is nonconvex, and $\tilde{C} = \text{conv}(C)$, then the relaxation is a convex problem, and gives a lower bound for the original one.

Part V — KKT Conditions

Solving Unconstrained OPs

Objective:

$$\underset{x \in \mathbb{R}^n}{\text{minimize}}\, f(x)$$

Necessary & Sufficinet Conditions for Optimality

x^* is a local minimum of $f(x)$ iff:

1. Zero gradient at x^*:
$$\nabla_x f(x^*) = 0$$
2. Hessian at x^* is positive semi-definite:
$$\nabla_x^2 f(x^*) \succeq 0$$

- For maximization, Hessian is negative semi-definite
- Example:
$$\min f(x) = 4x_1^2 - 4x_1x_2 + 2x_2^2$$
- Another example:
$$\min f(x) = 4x_1^2 - 4x_1x_2 + 2x_2^2 + 10x_1 - 7x_2 + 8$$

Solving Constrained OPs

- **Main objective:** find/compute minimum or a maximum of an objective function subject to equality and inequality constraints
- Formally, problem defined as finding the optimal x^*:

$$\begin{aligned} \min_{x} \quad & f(x) \\ \text{subject to} \quad & g(x) \leq 0 \\ & h(x) = 0 \end{aligned}$$

  - $x \in \mathbb{R}^n$
  - $f(x)$ is scalar function, possibly nonlinear, possibly nonconvex, possibly convex
  - $g(x) \in \mathbb{R}^m, h(x) \in \mathbb{R}^l$ are vectors of constraints

Main Principle

To solve constrained optimization problems: transform constrained problems to unconstrained ones. How? Augment the constraints to the cost function.

KKT Conditions

$$\begin{aligned} \min_{x} \quad & f(x) \\ \text{subject to} \quad & g(x) \leq 0 \\ & h(x) = 0 \end{aligned}$$

- Define the Lagrangian: $\mathcal{L}(x, \lambda, \mu) = f(x) + \lambda^T h(x) + \mu^T g(x)$

Optimality Conditions

The constrained optimization problem (above) has a local minimizer x^* iff there exists a unique μ^* such that:

1. $\nabla_x \mathcal{L}(x^*, \lambda^*, \mu^*) = \nabla_x f(x) + \lambda^{*T} \nabla_x h(x^*) + \mu^{*T} \nabla_x g(x^*) = 0$
2. $\mu_j^* \geq 0$ for $j = 1, \ldots, m$
3. $\mu_j^* g_j(x^*) = 0$ for $j = 1, \ldots, m$
4. $g_j(x^*) \leq 0$ for $j = 1, \ldots, m$
5. $h_i(x^*) = 0$ for $i = 1, \ldots, l$ (if x^*, μ^*, λ^* satisfy 1–5, they are candidates)
6. Second order necessary conditions (SONC): $\nabla_x^2 \mathcal{L}(x^*, \lambda^*, \mu^*) \succeq 0$

KKT Conditions — Example

Find the minimizer of the following optimization problem:

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & f(x) = (x_1 - 1)^2 + x_2 - 2 \\ \text{subject to} \quad & g(x) = x_1 + x_2 - 2 \leq 0 \\ & h(x) = x_2 - x_1 - 1 = 0 \end{aligned}$$

- First, find the Lagrangian function:

$$\mathcal{L}(x, \lambda, \mu) = (x_1 - 1)^2 + x_2 - 2 + \lambda(x_2 - x_1 - 1) + \mu(x_1 + x_2 - 2)$$

- Second, find the conditions of optimality (from previous slide):
  1. $\nabla_x \mathcal{L}(x^*, \lambda^*, \mu^*) = \begin{bmatrix} 2x_1^* - 2 - \lambda^* + \mu^* & 1 + \lambda^* + \mu^* \end{bmatrix}^\top = \begin{bmatrix} 0 & 0 \end{bmatrix}^\top$
  2. $\mu^*(x_1^* + x_2^* - 2) = 0$
  3. $\mu^* \geq 0$
  4. $x_1^* + x_2^* - 2 \leq 0$
  5. $x_2^* - x_1^* - 1 = 0$
  6. $\nabla_x^2 \mathcal{L}(x^*, \lambda^*, \mu^*) = \nabla_x^2 f(x^*) + \lambda^* \nabla_x^2 h(x^*) + \mu^* \nabla_x^2 g(x^*) \succeq 0$

$$= \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix} + \lambda^* \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} + \mu^* \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \succeq 0$$

Example — Cont'd

- To solve the system equations for the optimal x^*, λ^*, μ^*, we first try $\mu^* > 0$.
- Given that, we solve the following set of equations:
  1. $2x_1^* - 2 - \lambda^* + \mu^* = 0$
  2. $1 + \lambda^* + \mu^* = 0$
  3. $x_1^* + x_2^* - 2 = 0$
  4. $x_2^* - x_1^* - 1 = 0$

  $\Rightarrow$ $x_1^* = 0.5, x_2^* = 1.5, \lambda^* = -1, \mu^* = 0$
- But this solution contradicts the assumption that $\mu^* > 0$
- **Alternative:** assume $\mu^* = 0 \Rightarrow x_1^* = 0.5, x_2^* = 1.5, \lambda^* = -1, \mu^* = 0$
- This solution satisfies $g(x^*) \leq 0$ constraint, hence it's a candidate for being a minimizer
- We now verify the SONC: $L(x^*, \lambda^*, \mu^*) = \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix} \succeq 0$
- Thus, $x^* = \begin{bmatrix} 0.5 & 1.5 \end{bmatrix}^\top$ is a strict local minimizer

Other examples

Solve the following optimization problems

$$\text{minimize} \quad x_1 x_2 \quad \text{subject to} \quad x_1^2 + x_2^2 \leq 2, \quad x_{1,2} \geq 0$$

$$\text{minimize} \quad x_1 + x_2 + x_3^2 \quad \text{subject to} \quad x_1 = 1, \quad +x_1^2 + x_2^2 = 1$$

Part VI — Convergence

Optimization algorithm

Rule to provide a sequence $x^{(0)}, x^{(1)}, x^{(2)}, \ldots$

$$x^{(k+1)} = F^{(k)}(x^{(k)})$$

and a termination criterion.

Desirable properties of the algorithm

- If the algorithm is started at the problem solution, then it should not move
- Convergence
  Convergence can be expressed in several ways, e.g.,

$$\lim_{k\to\infty} f(x^{(k)}) = f(x^*)$$

$$\lim_{k\to\infty} x^{(k)} = x^*$$

- Convergence speed
  How fast does the algorithm converge?
  How many iterations does it take to guarantee that $|f(x^{(k)}) - f^*| \leq \varepsilon$?
- Complexity analysis
  How many operations (additions, multiplications, ...) does it take to converge, as a function of the problem size (number of variables, number of constraints)?

Unconstrained minimization algorithms

- Algorithms to solve

$$\text{minimize} \quad f(x)$$

- Interpreted as methods to solve $\nabla f(x) = 0$
- If the problem is convex, convergence to the global minimum
- If the problem is not convex, convergence to a point satisfying $\nabla f(x) = 0$ (which may or may not be local minimum)
- Points satisfying $\nabla f(x) = 0$ are called stationary points

Descent algorithms

$$x^{(k+1)} = x^{(k)} + t^{(k)} \Delta x^{(k)} \text{ with } f(x^{(k+1)}) < f(x^{(k)})$$

- $\Delta x^{(k)}$: Search direction (vector in $\mathbb{R}^n$)
- $t^{(k)} > 0$: Stepsize

General descent algorithm:

Require: Starting point $x \in \mathrm{dom}(f)$
  repeat
    Compute search direction Δx
    Line search: Compute step size t
    Update: $x^+ := x + t\Delta x$
  until convergence criterion is satisfied

Search directions for descent

- To guarantee descent, it is necessary to choose $\Delta x^{(k)}$ so that $\nabla f(x^{(k)})^T \Delta x^{(k)} < 0$
- You can see this geometrically for a simple function $f(x) = x^2$
- For a nonconvex function, use Taylor's 1st-order approximation

$$f(x^{(k+1)}) = f(x^{(k)} + t^{(k)} \Delta x^{(k)}) \approx f(x^{(k)}) + t^{(k)} \nabla f(x^{(k)})^T \Delta x^{(k)}$$

- **Bottomline:** If $\nabla f(x^{(k)})^T \Delta x^{(k)} < 0$ and $t^{(k)}$ is sufficiently small, descent can be guaranteed (even if the function is nonconvex)
- $t^{(k)}$ **sufficiently small is not enough for convergence**
- Proper value of $t^{(k)}$ must be selected through *line search*

Example

Solve this using steepest descent: $\min f(x) = 4x_1^2 - 4x_1x_2 + 2x_2^2$

1. Start with initial guess: $x^{(0)} = [2, 3]^T$
2. Compute gradient at this guess: $\nabla f(x^{(0)}) = \begin{bmatrix} 8x_1 - 4x_2 \\ -4x_1 + 4x_2 \end{bmatrix} = \begin{bmatrix} 4 \\ 4 \end{bmatrix}$
3. Find a t that minimizes $f(x^{(0)} - t\nabla f(x^{(0)})) = f\left(\begin{bmatrix} 2 - 4t \\ 3 - 4t \end{bmatrix}\right)$ or

$$t = \arg\min_{t>0} \underbrace{\left(4(2-4t)^2 - 4(2-4t)(3-4t) + 2(2-4t)^2\right)}_{g(t)}$$

4. To compute an optimal t, take derivative of $g'(t) = 0$ to obtain

$$g'(t) = 64t - 32 = 0 \Rightarrow t^* = 0.5, \quad g''(t) > 0$$

5. Compute next iterate: $x^{(1)} = x^{(0)} - t\nabla f(x^{(0)}) = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$
6. Go to Step 2 and repeat to get

$$x^{(0)} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}, \; x^{(1)} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \; x^{(2)} = \begin{bmatrix} 0.4 \\ 0.6 \end{bmatrix}, \; x^{(3)} = \begin{bmatrix} 0 \\ 0.2 \end{bmatrix}, \; x^{(4)} = \begin{bmatrix} 0.08 \\ 0.12 \end{bmatrix}$$

Search directions

1. **Steepest descent:** $\Delta x^{(k)} = -\nabla f(x^{(k)})$
2. **Newton's method:** $\Delta x^{(k)} = -(\nabla^2 f(x^{(k)}))^{-1} \nabla f(x^{(k)})$
3. **DFP**, **BFGS**, etc..

DFP Method

The **Davidon–Fletcher–Powell formula** (or **DFP**; named after William C. Davidon, Roger Fletcher, and Michael J. D. Powell) finds the solution to the secant equation that is closest to the current estimate and satisfies the curvature condition. It was the first quasi-Newton method to generalize the secant method to a multidimensional problem. This update maintains the symmetry and positive definiteness of the Hessian matrix.

Given a function $f(x)$, its gradient (∇f), and positive-definite Hessian matrix B, the Taylor series is

$$f(x_k + s_k) = f(x_k) + \nabla f(x_k)^T s_k + \frac{1}{2} s_k^T B s_k + \dots,$$

and the Taylor series of the gradient itself (secant equation)

$$\nabla f(x_k + s_k) = \nabla f(x_k) + B s_k + \dots$$

is used to update B.

The DFP formula finds a solution that is symmetric, positive-definite and closest to the current approximate value of B_k:

$$B_{k+1} = (I - \gamma_k y_k s_k^T) B_k (I - \gamma_k s_k y_k^T) + \gamma_k y_k y_k^T,$$

where

$$y_k = \nabla f(x_k + s_k) - \nabla f(x_k),$$
$$\gamma_k = \frac{1}{y_k^T s_k},$$

and B_k is a symmetric and positive-definite matrix.

The corresponding update to the inverse Hessian approximation $H_k = B_k^{-1}$ is given by

$$H_{k+1} = H_k - \frac{H_k y_k y_k^T H_k}{y_k^T H_k y_k} + \frac{s_k s_k^T}{y_k^T s_k}.$$

B is assumed to be positive-definite, and the vectors s_k^T and y must satisfy the curvature condition

$$s_k^T y_k = s_k^T B s_k > 0.$$

The DFP formula is quite effective, but it was soon superseded by the Broyden–Fletcher–Goldfarb–Shanno formula, which is its dual (interchanging the roles of *y* and *s*).[1]

Improved DFP: The BFGS Method

Consider the following unconstrained optimization problem

$$\underset{\mathbf{x}\in\mathbb{R}^n}{\text{minimize}} \quad f(\mathbf{x}),$$

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

From an initial guess $\mathbf{x}_0 \in \mathbb{R}^n$ and an initial guess of the Hessian matrix $B_0 \in \mathbb{R}^{n\times n}$ the following steps are repeated as $\mathbf{x}_k$ converges to the solution:

1. Obtain a direction $\mathbf{p}_k$ by solving $B_k\mathbf{p}_k = -\nabla f(\mathbf{x}_k)$.
2. Perform a one-dimensional optimization (line search) to find an acceptable stepsize α_k in the direction found in the first step. If an exact line search is performed, then $\alpha_k = \arg\min f(\mathbf{x}_k + \alpha\mathbf{p}_k)$. In practice, an inexact line search usually suffices, with an acceptable α_k satisfying Wolfe conditions.
3. Set $\mathbf{s}_k = \alpha_k\mathbf{p}_k$ and update $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{s}_k$.
4. $\mathbf{y}_k = \nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k)$.
5. $B_{k+1} = B_k + \dfrac{\mathbf{y}_k\mathbf{y}_k^{\mathrm{T}}}{\mathbf{y}_k^{\mathrm{T}}\mathbf{s}_k} - \dfrac{B_k\mathbf{s}_k\mathbf{s}_k^{\mathrm{T}}B_k^{\mathrm{T}}}{\mathbf{s}_k^{\mathrm{T}}B_k\mathbf{s}_k}$.

Convergence can be determined by observing the norm of the gradient; given some $\epsilon > 0$, one may stop the algorithm when $||\nabla f(\mathbf{x}_k)|| \leq \epsilon$. If B_0 is initialized with $B_0 = I$, the first step will be equivalent to a gradient descent, but further steps are more and more refined by B_k, the approximation to the Hessian.

The first step of the algorithm is carried out using the inverse of the matrix B_k, which can be obtained efficiently by applying the Sherman–Morrison formula to the step 5 of the algorithm, giving

$$B_{k+1}^{-1} = \left(I - \frac{\mathbf{s}_k\mathbf{y}_k^T}{\mathbf{y}_k^T\mathbf{s}_k}\right)B_k^{-1}\left(I - \frac{\mathbf{y}_k\mathbf{s}_k^T}{\mathbf{y}_k^T\mathbf{s}_k}\right) + \frac{\mathbf{s}_k\mathbf{s}_k^T}{\mathbf{y}_k^T\mathbf{s}_k}.$$

This can be computed efficiently without temporary matrices, recognizing that B_k^{-1} is symmetric, and that $\mathbf{y}_k^{\mathrm{T}}B_k^{-1}\mathbf{y}_k$ and $\mathbf{s}_k^{\mathrm{T}}\mathbf{y}_k$ are scalars, using an expansion such as

$$B_{k+1}^{-1} = B_k^{-1} + \frac{(\mathbf{s}_k^{\mathrm{T}}\mathbf{y}_k + \mathbf{y}_k^{\mathrm{T}}B_k^{-1}\mathbf{y}_k)(\mathbf{s}_k\mathbf{s}_k^{\mathrm{T}})}{(\mathbf{s}_k^{\mathrm{T}}\mathbf{y}_k)^2} - \frac{B_k^{-1}\mathbf{y}_k\mathbf{s}_k^{\mathrm{T}} + \mathbf{s}_k\mathbf{y}_k^{\mathrm{T}}B_k^{-1}}{\mathbf{s}_k^{\mathrm{T}}\mathbf{y}_k}.$$

Therefore, in order to avoid any matrix inversion, the **inverse** of the Hessian can be approximated instead of the Hessian itself: $H_k \overset{\text{def}}{=} B_k^{-1}$.[9]

From an initial guess $\mathbf{x}_0$ and an approximate **inverted** Hessian matrix H_0 the following steps are repeated as $\mathbf{x}_k$ converges to the solution:

1. Obtain a direction $\mathbf{p}_k$ by solving $\mathbf{p}_k = -H_k\nabla f(\mathbf{x}_k)$.
2. Perform a one-dimensional optimization (line search) to find an acceptable stepsize α_k in the direction found in the first step. If an exact line search is performed, then $\alpha_k = \arg\min f(\mathbf{x}_k + \alpha\mathbf{p}_k)$. In practice, an inexact line search usually suffices, with an acceptable α_k satisfying Wolfe conditions.
3. Set $\mathbf{s}_k = \alpha_k\mathbf{p}_k$ and update $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{s}_k$.
4. $\mathbf{y}_k = \nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k)$.
5. $H_{k+1} = H_k + \dfrac{(\mathbf{s}_k^{\mathrm{T}}\mathbf{y}_k + \mathbf{y}_k^{\mathrm{T}}H_k\mathbf{y}_k)(\mathbf{s}_k\mathbf{s}_k^{\mathrm{T}})}{(\mathbf{s}_k^{\mathrm{T}}\mathbf{y}_k)^2} - \dfrac{H_k\mathbf{y}_k\mathbf{s}_k^{\mathrm{T}} + \mathbf{s}_k\mathbf{y}_k^{\mathrm{T}}H_k}{\mathbf{s}_k^{\mathrm{T}}\mathbf{y}_k}$.

Stepsize selection methods

1. Exact line search: t solves $\min_{t>0} f(x + t\Delta x)$
   - Inexact line search: Minimize a quadratic or cubic interpolation of $g(t) = f(x + t\Delta x)$ with respect to t
2. **Backtracking line search (Armijo rule)**
   - **In contrast to exact line search, backtracking only needs function evaluations**
3. Constant stepsize: $t^{(k)} = t > 0$ for all $k = 0, 1, 2, \ldots$
   - Simple choice: no need for minimization as with the exact line search, no need for function evaluations as with backtracking
   - Convergence to stationary point under certain conditions
   - In practice, set t to a small value; in case of divergence, reduce t

Backtracking line search

Parameters: $0 < \beta < 1$, $0 < \alpha < 1/2$

- Start with $t = 1$
- If $f(x + t\Delta x) < f(x) + \alpha t \nabla f(x)^T \Delta x$, accept t and exit
- Otherwise, set $t := \beta t$, and check the condition again
- As long as $\nabla f(x)^T \Delta x < 0$, the set of acceptable stepsizes t will always contain an interval of the form $[0, \delta]$, and the line search will finish in a finite number of steps

Convergence of gradient methods

- Generally, convergence to a *stationary point* is guaranteed when
  1. exact/inexact line minimization or backtracking line search is used, and
  2. the search direction is chosen so that $\nabla f(x^{(k)})^T \Delta x^{(k)} < 0$
- Additional conditions on f can guarantee convergence under the (simpler) constant or dimishing stepsize rules.

Convergence: Convexity vs. nonconvexity

- The statements on the previous slide apply to convex as well as nonconvex problems
- For convex problems, stationary points ($\nabla f(x) = 0$) are optimal
- For nonconvex problems, convergence to a stationary point is guaranteed (under conditions)
- For nonconvex problems, the convergence point may be local minimum and in general depends on the initialization point x^0
- Running the algorithm with multiple initializations and selecting the convergence point that yields the smallest objective value is an option in this case

Convergence (termination) criteria

1. Scale-free
   - $\|\nabla f(x^{(k)})\| \leq \epsilon \|\nabla f(x^{(0)})\|$
   - $\|x^{(k)} - x^{(k-1)}\| \leq \epsilon \|x^{(k)}\|$
2. Scale-dependent
   - $\|\nabla f(x^{(k)})\| \leq \epsilon$
   - $\|x^{(k)} - x^{(k-1)}\| \leq \epsilon$

- Scale-free are preferable to scale-dependent
- Setting of ϵ may require trial and error
- Algorithm specific criteria also exist

Newton's method

- Newton direction

$$\Delta x = -(\nabla^2 f(x))^{-1} \nabla f(x)$$

- If $\nabla^2 f(x) \succ 0$, then Δx is a descent direction

$$-\nabla f(x)^T (\nabla^2 f(x))^{-1} \nabla f(x) < 0$$

  unless $\nabla f(x) = 0$, which implies that x is a stationary point

Newton's method:

Require: Starting point $x \in \mathrm{dom}(f)$

  repeat

    Compute search direction: Solve the linear system $\nabla^2 f(x) \Delta x = -\nabla f(x)$

    Line search: Compute step size t by line search

    Update: $x^+ := x + t\Delta x$

  until convergence criterion is satisfied

Example: minimizing the Powell function

$$f(x_1, x_2, x_3, x_4) = (x_1 + 10x_2)^2 + 5(x_3 - x_4)^2 + (x_2 - 2x_3)^4 + 10(x_1 - x_4)^4$$

$$x^0 = [3, -1, 0, 1]^\top$$

Solution: We are going to do three iterations with fixed step size $t = 1$. Note that $f(x^0) = 215$, and we have

$$\nabla f(x) = \begin{bmatrix} 2(x_1 + 10x_2) + 40(x_1 - x_4)^3 \\ 20(x_1 + 10x_2) + 4(x_2 - 2x_3)^3 \\ 10(x_3 - x_4) - 8(x_2 - 2x_3)^3 \\ -10(x_3 - x_4) - 40(x_1 - x_4)^3 \end{bmatrix}$$

And

$$\nabla^2 f(x) = \begin{bmatrix} 2 + 120(x_1 - x_4)^2 & 20 & 0 & -120(x_1 - x_4)^2 \\ 20 & 200 + 12(x_2 - 2x_3)^2 & -24(x_2 - 2x_3)^2 & 0 \\ 0 & -24(x_2 - 2x_3)^2 & 10 + 48(x_2 - 2x_3)^2 & -10 \\ -120(x_1 - x_4)^2 & 0 & -10 & 10 + 120(x_1 - x_4)^2 \end{bmatrix}$$

Iteration 1

$$\nabla f(x^0) = [306, -144, -2, -310]^\top$$

$$\nabla^2 f(x^0) = \begin{bmatrix} 482 & 20 & 0 & -480 \\ 20 & 212 & -24 & 0 \\ 0 & -24 & 58 & -10 \\ -480 & 0 & -10 & 490 \end{bmatrix},$$

$$(\nabla^2 f(x^0))^{-1} = \begin{bmatrix} 1126 & -0.0089 & 0.0154 & 0.1106 \\ -0.0089 & 0.0057 & 0.008 & -0.0087 \\ 0.0154 & 0.008 & 0.0203 & 0.0155 \\ 0.1106 & -0.0087 & 0.0155 & 0.1107 \end{bmatrix} \Rightarrow \textbf{MATLAB}$$

$$\Delta x^0 = [1.4127, -0.8413, -0.254, -0.746]^\top, \text{ hence}$$

$$x^1 = x^0 - \Delta x^0 = [1.5873, -0.1587, 0.254, 0.254]^\top,$$

$$f(x^1) = 31.8$$

Iteration 2

Calculate $\nabla f(x^1), \nabla^2 f(x^1), (\nabla^2 f(x^1))^{-1}$

$$\Delta x^1 = [0.5291, -0.0529, 0.0846, 0.0846]^\top, \text{ Hence,}$$
$$x^2 = x^1 - \Delta x^1 = [1.0582, -0.1058, 0.1694, 0.1694]^\top,$$
$$f(x^2) = 6.28$$

Iteration 3

Calculate $\nabla f(x^{(2)}), \nabla^2 f(x^{(2)}), (\nabla^2 f(x^{(2)}))^{-1}, \Delta x^{(2)}$

$$x^{(3)} = x^{(2)} - \Delta x^{(2)} = [0.7037, -0.0704, 0.1121, 0.1111]^\top,$$
$$f(x^{(3)}) = 1.24$$

Levenberg-Marquardt Modification of Newton's method

- Newton's method can fail if $\nabla^2 f(x)$ is singular (Newton direction is not defined) or $\nabla^2 f(x) \prec 0$ (Newton direction is not a descent direction)
- Newton's method: $x^{(k+1)} = x^{(k)} - t^{(k)}[\nabla^2 f(x^{(k)})]^{-1}\nabla f(x^{(k)})$
- Levenberg-Marquardt modification:
  $x^{(k+1)} = x^{(k)} - t^{(k)}[\mu_k I + \nabla^2 f(x^{(k)})]^{-1}\nabla f(x^{(k)})$ with $\mu_k > 0$
- Search direction $\Delta x^{(k)} = -[\mu_k I + \nabla^2 f(x^{(k)})]^{-1}\nabla f(x^{(k)})$ can always be made to be a descent direction for sufficiently large μ_k
- Method approaches the behavior of Newton's method for $\mu_k \to 0$ and of a gradient method for $\mu_k \to \infty$
- Implementation aspects
  - Start with a small value of μ_k and increase it slowly until the iteration is descent: $f(x^{(k+1)}) < f(x^{(k)})$
  - Search direction computed as solution to $(\mu I + \nabla^2 f(x))\Delta x = -\nabla f(x)$
  - Combine with line search

Some observations

- Newton's method has several advantages over gradient and steepest descent
  - Convergence of Newton's method is rapid in general and quadratic near x^*
  - Once quadratic convergence phase is reached, at most six or so iterations are required to produce a solution
  - Newton's method is *affine invariant*: it is insensitive to the choice of coordinates, or the condition number of the sublevel sets of the objective
  - Newton's method scales well with problem size: its performance on problems in $\mathbb{R}^{10000}$ is similar to its performance on problems in $\mathbb{R}^{10}$, with only a small increase in the number of steps required
  - The good performance of Newton's method is not dependent on the choice of algorithm parameters
  - In contrast, the choice of norm for steepest descent plays a critical role in its performance
- Main disadvantage: the cost of forming and storing the Hessian, cost of computing the Newton step, which requires solving a set of linear equations.
- But you can exploit problem structure

Conjugate Gradient Algorithms

- How about an algorithm that is an intermediate between the method of steepest descent and Newton's method?
  - Solves quadratics of n variables in n steps
  - Requires no Hessian matrix evaluations
  - No matrix inversion and no storage of an $n \times n$ matrix required
- This method performs better than steepest descent, but not as well as Newton's
- The crucial factor in the efficiency of an iterative search method is the direction of search
- Won't be covered here in this class

Module Summary (wohoooo we're done)

What did we learn?

Questions And Suggestions?

Thank You!

Please visit

https://lab.vanderbilt.edu/taha/

IFF you want to know more ☺