# MATH 560 - Optimization

Aram-Alexandre Pooladian
Department of Mathematics
McGill University
Montreal, QC, Canada

September 22, 2018

# Contents

# Chapter 1

# Introduction

## 1.1 Basics

A basic optimization problem follows the following structure:

$$\min_x \ f(x) \tag{1.1}$$

where $x \in \mathbb{R}^n$ and $f$ is a function $f : \mathbb{R}^n \to \mathbb{R}$. The above problem is unconstrained (such an example is the "Least Squares" Problem). If (1) is subject to one or both of the following types of constraints, it becomes a constrained problem.

$$c_i(x) = 0 \quad i \in \mathcal{E}$$
$$c_i(x) \geq 0 \quad i \in \mathcal{I}$$

**Example 1.1.1.**

$$\min f(x) = (x_1 - 2)^2 + (x_2 - 1)^2$$

where $x = [x_1, x_2]^T$ such that

$$x_1^2 - x_2 \leq 0 \text{ and } x_1 + x_2 \leq 2$$

thus

$$c(x) = [c_1(x), c_2(x)]^T = [-x_1^2 + x_2, -x_1 - x_2 + 2]^T, \ \mathcal{I} = \{1, 2\}, \ \mathcal{E} = \emptyset$$

**Definition 1.1.1.** The **feasible region** is the set of points satisfying all the constraints. The **infeasible region** follows.

- If a problem is overspecified with too many constraints, the problem becomes **infeasible** and the feasible set is empty.

# Chapter 2

# Fundamentals of Unconstrained Optimization

## 2.1 What is a Solution?

**Definition 2.1.1.** A point $x^\star$ is a **global minimizer** if $f(x^\star) \leq f(x)\ \forall x \in \mathbb{R}^n$

- Global minimizer is difficult to find (usually) because our knowledge of $f$ is usually locally known or defined.

**Definition 2.1.2.** A point $x^\star$ is a **local minimizer** if there is a neighborhood $\mathcal{N}$ of $x^\star$ such that $f(x^\star) \leq f(x)\ \forall x \in \mathcal{N}$. It is often called a *weak* local minimizer.

- Note: $\mathcal{N}$ of $x^\star = V_\epsilon(x^\star) = ]x^\star - \epsilon, x^\star + \epsilon[$

**Definition 2.1.3.** A point $x^\star$ is a **strict local minimizer** if there is a neighborhood $\mathcal{N}$ of $x^\star$ such that $f(x^\star) < f(x)\ \forall x \in \mathcal{N} \setminus x^\star$.

**Definition 2.1.4.** A point $x^\star$ is a **isolated local minimizer** if there is a neighborhood $\mathcal{N}$ of $x^\star$ such that $x^\star$ is the only local minimizer in $\mathcal{N}$. All isolated local minimizers are strict but not all strict local minimizers are isolated.

**Example 2.1.1.** $f(x) = 1 \Rightarrow$ any point $x \in \mathbb{R}^n$ is both a local and a global minimizer

**Example 2.1.2.** $f(x) = (x-2)^2 \Rightarrow$ any point $x^\star = 2$ is a strict local minimizer as well as a global minimizer.

**Example 2.1.3.** There are scenarios where a strict local minimizer is **not** isolated. For example, if we take $f(x) = x^4 \cos(1/x) + 2x^4$ and $f(0) = 0$, which is twice continuously differentiable ($f \in C^2$) and has a strict local minimizer at $x^\star = 0$. However, there are strict local minimizers at many nearby points $x_j$ and we can label these points such that $x_j \to 0$ as $j \to \infty$

**Example 2.1.4.** $f(x) = 1 \Rightarrow$ any point $x \in \mathbb{R}^n$ is both a local and a global minimizer

## 2.2 Recognizing a Local Minimum

**Theorem 2.2.1.** *(Taylor's Theorem) Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is continously differentiable and that $p \in \mathbb{R}^n$. Then,*

$$f(x+a) = f(x) + \nabla f(x+at)^T a \tag{2.1}$$

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

$$\nabla f(x+a) = \nabla f(x) + \int_0^1 \nabla^2 f(x+at) a dt \tag{2.2}$$

*and finally:*

$$f(x+a) = f(x) + \nabla f(x)^T a + \frac{1}{2} a^T \nabla^2 f(x+at) a \tag{2.3}$$

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

**Theorem 2.2.2.** *(First Order Necessary Conditions) If $x^\star$ is a local minimizer and $f$ is continuously differentiable in an open neighborhood of $x^\star$, then $\nabla f(x^\star) = 0$*

*Proof.* Suppose not. Then we have that $\nabla f(x^\star) \neq 0$ but $x^\star$ is still a local minimizer (simply derive a contradiction). If we define the $a = -\nabla f(x^\star)$, then the expression $a^T \nabla f(x^\star) = -\|\nabla f(x^\star)\|^2 < 0$. Since $\nabla f$ is continous in the neighborhood of $x^\star$, there is a scalar value $\tau > 0$ such that

$$a^T \nabla f(x^\star + at) < 0 \quad \forall t \in [0, \tau]$$

So for any $\bar{t} \in ]0, \tau]$, using the first statement from *Taylor's Theorem* and by letting $a \equiv a\bar{t}$

$$f(x^\star + a\bar{t}) = f(x^\star) + (a\bar{t})^T \nabla f(x^\star + t(a\bar{t})) = f(x^\star) + \bar{t} a^T \nabla f(x^\star + at), \qquad \text{for some } t \in ]0, \bar{t}[$$

where in the last equality, the $t$ value that was initially between $]0,1[$ absorbed the value $\bar{t}$. Therefore, since the last term is negative, $f(x^\star + a\bar{t}) < f(x^\star)$, but $x^\star$ was supposed to be the local minimizer; we have a contradiction. □

**Definition 2.2.1.** A point $x^\star$ is called a **stationary point** if $\nabla f(x^\star) = 0$ and, by the previous theorem, any local minimizer *must* be a stationary point.

**Example 2.2.1.** $f(x) = x^3 \to x^\star = 0$ has a saddle point and, consequently, not a minimum. Thus the condition is necessary but not sufficient!

**Theorem 2.2.3.** *(Second Order Necessary Conditions) If $x^\star$ is a local minimizer of $f$ and $\nabla^2 f$ exists and is continuous in an open neighborhood of $x^\star$, then $\nabla f(x^\star) = 0$ and $\nabla^2 f(x^\star)$ is positive semidefinite*

*Proof.* Suppose not. Then we have that $\nabla^2 f(x^\star)$ is not positive semidefinite. If we define the vector $a$ such that $a^T \nabla^2 f(x^\star) a < 0$, and since $\nabla^2 f$ is continuous near $x^\star$, there is a scalar $\tau > 0$ such that

$$a^T \nabla^2 f(x^\star + at) a < 0 \quad \forall t \in [0, \tau]$$

Performing the Taylor Series Expansion around $x^\star$, for all $\bar{t} \in ]0, \tau]$ and for some $t \in ]0, \bar{t}$ that

$$f(x^\star + a\bar{t}) = f(x^\star) + (a\bar{t})^T \nabla f(x^\star) + \frac{1}{2}(a\bar{t})^T \nabla^2 f(x^\star + at)(a\bar{t}) = f(x^\star) + \bar{t} a^T \nabla f(x^\star) + \frac{1}{2}\bar{t}^2 a^T \nabla^2 f(x^\star + at) a$$

where the middle term is 0 and the last term is negative thus, once again, we have found a direction away from $x^\star$ such that $f$ is still decreasing, thus $x^\star$ is no longer the local minimizer. □

**Theorem 2.2.4.** *(Second Order Sufficient Conditions) Suppose that $\nabla^2 f$ is continuous in an open neighborhood of $x^\star$ and that $\nabla f(x^\star) = 0$ and $\nabla^2 f(x^\star)$ is positive definite. Then $x^\star$ is a strict local minimizer of $f$*

*Proof.* Since $\nabla^2 f$ is continuous and positive definite when evaluated at $x^\star$, we can choose a radius $\rho > 0$ so that $\nabla^2 f(x)$ remains positive definite $\forall x \in \mathcal{D} = \{z : \|z - x^\star\| < \rho\}$, an open ball. If we take a non-zero vector $a$ such that $\|a\| < \rho$ (contained in the open ball), we then have $x^\star + a \in \mathcal{D}$, thus we can define $z = x^\star + at$, where $t \in ]0, 1[$ and so:

$$\begin{aligned} f(x^\star + a) &= f(x^\star) + a^T \nabla f(x^\star) + \frac{1}{2} a^T \nabla^2 f(z) a \\ &= f(x^\star) + \frac{1}{2} a^T \nabla^2 f(z) a \end{aligned}$$

where the second term is positive definite because $z \in \mathcal{D}$. This ultimately says that $f(x^\star + a) > f(x^\star)$, which is what we wanted to show. □

**Remark** Note that the previous theorem guarantees something stronger than the necessary conditions; namely that the minimizer is a *strict* local minimizer. Also, the second-order sufficient conditions are **not** necessary. That is to say: a point $x^\star$ may be a strict local minimizer and yet might fail the sufficient conditions.

**Example 2.2.2.** Take $f(x) = x^4$, thus $x^\star = 0$ as $\nabla f(x^\star) = 0$ but also the Hessian matrix vanishes, thus no longer positive definite

**Recap**

- **If** $x^\star$ is a local minimizer, then $\nabla f(x^\star) = 0$ and $\nabla^2 f(x^\star) \succeq 0$ (i.e. positive semidefinite)
- **If** we have a point $x^\star$ such that $\nabla f(x^\star) = 0$ and $\nabla^2 f(x^\star) \succ 0$ (i.e. positive definite), **then** $x^\star$ is a strict local minimizer

**Remark on Positive Definite and Semidefinite Matrices**

**Definition 2.2.2.** A matrix $B$ is positive definite if $x^T B x > 0,\ \forall x \in \mathbb{R}^n, x \neq 0$. Likewise, it is positive semidefinite if $x^T B x \geq 0,\ \forall x \in \mathbb{R}^n$

**Remark** In either case, $B$ is a symmetric and square matrix. Note that, if a function $f$ is twice continuously differentiable, then the Hessian matrix is symmetric. If a matrix is symmetric, it has what is called an *eigenvalue decomposition*.

**Definition 2.2.3.** If a matrix $A$ is symmetric, it has an **eigenvalue decomposition**, which means it can have the form:

$$A = U \Lambda U^{-1} = U \Lambda U^T$$

where $\Lambda$ is a diagonal matrix composed of the eigenvalue of $A$, and $U$ is an *orthogonal* matrix (both of these matrices are in $\mathbb{R}^{nxn}$). Recall that a matrix is orthogonal when $UU^T = U^T U = I \Rightarrow U^T = U^{-1}$, which implies the columns are orthonormal.

**Remarks**

- $A^{-1} = U \Lambda^{-1} U^T$ exists if $\lambda_i \neq 0$ for all $i$ (since $\Lambda$ needs to be invertible).
- If $A \succeq 0$, then $A^{-1} \succeq 0$
- If $A \succ 0$, then all its eigenvalues are positive. To see this:

$$0 < x^T A x = x^T (U \Lambda U^T) x = (U^T x)^T \Lambda (U^T x) \equiv y^T \Lambda y = \sum_{i=1}^{n} \lambda_i \cdot (y_i)^2 \Rightarrow \lambda_i > 0 \ \forall i$$

### 2.2.1 Convexity

**Definition 2.2.4.** A function $f$ is **convex** if $\forall x, y \in \mathbb{R}^n$

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y) \quad \forall \alpha \in [0,1]$$

**Definition 2.2.5.** A function $f$ is **concave** if $-f$ is convex

**Example 2.2.3.** $f(x) = a^T x + b \quad \forall b \in \mathbb{R} \Rightarrow$ is both concave and convex (also linear)

**Example 2.2.4.** $f(x) = x^T Q x + q^T x + b$ is convex if $Q \succeq 0$ (again, $\forall b \in \mathbb{R}$)

**Remark** Shifting the function by a constant does not change whether or not it is convex. It only changes the value of $f(x^\star)$ but not the actual location of the minimizer

**Theorem 2.2.5.** *When $f$ is convex, any local minimizer $x^\star$ is also a global minimizer of $f$.*

*Proof.* (In Lecture 3, a proof-by-picture was given) Suppose that $x^\star$ is a local minimizer while not being a global minimizer. The we can find a point $z \in \mathbb{R}^n$ with $f(z) < f(x^\star)$. Consider the line segment that joins both $x^\star$ and $z$:

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

And by the convexity property for $f$, we have:

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

Thus, any neighborhood $\mathcal{N}$ of $x^\star$ contains of the piece of the line segment, so there will always be points $x \in \mathcal{N}$ such that $f(x) < f(x^\star)$. Hence, $x^\star$ is not a local minimizer. Contradiction □

**Theorem 2.2.6.** *If $f$ is convex, any stationary point is also a global minimizer.*

*Proof.* We start by implementint the convexity of $f$, namely:

$$f(\alpha x + (1-\alpha)y) = f(y + \alpha(x-y)) \leq \alpha f(x) + (1-\alpha)f(y)$$

By rearranging a little, we get:

$$f(x) \geq f(y) + \frac{f(y + \alpha(x-y)) - f(y)}{\alpha}$$

And taking the limit of $\alpha \to 0$ of the right-most term gives the directional derivative of $f$:

$$f(x) \geq f(y) + \nabla f(y)^T (x-y)$$

However, let $y = x^\star$ be a stationary point. Then then right-most term disappears and what remains is the following contradiction:

$$f(x) \geq f(x^\star) \quad \forall x$$

□

### 2.2.2 Recognizing Convex Functions

There are multiple ways to notice if a function is convex (Boyd and Vandenburgh — Chp 3). Listed are a couple of ways:

1. By the definition
2. $f$ is convex iff $f(x) \geq f(y) + \nabla f(y)^T (x-y)$
3. $f$ is convex iff $\nabla^2 f(x) \succeq 0 \ \forall x$

4. $f$ is convex iff, $\forall x, v \in \mathbb{R}^n$, the function $g(t) = f(x + vt)$ is convex in $t$ (an example follows).

**Example 2.2.5.** $f(x) = x^T Q x + q^T x$, where $Q \succeq 0$. We fix $x_0, v \in \mathbb{R}^n$. Applying the method:

$$\begin{aligned} g(t) = f(x_0 + vt) &= (x_0 + vt)^T Q(x_0 + vt) + q^T(x_0 + vt) \\ &= x_0^T Q x_0 + 2t(x_0^T Q v) + t^2(v^T Q v^T) + q^T x_0 + t(q^T v) \end{aligned}$$

where the second term arises because of symmetry. Note that this is analogous to having the function $g$ having the form $g(t) = at^2 + bt + c$, which is convex iff $a \geq 0$. In this case, $a \equiv v^T Q v \geq 0$ because $v$ is fixed and $Q \succeq 0$

## 2.3 Overview of Optimization Algorithms

### 2.3.1 Line Search Method

The simple line search method has the following form:

$$x_{k+1} = x_k + \alpha_k p_k$$

where $p_k$ is in $\mathbb{R}^n$ is the **search direction** and $\alpha_k$ is a scalar known as the **step length**. In the *line search* strategy, the algorithm chooses a direction $p_k$ and searches along this direction from the current iterate $x_k$ for a new iterate with a lower function value.

#### Gradient Descent

Let $z = x_k + \alpha p_k$, then:

$$f(z) = f_k + \alpha \nabla f_k^T p_k + \frac{1}{2}\alpha^2 p_k^T \nabla^2 f_z p_k \tag{2.4}$$

and if $\alpha < 1$, the last term becomes negligible and the second term dominates. To observe the greatest decrease (i.e. largest negative number) we do:

$$\min_{p \in \mathbb{R}^n} \nabla f_k^T p \quad \text{such that } \|p\| = 1$$

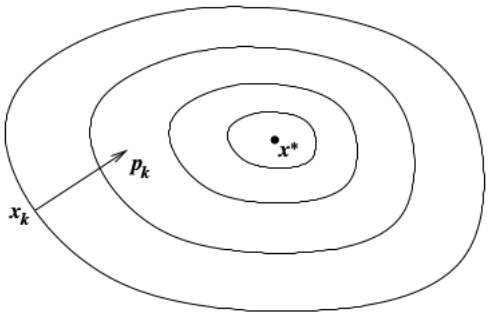
and, by definition of a the dot product, becomes:

$$\min_{p \in \mathbb{R}^n} \|\nabla f_k^T\| \cdot \|p\| \cos(\theta) \quad \text{such that } \|p\| = 1$$

which is minimized when $\cos(\theta) = -1 \Rightarrow \theta = \pi \Rightarrow$ when $\nabla f_k$ is opposite in direction to $p$. To then normalize $p$, we are left with:

$$p = \frac{-\nabla f_k}{\|\nabla f_k\|}$$

Figure 2.1: Steepest descent direction for a function of two variables



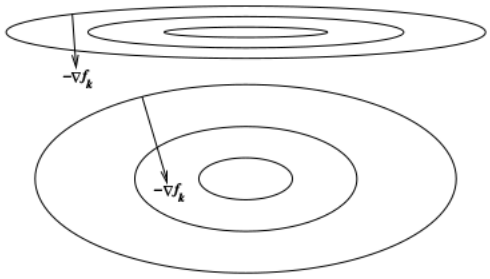



Figure 2.2: Poorly scaling and well scaled problems, and performance of the steepest descent direction

**Scaling** The performance of an algorithm may depend crucially on how the problem is formulated. One important issue is the *scaling* of the problem. An example is minimizing the function $f(\vec{x}) = 10^9x^2 + y^2$, which is very sensitive to changes in $x$ but virtually none in $y$

**Newton's Method**

Given a smooth, not-necessarily entirely convex function, one can approximate a point $f(x_k)$ using Taylor's theorem:

$$f(x_k + a) \simeq f_k + \nabla f_k^T a + \frac{1}{2} a^T \nabla^2 f_k a \equiv m_k(a)$$

where $m_k(a)$ is called the *quadratic model*. Assuming for the moment that $\nabla^2 f_k \succ 0$, we obtain the Newton direction by finding the vector $a$ that minimizes the quadratic model. By setting the derivative to zero and solving, we get:

$$a_k^N = -(\nabla^2 f_k)^{-1} \nabla f_k$$

**Remark** The difference between the quadratic model above and the actual Taylor expansion is the last term, which is evaluated at $x_k$ in the model but at $x_k + at$ in the real Taylor expansion. If $\nabla^2 f$ is sufficiently smooth, this difference introduces a perturbation of only $O(\|a\|^3)$ into the expansion, so that when $\|a\|$ is small, the approximation $f(x_k + a) \simeq m_k(a)$ is quite accurate

# Chapter 3

# Line Search Methods

Each iteration of a line search method computes a search direction $p_k$ and then decides how far along to move along that direction. The iteration is given by

$$x_{k+1} = x_k + \alpha_k p_k \tag{3.1}$$

as mentioned earlier. The success of a line search method depends on effective choices of both the direction and step length.
Most algorithms require $p_k$ to be a *descent direction*, such that $p_k^T \nabla f_k < 0$ which guarantees that the function $f$ can be reduced along this direction (until the gradient is zero). Moreover, the search direction often has the form

$$p_k = -B_k^{-1} \nabla f_k \tag{3.2}$$

where $B_k$ is a symmetric, nonsingular matrix. In the *Gradient Descent* method, $B_k = I_k$ while in the *Newton* method, $B_k = \nabla^2 f_k$ and finally, if $B_k \simeq \nabla^2 f_k \succ 0$, this is the *Quasi-Newton* method. If $p_k$ is defined as above and $B_k \succ 0$, then:

$$p_k^T \nabla f_k = -\nabla f_k^T B_k^{-1} \nabla f_k < 0 \tag{3.3}$$

thus $p_k$ is a descent direction. This chapter discusses how to choose $\alpha_k$ and $p_k$ to promote convergence from remote starting positions.

## 3.1 Step Length (Selection)

In computing the step length $\alpha_k$, we face a tradeoff. We want to substantially reduce $f$ but not spend too much time making this decision. Ideally, we want to pick the global minimizer of the function:

$$\phi(\alpha) = f(x_k + \alpha p_k), \quad \alpha > 0 \tag{3.4}$$

but in general, it is too expensive to identify this value.
Typical line search algorithms try out a sequence of candidate values for $\alpha$, stopping to accept one of these values when certain conditions are satisfied (this gets complicated so we'll just start with the basics).
We now discuss various termination conditions for line search algorithms and show that effective step lengths need **not** lie near minimizers of $\phi(\cdot)$ defined above.
A simple condition to impose on $\alpha_k$ is that it **must** require a reduction in $f$, that is:

$$f(x_k + \alpha_k p_k) < f(x_k)$$

however this is not sufficient enough for convergence. (write textbook example) To have convergence, we need to enforce a *sufficient decrease* condition.

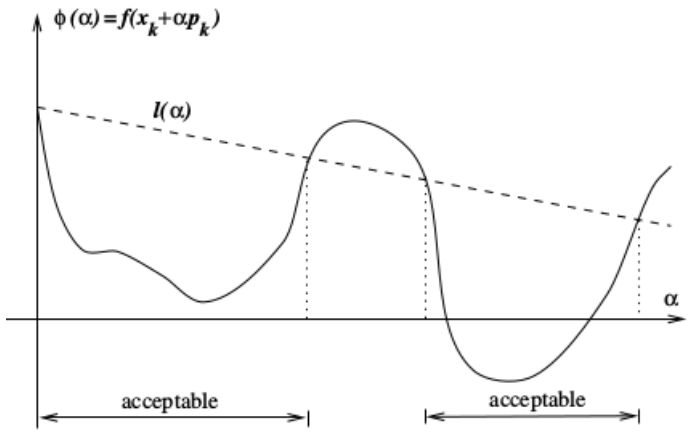
### 3.1.1 Wolfe Conditions

A popular inexact line search condition stipulates that $\alpha_k$ should first of all give a *sufficient decrease* in the objective function $f$, as stated by:

$$f(x_k + \alpha p_k) \leq f(x_k) + c_1 \alpha \nabla f_k^T p_k \quad c_1 \in (0, 1) \tag{3.5}$$

(called the *Arminjo Condition*). This says that the reduction in $f$ should be proportional to the step length $\alpha_k$ and the directional derivative $\nabla f_k^T p_k$. This condition is illustrated in the following figure, where $l(\alpha)$ is given by the RHS of the above equation.

Figure 3.1: Sufficient Decrease Condition

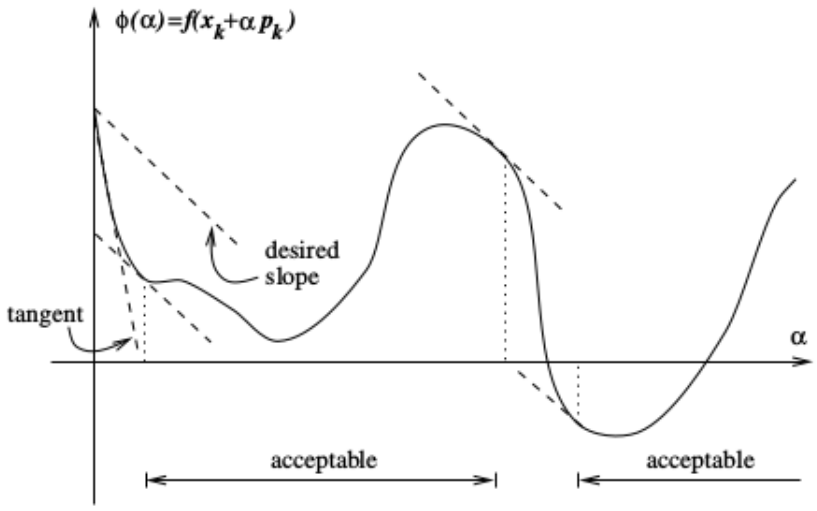



The sufficient decrease condition is not enough by itself to ensure that the algorithm makes reasonable progress because, as we see from the above figure, it is satisfied for sufficiently small values of $\alpha$, since $\phi(\alpha)$ almost always (if not always) starts off decreasing. To rule out ridiculously short steps, we introduce a second requirement, called the *curvature condition*, which requires $\alpha_k$ to satisfy

$$\nabla f(x_k + \alpha_k p_k)^T p_k \geq c_2 \nabla f_k^T p_k \quad c_2 \in (c_1, 1) \tag{3.6}$$

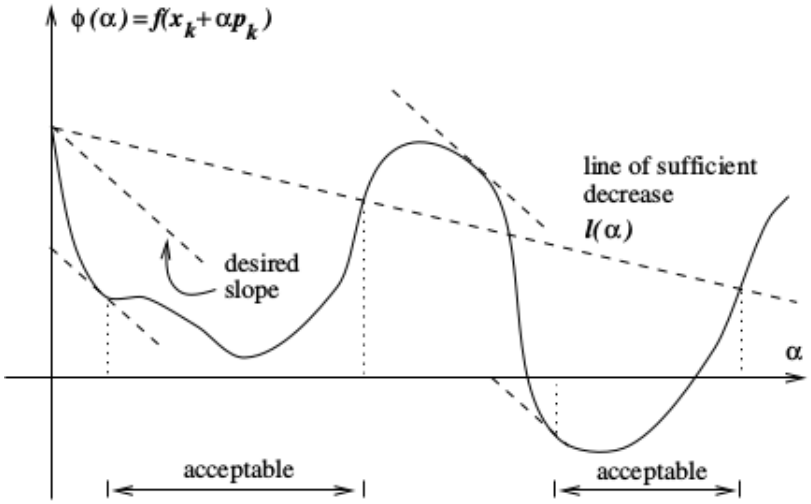
where the LHS is simply the derivative $\phi'(\alpha_k)$, so the curvature condtion ensures that the slope of $\phi$ at $\alpha_k$ is greater than $c_2$ times the initial slope $\phi'(0)$. This makes sense since the slope $\phi'(\alpha)$ is strongly negative, we have an indication that we can reduce $f$ significantly by moving further along the chosen direction.

Figure 3.2: The Curvature Condition




On the other hand, if $\phi'(\alpha_k)$ is only slightly negative or even positive, it is a sign that we cannot expect much more decrease in $f$ in this direction, so it makes sense to terminate the line search. These two conditions are known as the **Wolfe Conditions** and their collective impact is illustrated in the next figure.

Figure 3.3: The Curvature Condition




**Technical Assumptions**

1. $f$ is bounded from below $\Leftrightarrow f(x) > -\infty \quad \forall x \in \mathbb{R}^n$
2. $\nabla f(x)$ is *Lipschitz continuous* on $\mathcal{N}$ (i.e. locally lipschitz) if there exists a constant $L > 0$ such that $\|\nabla f(x) - \nabla f(\tilde{x})\| \leq L\|x - \tilde{x}\| \quad \forall x, \tilde{x} \in \mathcal{N}$

**Definition 3.1.1.** For a given $x_0$, the **level set** $\mathcal{L} = \{x \in \mathbb{R}^n : f(x) \leq f(x_0)\}$

**Theorem 3.1.1.** *(Zontendijk) Run line search with $p_k$ being a descent direction and $\alpha_k$ satisfying the two Wolfe Conditions for all $k \geq 0$. Under the technical assumptions and the following definition:*

$$\cos(\theta_k) = \frac{\nabla f_k^T p_k}{\|\nabla f_k\|\|p_k\|}$$

*then:*

$$\sum_{k\geq 0}(\cos(\theta_k))^2\|\nabla f_k\|^2 = \lim_{t\to\infty}\sum_{k=0}^{t}(\cos(\theta_k))^2\|\nabla f_k\|^2 < \infty \tag{3.7}$$

**Remark** For gradient descent, $p_k = -\nabla f_k \Rightarrow \cos(\theta_k) < 0$ unless $\nabla f_k = 0$, then $\cos(\theta_k) = 0$. Thus $(\cos(\theta_k))^2 > 0$. According to the *Zontendijk Theorem*, terms must start going to zero in order to get to a finite sum. Thus, $\|\nabla f_k\| \to 0, thus \nabla f_k \to 0$ as $k \to \infty$

*Proof.* This will essentially be a proof by definition.
Starting from the second Wolfe condition, we have:

$$\nabla f_{k+1}^T p_k - \nabla f_k^T p_k = (\nabla f_{k+1} - \nabla f_k)^T p_k \geq (c_2 - 1)\nabla f_k^T p_k$$

Also:

$$\begin{aligned}(\nabla f_{k+1} - \nabla f_k)^T p_k &\leq |(\nabla f_{k+1} - \nabla f_k)^T p_k| \\ &\leq \|(\nabla f_{k+1} - \nabla f_k)\| \cdot \|p_k\| \\ &\leq L\|x_{k+1} - x_k\| \cdot \|p_k\| = L\|x_k + \alpha_k p_k - x_k\| \cdot \|p_k\| \\ &= \alpha_k L\|p_k\|^2\end{aligned}$$

which we got using *Lipschitz continuous*. Combining these two:

$$(\nabla f_{k+1} - \nabla f_k)^T p_k \leq L\alpha_k\|p_k\|^2 \geq (c_2 - 1)\nabla f_k^T p_k \Rightarrow \alpha_k \geq \frac{(c_2 - 1)}{L}\frac{\nabla f_k^T p_k}{\|p_k\|^2}$$

Now plugging this into the first Wolfe Condition

$$\begin{aligned} f_{k+1} &\leq f_k + c_1\alpha_k \nabla f_k^T p_k \\ &\leq f_k - \frac{c_1(1-c_2)}{L}\frac{(\nabla f_k^T p_k)^2}{\|p_k\|^2} \\ &\equiv f_k - A(\cos(\theta_k))^2\|\nabla f_k\|^2 \end{aligned}$$

such that $A \equiv \frac{c_1(1-c_2)}{L} > 0$. Iteratively; we get the following relation:

$$f_{k+1} \leq f_0 - c\sum_{j=0}^{k}(\cos(\theta_j))^2\|\nabla f_j\|^2$$

Recall that since $f(x)$ is bounded below by $b$, i.e. $f(x) \geq b > -\infty$, which implies the following:

$$f_0 - f_{k+1} \leq f_0 - b < \infty$$

thus:

$$\lim_{k\to\infty}\sum_{j=0}^{k}(\cos(\theta_j))^2\|\nabla f_j\|^2 < \infty$$

□

### 3.1.2 Step Length Selection Algorithms

The first algorithm outlined in is called **Back-Tracking Line Search** and the algorithm is as follows:

Choose $\bar{\alpha} > 0$, $\rho \in (0,1)$ and $c \in (0,1)$;
Set $\alpha \leftarrow \bar{\alpha}$;
**repeat**
  $\alpha \leftarrow \rho\alpha$;
**until** $f(x_k + \alpha p_k) \leq f(x_k) + c\alpha\nabla f_k^T p_k$;
Terminate with $\alpha_k = \alpha$

Problems with the Wolfe Conditions are based on $\rho$ being too large and missing the proper interval or picking a terrible choice for $\bar{\alpha}$ and moving only to the left (and approaching 0 again).

## 3.2 Quantifying Rates of Convergence

Suppose $\lim_{k\to\infty} a_k = a^\star$

**Definition 3.2.1.** A type of convergence is called **linear convergence** if $\exists\epsilon \in (0,1), \exists K \in \mathbb{R}$ such that $\forall k \geq K$

$$\lim_{k\to\infty}\frac{\|a_{k+1} - a^\star\|}{\|a_k - a^\star\|} \leq \epsilon$$

**Example 3.2.1.** $a_k = 1 - (1/2)^k$ then choose $\epsilon = 1/2$

**Definition 3.2.2.** A type of convergence is called **super linear convergence** if $\exists\epsilon \in (0,1), \exists K \in \mathbb{R}$ such that $\forall k \geq K$

$$\lim_{k\to\infty}\frac{\|a_{k+1} - a^\star\|}{\|a_k - a^\star\|} = 0$$

**Example 3.2.2.** $a_k = 1 - (k)^{-k}$

**Definition 3.2.3.** A type of convergence is called **quadratic convergence** if $\exists\epsilon \in (0,1), \exists K \in \mathbb{R}$ such that $\forall k \geq K$

$$\lim_{k\to\infty}\frac{\|a_{k+1} - a^\star\|}{\|a_k - a^\star\|^2} \leq \epsilon$$

**Example 3.2.3.** $a_k = 1 - (1/2)^{2^k}$

## 3.3 Review of Miscellaneous things

Let $A = A^T$ (be symmetric), with $n$ eigenvalues $\lambda_1 < \lambda_2 < \cdots < \lambda_n$ (sorted the eigenvalues in ascending order), then we can take the "rated quotient", defined as:

$$\frac{x^T A x}{x^T x}$$

and we can write the following:

$$\sup_{x \neq 0} \frac{x^T A x}{x^T x} \leq \lambda_n \qquad \text{and } \inf_{x \neq 0} \frac{x^T A x}{x^T x} \leq \lambda_1$$

so both of these imply that the rated quotient is bounded by both of these eigenvalues. In turn, we have:

$$\lambda_1 \leq \frac{x^T A x}{x^T x} \leq \lambda_n \Rightarrow \lambda_1 \|x\|_2^2 \leq x^T A x \leq \lambda_n \|x\|_2^2$$

where $\|\cdot\|$ is the Euclidean norm.

**Definition 3.3.1.** A function $\|\cdot\|$ is a **vector norm** if:

- $\|x\| = 0$ iff $x = 0$
- $\|x + y\| \leq \|x\| + \|y\|$
- $\|\alpha x\| = |\alpha| \|x\| \quad \forall \alpha \in \mathbb{R}$

**Definition 3.3.2.** The **p-norm** of a vector, such that $p \geq 1$, is defined as:

$$\|x\|_p = \big( \sum_{i=1}^{n} |x_i|^p \big)^{\frac{1}{p}}$$

Other norms are:

1. $p = 2$ is called the Euclidean Norm
2. $p = 1$: $\|x\|_1 = \sum_{i=1}^{n} |x_i|$
3. $p = \infty$: $\|x\|_\infty = \max(|x_1|, \ldots, |x_n|)$

A function $\|\cdot\|$ is a **maxtrix norm** if:

- $\|A\| \geq 0$ and $\|A\| = 0$ iff $A = 0$
- $\|A + B\| \leq \|A\| + \|B\|$
- $\|\alpha A\| = |\alpha| \|A\| \quad \forall \alpha \in \mathbb{R}$

The matrix norm induced by a vector norm $\|\cdot\|_p$ is:

$$\|A\|_p = \sup_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_p} = \sup_{x : \|x\| = 1} \|Ax\|_p$$

so then:

$$\|A\|_2 = \sup_{x \neq 0} \left( \frac{x^T A^T A x}{x^T x} \right)^{\frac{1}{2}}$$

**Special Properties**

1. If $A = A^T$, then $\|A\|_2 = \max_{i=1,\dots,n} |\lambda_i|$
2. If $A = A^T \succ 0$ (positive definite), then $\|A\|_2 = \lambda_n$
3. Specifically for the $p = 2$ norm, we have that $\|Ax\|_2 \leq \|A\|_2 \|x\|_2$

## 3.4 Rate of Convergence of Gradient Descent (GD)

Recall:

$$x_{k+1} = x_k - \alpha_k \nabla f_k \tag{3.8}$$

and $x_k \to x_\star$, where $x_\star$ is the local minimizer, happens linearly (that is, it converges linearly) if there is $\rho \in (0, 1)$ and

$$\frac{\|x_{k+1} - x_\star\|}{\|x_k - x_\star\|} \leq \rho$$

where there will be a primary focus on $f(x) = \frac{1}{2} x^T Q x - b^T x$, where $Q = Q^T \succ 0$, so now $f$ is a convex function. We have that the gradient is:

$$\nabla f(x) = Qx - b$$

and then the local minimizer $x_\star$ is given by: $Qx_\star = b$.
Ideally (through exact line search), we have that:

$$\alpha_k = \min_{\alpha > 0} \phi(\alpha) := f(x_k - \alpha \nabla f_k) \tag{3.9}$$

whereby minimizing the function $\phi(\alpha)$, we get:

$$\frac{d}{d\alpha}\phi(\alpha) = 0 \Rightarrow \alpha_k = \frac{\nabla f_k^T \nabla f_k}{\nabla f_k^T Q \nabla f_k} \tag{3.10}$$

which looks very similar to the forms of the rated quotient. We now make the following definition: $\|x\|_Q^2 := x^T Q x$ and we observe:

$$\begin{aligned} \|x_k - x_\star\|_Q^2 &= (x_k - x_\star)^T Q (x_k - x_\star) \\ &= x_k^T Q x_k - 2 x_k^T Q x_\star + x_\star^T Q x_\star \\ &= 2 f(x_k) + x_\star^T Q x_\star \end{aligned}$$

since $Qx_\star = b$. We also have that:

$$\begin{aligned} \|x_{k+1} - x_\star\|_Q^2 &= 2 f(x_k - \alpha_k \nabla f_k) + x_\star^T Q x_\star \\ &= 2\big[\frac{1}{2}(x_k - \alpha_k \nabla f_k)^T Q (x_k - \alpha_k \nabla f_k) - b^T (x_k - \alpha_k \nabla f_k)\big] + x_\star^T Q x_\star \\ &= x_k^T Q x_k - 2\alpha_k \nabla f_k^T Q x_k + \alpha^2 \nabla f_k^T Q \nabla f_k - 2 b^T x_k + 2\alpha_k b^T \nabla f_k + x_\star^T Q x_\star \end{aligned}$$

Now we want to look at the difference between the two norms we just defined. This is outlined below. Note that we make use of the closed-form solution of $\alpha_k$ written above. So:

$$\begin{aligned} \|x_k - x_\star\|_Q^2 - \|x_{k+1} - x_\star\|_Q^2 &= 2\alpha_k \nabla f_k^T Q x_k - 2\alpha_k \nabla f_k^T b - \alpha_k^2 \nabla f_k^T Q \nabla f_k \\ &= 2\alpha_k \nabla f_k^T (Q x_k - b) - \alpha_k^2 \nabla f_k^T Q \nabla f_k \\ &= 2\alpha_k \nabla f_k^T \nabla f_k - \alpha_k^2 \nabla f_k^T Q \nabla f_k \\ &= 2\frac{(\nabla f_k^T \nabla f_k)^2}{\nabla f_k^T Q \nabla f_k} - \frac{(\nabla f_k^T \nabla f_k)^2}{\nabla f_k^T Q \nabla f_k} \\ &= \frac{(\nabla f_k^T \nabla f_k)^2}{\nabla f_k^T Q \nabla f_k} \end{aligned}$$

Now we look at some other fraction:

$$\frac{\|x_k - x_\star\|_Q^2 - \|x_{k+1} - x_\star\|_Q^2}{\|x_k - x_\star\|_Q^2} = \frac{(\nabla f_k^T \nabla f_k)^2}{(\nabla f_k^T Q \nabla f_k)\|x_k - x_\star\|_Q^2}$$

and, using the the following nice relationships:

$$\begin{aligned} \nabla f_k = Qx_k - b \Rightarrow \nabla f_\star = Qx_\star - b = 0 \quad \text{thus} \\ \nabla f_k = \nabla f_k - \nabla f_\star = Q(x_k - x_\star) \quad \text{thus} \\ \|x_k - x_\star\|_Q^2 = (x_k - x_\star)^T Q^T Q^{-1} Q(x_k - x_\star) = \nabla f_k^T Q^{-1} \nabla f_k \end{aligned}$$

and so we have that:

$$\begin{aligned} 1 - \frac{\|x_{k+1} - x_\star\|_Q^2}{\|x_k - x_\star\|_Q^2} &= \frac{\nabla f_k^T \nabla f_k}{\nabla f_k^T Q \nabla f_k} \frac{\nabla f_k^T \nabla f_k}{\nabla f_k^T Q^{-1} \nabla f_k} \\ &\geq \frac{1}{\lambda_n} \lambda_1 \end{aligned}$$

which is because, from the rated quotient, we have that $\frac{a^T Q^{-1} a}{a^T a} \leq \frac{1}{\lambda_n}$ and FINALLY:

$$\frac{\|x_{k+1} - x_k\|_Q^2}{\|x_k - x_\star\|_Q^2} \leq 1 - \frac{\lambda_1}{\lambda_n} := 1 - \frac{1}{\kappa} \tag{3.11}$$

where $\kappa := \frac{\lambda_n}{\lambda_1}$ is called the **condition number** of the matrix Q. Note that $\kappa \geq 1$, since $\lambda_n \geq \lambda_1$. This is a very important concept regarding scaling.

**Example 3.4.1.**

$$Q = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \Rightarrow \kappa = 1$$

which gives a perfect circle and the problem is decently scaled.

**Example 3.4.2.**

$$Q = \begin{bmatrix} 10^3 & 0 \\ 0 & 1 \end{bmatrix} \Rightarrow \kappa = 10^3$$

which gives an ellipse ellongated around the y-axis (i.e. going up along the y axis) and the problem is poorly scaled.

Now we have that:

$$\begin{aligned} \|x_{k+1} - x_\star\|_Q^2 &\leq \left(1 - \frac{1}{\kappa}\right) \|x_k - x_\star\|_Q^2 \\ 2f(x_{k+1}) + x_\star^T Q x_\star &\leq \left(1 - \frac{1}{\kappa}\right)(2f(x_k) + x_\star^T Q x_\star) \\ \Rightarrow f(x_{k+1}) &\leq \left(1 - \frac{1}{\kappa}\right) f(x_k) - \frac{1}{\kappa} x_\star^T Q x_\star \end{aligned}$$

we also have a relationship between the two norms ($p = 2$ and the $Q$-norm):

$$\|x\|_Q^2 = x^T Q x \leq \lambda_n \|x\|_2^2 \quad \text{and } \lambda_1 \|x\|_2^2 \leq \|x\|_Q^2 \Rightarrow \lambda_1 \|x\|_2^2 \leq \|x\|_Q^2 \leq \lambda_n \|x\|_2^2$$

## 3.5 Rate of Convergence of Newton's Method

This is another line search method with the following conditions: $p_k = -(\nabla^2 f_k)^{-1} f_k$ and $\alpha_k = 1$ thus the updating step is:

$$x_{k+1} = x_k - (\nabla^2 f_k)^{-1} f_k$$

Recall that $x_k \to x^\star$ *convergences quadratically* if, for large enough $k$, there exists a constant $M > 0$ such that:

$$\lim_{k \to \infty} \frac{\|a_{k+1} - a^\star\|}{\|a_k - a^\star\|^2} \leq M$$

**Technical Assumptions** Let $x^\star$ be a strict local minimizer. Therefore: $\nabla f(x^\star) = 0$ and $\nabla^2 f(x^\star) \succ 0$. For $r > 0$, there exists a neighborhood $\mathcal{N} = \{x \mid \|x - x^\star\| < r\}$ such that:

1. $x_0 \in \mathcal{N}$
2. $\exists L > 0$ such that $\|\nabla^2 f(x) - \nabla^2 f(y)\|_2 \leq L\|x - y\| \ \ \forall x, y \in \mathcal{N}$ (i.e. Lipschitz). This says that the eigenvalues of these Hessian matrices do not change too much in this neighborhood
3. $\|\nabla^2 f(x)^{-1}\|_2 \leq 2\|\nabla^2 f(x^\star)^{-1}\| \ \ \forall x \in \mathcal{N}$

**Recall Taylor's Expansion**

$$\nabla f_k - \nabla f_\star = \int_0^1 \nabla^2 f(x_k + \tau(x^\star - x_k))(x_k - x^\star) d\tau$$

**Showing the Quadratic Convergence of Newton's Method**
A bunch of algebraic manipulation follows to show this quadratic convergence.

$$\begin{aligned} x_{k+1} - x^\star &= x_k - x^\star - \nabla^2 f_k^{-1} \nabla f_k \\ &= \nabla^2 f_k^{-1} \big[\nabla^2 f_k (x_k - x^\star) - (\nabla f_k - \nabla f_\star)\big] \end{aligned}$$

We then take the $\|\cdot\|_2$ of the term in the brackets and apply Taylor's Theorem:

$$\begin{aligned} \|\nabla^2 f_k (x_k - x^\star) - (\nabla f_k - \nabla f_\star)\|_2 &= \left\| \int_0^1 \big[\nabla^2 f_k - \nabla^2 f(x_k + \tau(x^\star - x_k))\big](x_k - x^\star) d\tau \right\|_2 \\ &\leq \int_0^1 \|\nabla^2 f_k - \nabla^2 f(x_k + \tau(x^\star - x_k))\|_2 \cdot \|x_k - x^\star\|_2 d\tau \\ &\leq \|x_k - x^\star\|_2 \int_0^1 L\|\tau(x^\star - x_k)\|_2 d\tau \\ &\leq \frac{L}{2}\|x_k - x^\star\|_2^2 \end{aligned}$$

We go back to the first set of equations that we started with, take the norm and apply these new results:

$$\begin{aligned} \|x_{k+1} - x^\star\|_2 \leq \|\nabla^2 f_k^{-1}\|_2 \cdot \frac{L}{2}\|x_k - x^\star\|_2^2 \Leftrightarrow \frac{\|x_{k+1} - x^\star\|_2}{\|x_k - x^\star\|_2^2} &\leq \frac{L}{2}\|\nabla^2 f_k^{-1}\|_2 \\ &\leq L\|\nabla^2 f_k^{-1}\|_2 \end{aligned}$$

Thus we have that the Newton Method converges quadratically when "close enough" to $x^\star$ $\square$

**Example 3.5.1.** $a_k = 1 - (1/2)^{2^k}$. Then $|x_{k+1} - x^\star| = (\frac{1}{2})^{2^{k+1}}$ and $|x_k - x^\star| = (\frac{1}{2})^{2^k}$ thus we have that: $|x_k - x^\star|^2 = (\frac{1}{2})^{2^k} \cdot (\frac{1}{2})^{2^k}$ and:

$$\frac{|x_{k+1} - x^\star|}{|x_k - x^\star|^2} = \left[\left(\frac{1}{2}\right)^{2^k} \cdot \left(\frac{1}{2}\right)^2\right] \Big/ \left[\left(\frac{1}{2}\right)^{2^k} \cdot \left(\frac{1}{2}\right)^{2^k}\right] = 1$$

(**CHECK**)

## 3.6 Quasi-Newton Methods

Another line search method with the general form for the search direction $p_k = -B_k^{-1}\nabla f_k$ and we enforce that $B_k = B_k^T \succ 0$ (so $B_k$ is symmetric) and we use the following model:

$$f(x_k + p) \simeq m_k(p) = f_k + \nabla f_k^T p + \frac{1}{2}p^T B_k p$$

and, in addition:

$$f(x_{k+1} + p) \simeq m_{k+1}(p) = f_{k+1} + \nabla f_{k+1}^T p + \frac{1}{2}p^T B_{k+1} p$$

While performing the algorithm, we have $B_k$, $x_k$, $x_{k+1}$, $\nabla f_k$ and $\nabla f_{k+1}$. We should (and will) choose $B_k$ such that: $\nabla m_{k+1}$ matches $\nabla f$ at $x_k$ and $x_{k+1}$, thus

$$\nabla m_{k+1}(p) = \nabla f_{k+1} + B_{k+1}p$$

so at $x_{k+1}$, we have $\nabla m_{k+1}(0) = \nabla f_{k+1}$ and, at $x_k$, $\nabla m_{k+1}(-\alpha_k p_k) = \nabla f_{k+1} - B_{k+1}(\alpha_k p_k) = \nabla f_k$ THUS

$$\nabla f_{k+1} - \nabla f_k = B_k(x_{k+1} - x_k) :\Rightarrow y_k = B_{k+1}s_k$$

ans so to recap, we want $B_{k+1} = B_{k+1}^T \succ 0$ and $y_k = B_{k+1}s_k$, where $y_k$ and $s_k$ are defined above (and also, $s_k = \alpha_k p_k$). In turn, this implies that $s_k^T y_k = s_k^T B_{k+1} s_k > 0$, which always holds if $\alpha_k$ satisfies the Wolfe Conditions

## 3.7 BFGS Algorithm

Focuses on finding $H_k := B_k^{-1}$ directly. From above, we have that $H_k$ must satisfy $H_{k+1}y_k = s_k$. (Not important the previous conditin and $H_{k+1} \succ 0 \Rightarrow y_k^T s_k > 0$)
So we hope to find a solution $H_{k+1}$ that is the solution to the following:

$$\min_H \|H - H_k\|_{\bar{G}_k} \quad \text{such that } H = H^T \text{ and } Hy_k = s_k$$

where

$$\bar{G}_k = \int_0^1 \nabla^2 f(x_k + \tau(x_{k+1} - x_k))d\tau$$

and note that

$$\|A\|_G = \|G^{\frac{1}{2}}AG^{\frac{1}{2}}\|_F$$

where the **Frobenius Norm** is defined as:

$$\|C\|_F^2 = \sum_{i=1}^{n}\sum_{j=1}^{n}(C_{ij})^2$$

and the final result of all this is:

$$H_{k+1} = (I - \rho_k s_k y_k^T)H_k(I - \rho_k y_k s_k^T) + \rho_k s_k s_k^T) \quad \text{where } \rho_k = \frac{1}{s_k^T y_k}$$

The BFGS algorithm:

Given $x_0$, $\epsilon > 0$;
Set $H_0 \simeq I$ (proportional to $I$);
**while** $\|\nabla f_k\| > \epsilon$ **do**
  $p_k = -H_k \nabla f_k$;
  $\alpha_k$ via **line search** (Both Wolfe Conditions);
  $x_{k+1} = x_k + \alpha_k p_k$;
  $s_k = x_{k+1} - x_k = \alpha_k p_k$;
  $y_k = \nabla f_{k+1} - \nabla f_k$;
  $\rho_k = (s_k^T y_k)^{-1}$;
  $H_{k+1} = (I - \rho_k s_k y_k^T)H_k(I - \rho_k y_k s_k^T) + \rho_k s_k s_k^T$;
**end**

### 3.7.1 Convergence Analysis for BFGS

Questions to ask:

- Is $p_k$ a descent direction?
  $\Rightarrow -(\nabla f_k^T)p_k > 0$, thus: $-(\nabla f_k^T)(-H_k \nabla f_k) = \nabla f_k^T H_k \nabla f_k > 0$ if $H_k \succ 0$. By choice, $H_0 = \gamma I \succ 0$, where $\gamma > 0$.
- Suppose $H_k \succ 0$. Is $H_{k+1} \succ 0$?
  For arbitrary $z \neq 0$, $z^T H_{k+1} z = w^T H_k w + \rho_k (z^T s_k)^2$ where $w = z - \rho_k y_k (s_k^T z)$. Since $\rho_k > 0 \ \forall k$ and since $H_k \succ 0$, we have that $H_{k+1} \succ 0$ (basically simple induction argument).
- Does $\nabla f_k \to 0$? (i.e. does $x_k$ approach a stationary point?)
- Rate of convergence?

#### (Recall) Zoutendijk's Theorem

**Theorem 3.7.1.** *If $f$ is "well-behaved" and $\{\alpha_k\}_{k=1}^{\infty}$ satisfy the Wolfe Conditions and $\{p_k\}_{k=1}^{\infty}$ are descent directions, then we have*

$$\cos(\theta_k) = \frac{\nabla f_k^T p_k}{\|\nabla f_k\| \|p_k\|}$$

*and finally:*

$$\sum_{k \geq 0} (\cos(\theta_k))^2 \|\nabla f_k\|^2 = \lim_{t \to \infty} \sum_{k=0}^{t} (\cos(\theta_k))^2 \|\nabla f_k\|^2 < \infty \tag{3.12}$$

**Example 3.7.1.** For Gradient Descent, $p_k = -\nabla f_k$, thus $\cos(\theta_k) = 1 \ \forall k$, thus $\sum_{k \geq 0} \|\nabla f_k\|^2 < \infty$ implies that $\|\nabla f_k\|^2 \to 0$

**Example 3.7.2.** So for BFGS, we need to show that $\cos(\theta_k)^2 \geq \delta > 0$ infinitely often. Assume $m\|z\|_2^2 \leq z^T \nabla^2 f(x) z \leq M \|z\|_2^2$. It can be shown that $\sum_{k \geq 0} \|x_k - x^\star\|_2 < \infty$. Under all this crap, BFGS converges super-linearly.

#### BFGS Iteration Complexity

Suppose we have that

$$H_{k+1} = V_k^T H_k V_k + \rho s_k s_k^T$$

where $s_k = x_{k+1} - x_k$,$y_k = \nabla f_{k+1} - \nabla f_k$, $\rho_k 1/(s_k^T y_k)$ and $V_k = I - \rho_k y_k s_k^T$. By default, we have:

- Memory Complexity: $O(n^2)$
- Time Complexity: $O(n^2)$ because matrix multiplication or something

However, we observe that $H_{k+1}$ is a function of $H_0 = \gamma I$ and $\{(s_i, y_i)\}_{i=0}^{k} \Rightarrow 2nk$ storage.

### 3.7.2 Limited-Memory BFGS (L-BFGS)

We keep the $m$ most recent pairs $\{(s_i, y_i)\}_{i=k-m+1}^{k}$, with $m < n$. We have that, by approximating $H_k^0 \simeq H_{k-m}$:

$$\begin{aligned} H_k = &\big(V_{k-1}^T V_{k-2}^T \cdots V_{k-m}^T\big) H_k^0 \big(V_{k-m} \cdots V_{k-2} V_{k-1}\big) \\ &+ \rho_{k-m} \big(V_{k-1}^T \cdots V_{k-m+1}^T\big) s_{k-m} s_{k-m}^T \big(V_{k-m+1} \cdots V_{k-1}\big) \\ &+ \rho_{k-m+1} \big(V_{k-1}^T \cdots V_{k-m+2}^T\big) s_{k-m+1} s_{k-m+1}^T \big(V_{k-m+2} \cdots V_{k-1}\big) \\ &+ \ldots + \rho_{k-1} s_{k-1}^T s_{k-1} \end{aligned}$$

Input: $\{(s_i, y_i)\}_{i=k-m}^{k-1}$, $\nabla f_k$;
Set $q \leftarrow \nabla f_k$;
**for** $i = k-1, \ldots, k-m$ **do**
  $\alpha_i = \rho_i s_i^T q$;
  $q \leftarrow q - \alpha_i y_i$
**end**
Set $r \leftarrow H_k^0 q$ **for** $i = k-m, \ldots, k-1$ **do**
  $\beta = \rho_i y_i^T r$;
  $r \leftarrow r + s_i(\alpha_i + \beta)$
**end**
Output: $r = H_k \nabla f_k$

- Storage Complexity: $O(nm)$
- Computational Complexity: $O(nm)$
- Typically, $H_k^0 = \frac{s_{k-1}^T y_{k-1}}{y_{k-1}^T y_{k-1}} I$

Now, we recall that:

$$\bar{G}_k = \int_0^1 \nabla^2 f(x_k + \tau \alpha_k p_k) d\tau$$

and by applying Taylor's theorem:

$$\begin{aligned} y_k = \nabla f_{k+1} - \nabla f_k &= \int_0^1 \nabla^2 f(x_k + \tau \alpha_k p_k)(x_{k+1} - x_k) d\tau \\ &= \bar{G}_k (x_{k+1} - x_k) := s_k \end{aligned}$$

and we define $z_k = \bar{(G)}_k^{1/2} s_k$ thus we get that

$$\frac{s_k y_k}{y_n^T y_n} = \frac{s_k^T \bar{G}_k s_k}{s_k^T \bar{G}_k \bar{G}_k s_k} = \frac{z_k^T z_k}{z_k \bar{G}_k z_k}$$

## 3.8 Conjugate Direction Methods

**Recall** Minimizing the quadratic equation

$$f(x) = \frac{1}{2} x^T A x - b^T x$$

where $A = A^T \succ 0$ is equivalent to solving $Ax = b$. This is because $f(x)$ is a convex function, this it's local minimium is the global minimum. We also have that the gradient of the function is

$$\nabla f(x) = Ax - b := r(x) \text{ called the residual at x } \Rightarrow \nabla f(x) = 0 \Leftrightarrow Ax = b$$

**Definition 3.8.1.** A set of non-zero vectors $\{p_0, p_1, \ldots, p_l\}$ are said to be conjugate with respect to $A$ if $p_i^T A p_j = 0$ for all $i \neq j$

**Lemma 3.8.0.1.** *Conjugacy implies that* $\{p_0, p_1, \ldots, p_l\}$ *are linearly independent*

*Proof.* Suppose that they are not linearly independent. Without loss of generality, suppose that $p_l = \sigma_0 p_0 + \sigma_1 p_1 + \ldots + \sigma_{l-1} p_{l-1}$, where at least one of the $\sigma_i \neq 0$. In particular, suppose that $\sigma_j \neq 0$. Then

$$\begin{aligned} p_j^T A p_l &= p_j^T A(\sigma_0 p_0 + \sigma_1 p_1 + \ldots + \sigma_{l-1} p_{l-1}) \\ &= \text{ proof left incomplete} \end{aligned}$$

□

## Conjugate Direction Algorithm

Given $x_0$ and $\{p_0, p_1, \ldots, p_{n-1}\}$;
**for** $k = 0, \ldots n-1$ **do**
  Set $x_{k+1} = x_k + \alpha_k p_k$;
  where $\alpha_k = \frac{-r_k^T p_k}{p_k^T A p_k}$ and $r_k = Ax_k - b$
**end**

**Claim 1.** *$x_k$ converges (i.e. $x_k = x_\star$) in at most $n$ iterations. For a given $x_0$, there is a $K \leq n$ such that $x_k = x_\star$ for all $k \geq K$*

*Proof.* Since $\{p_0, p_1, \ldots, p_{n-1}\}$ are conjugate (for a particular matrix $A$), they are linearly independent; so they span $\mathbb{R}^n$. Thus we have that:

$$x_\star - x_0 = \sigma_0 p_0 + \sigma_1 p_1 + \ldots + \sigma_{n-1} p_{n-1}$$

We also have that, by conjugacy:

$$p_i^T A(x_\star - x_0) = \sigma_i p_i^T A p_i \Rightarrow \sigma_i = \frac{p_i^T A(x_\star - x_0)}{p_i^T A p_i}$$

Also, because of the spanning property, we have that:

$$x_k = x_0 + \alpha_0 p_0 + \ldots \alpha_{k-1} p_{k-1} \quad \text{and } p_k^T A(x_k - x_0) = 0$$

Thus:

$$\begin{aligned} p_i^T A(x_\star - x_0) &= p_i^T A((x_\star - x_i) + (x_i - x_0)) \\ &= p_i^T A(x_\star - x_i) \quad \text{see above property} \end{aligned}$$

and since $\nabla f(x) = Ax - b$, we have that $\nabla f(x_\star) = Ax_\star - b = 0 \Rightarrow Ax_\star = b$, thus:

$$p_i^T A(x_\star - x_0) = p_i^T(b - Ax_i) = -p_i^T r_i$$

Thus:

$$\sigma_k = \frac{-p_k^T r_k}{p_k^T A p_k} = \alpha_k$$

Therefore, $x_n = x_\star$ □

**Example 3.8.1.**

$$A = \begin{bmatrix} 10 & 0 \\ 0 & 1 \end{bmatrix} \text{ and } b = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

which looks like an ellipsoid ellongated around the y-axis. We claim that $p_0 = [1, 0]^T$ and $p_1 = [0, 1]^T$ are conjugate. Literally in two iterations, our $x_0$ moves directly to the minimum.

**Claim 2.** *For $k \geq 1$, we have that $r_k^T p_i = 0$ for all $i = 0, 1, \ldots k-1$*

*Proof.*
Base case:

$$\begin{aligned} r_1^T p_0 &= (Ax_1 - b)^T p_0 \\ &= (A(x_0 + \alpha_0 p_0) - b)^T p_0 \\ &= (\underbrace{Ax_0 - b}_{r_0} + \alpha_0 A p_0)^T p_0 \\ &= r_0^T p_0 + \alpha_0 p_0^T A p_0 \\ &= 0 \text{ (by definition of } \alpha_0) \end{aligned}$$

Hypothesis: $r_{k-1}^T p_i = 0$ holds for all $i = 0, 1, \ldots, k-2$
Induction:
By definition of $r_k$

$$\begin{aligned} r_k^T p_i &= p_i^T (Ax_k - b) \\ &= p_i^T (A(x_{k-1} + \alpha_{k-1} p_{k-1}) - b) \\ &= p_i^T (r_{k-1} + \alpha_{k-1} A p_{k-1}) \end{aligned}$$

if $i = k-1$

$$p_{k-1}^T r_k = p_{k-1}^T r_{k-1} + \alpha_{k-1} p_{k-1}^T A p_{k-1} = 0(\text{ by choice of } \alpha_{k-1})$$

if $i = 0, 1, \ldots, k-2$

$$p_i^T r_k = p_i^T r_{k-1} + \alpha_{k-1} p_i^T A p_{k-1} = 0$$

where the first term is 0 by hypothesis and the second term by conjugacy □

**Claim 3.** *$x_k$ minimizes $f(x) = \frac{1}{2} x^T A x - b^T x$ over $\{x \in \mathbb{R}^n \mid x \in x_0 + span\{p_0, \ldots, p_{k-1}\}\}$*

Why is this useful? We can focus on $x(\sigma) = x_0 + \sigma_0 p_0 + \ldots \sigma_{k-1} p_{k-1} = x_0 + P_k \sigma$, where $\sigma \in \mathbb{R}^k$ and $P_k$ is a matrix where the columns are $p_i$, with $i = 0, 1, \ldots k-1$. This implies that the minimization problem in **Claim 3** is equivalent to:

$$\min_{\sigma \in \mathbb{R}^l} f(x_0 + P_k \sigma) = g(\sigma)$$

and from assignment 1, we have that:

$$\begin{aligned} \nabla g(\sigma) &= P_k^T \nabla f(x_0 + P_k \sigma) \\ &= P_k^T (A(x_0 + P_k \sigma) - b) \\ &= P_k^T (A(x_0 + \sigma_0 p_0 + \ldots + \sigma_{k-1} p_{k-1}) - b) \\ &= P_k^T (r_0 + \sigma_0 A p_0 + \ldots + \sigma_{k-1} A p_{k-1}) \end{aligned}$$

and we note that $\nabla g(\sigma_\star) = 0$ where each component is represented by:

$$[\nabla g(\sigma_\star)]_i = P_{i-1}^T (r_0)$$

## 3.9 Conjugate Gradient Method

To recall, we're given a symmetric matrix $A = A^T \succ 0$ and $b$ and we wish to find $x$ such that $Ax = b \Leftrightarrow x^* = A^{-1} b$. The problem arises when the size of $A$ is large; since the order of inversion is $O(n^3)$. If, instead, we had a function $\phi(x) = \frac{1}{2} x^T A x - b^T x$, then it's gradient $\nabla \phi(x) = Ax - b$ has the same solution as above; thus $Ax^* = b$. Recall the notion of "rate of convergence" and apply it to this problem:

$$\begin{aligned} \phi(x) - \phi(x^*) &= \frac{1}{2} x^T A x - b^T x - \frac{1}{2} (x^*)^T A x^* + b^T x^* \\ &= \frac{1}{2} (x - x^*)^T A (x - x^*) = \frac{1}{2} \|x - x^*\|_A^2 \end{aligned}$$

**Example 3.9.1.** As an example, the *Newton Direction* $p_k$ is a solution to $\nabla^2 f_k p_k = -\nabla f_k$)

**Definition 3.9.1.** The **residual** at a particular point $x$ is denoted $r(x)$ and given by $Ax - b$. It tells us how close we're getting to solving the linear system

In terms of the previous problem, we have that $r_k = \nabla\phi(x_k) = A(x_k - x^*)$. So the convergence equations become something like:

$$\begin{aligned}\phi(x) - \phi(x^*) &= \frac{1}{2}(x - x^*)^T A(x - x^*) \\ &= \frac{1}{2} r_k^T A^{-1} r_k = \frac{1}{2}\|r_k\|^2_{A^{-1}}\end{aligned}$$

Again, the issue with this that we do not explicitly know $x^*$ (because hard to inverse matrix) and we can't compute this norm either (for the same reason). Instead we note that it is common to measure relative improvement as $\|r_k\|_2/\|b\|_2$ (relative residual using $x_0 = 0$).

**Recall: Conjugacy** that a set of non-zero vectors $\{p_0, \ldots, p_l\}$ are conjugate with respect to a symmetric positive definite matrix $A$ if $p_i^T A p_j = 0$ for all $i \neq j$; conjugacy implies that the vectors $\{p_0, \ldots, p_l\}$ are linearly independent.

**Recall: Conjugate Direction Method** Given $x_0$ and conjugate directions $\{p_0, p_1, \ldots, p_{n-1}\}$, we repeat:

$$x_{k+1} = x_k + \alpha_k p_k \qquad \text{with } \alpha_k = \frac{-r_k^T p_k}{p_k^T A p_k}$$

where $r_k = r(x_k)$. We also showed that $x_k = x^*$, or $r(x_k) = 0$ after at most $n$ steps

**Residuals and Expanding Subspaces:** For any starting point $x_0$, after $k$ iterations of the conjugate direction method, we have that $r_k^T p_i = 0$ for $i = 0, 1, \ldots, k-1$. This was proved by induction in class. We can use this to show that $x_k$ minimizes $\phi(x) = \frac{1}{2}x^T Ax - b^T x$ over

$$\{x \in \mathbb{R}^n \mid x \in x_0 + \text{span}\{p_0, p_1, \ldots, p_{k-1}\}\} = \{x \in \mathbb{R}^n \mid x \in x_0 + \sigma_0 p_0 + \sigma_1 p_1 + \cdots + \sigma_{k-1} p_{k-1}\}$$

**Problem:** How do we even get these conjugate vectors in the first place?! Using eigenvalue decomposition of $A$ is far too expensive — $O(n^3)$

**The Method**

1. Given $x_0$, set $p_0 = -r_0 = \nabla\phi(x_0)$
2. Update $x_{k+1} = x_k + \alpha p_k$, where $\alpha_k = (-r_k^T p_k)/(p_k^T A p_k)$ (which minimizes $\phi(x_k + \alpha p_k)$ over $\alpha$)
3. For $k \geq 1$, set $p_k = -r_k + \beta_k p_{k-1}$ where $\beta_k$ is chosen so that $p_{k-1}^T A p_k = 0$ (i.e. two successive step directions are conjugate to each other), where $\beta_k = (r_k^T A p_{k-1})/(p_{k-1}^T A p_{k-1})$

### 3.9.1 Krylov Subspaces

We can verify by induction that $p_k^T A p_i = 0$ for all $i = 1, \ldots, k-1$
Recall that since $x_{k+1} = x_k + \alpha_k p_k$, then this implies that $x_{k+1} = (x_0 + \alpha_0 p_0 + \cdots) + \alpha_k p_k$, since to get to the $x_k$ point, a bunch of steps were needed as well.
We have that $r_k = Ax_k - b$ and so $r_{k+1} = r_k + \alpha_k A p_k$.
Also $p_0 = -r_0$ and $p_k = -r_k + \beta_k p_{k-1}$
Using all these, we can show that:

$$\begin{aligned}x_{k+1} - x_0 &\in \text{span}\{p_0, p_1, \ldots, p_k\} \\ &= \text{span}\{r_0, r_1, \ldots, r_k\} \\ &= \text{span}\{r_0, Ar_0, A^2 r_0, \ldots, A^k r_0\} \quad \textbf{Krylov Subspace}\end{aligned}$$

so $x_{k+1} \in x_0 + \text{span}\{r_0, Ar_0, A^2 r_0, \ldots, A^k r_0\}$.

**Definition 3.9.2.** A **Krylov Subspace** is a subspace with the following form:

$$\begin{aligned}\mathcal{K}_k(r_0) &= \text{span}\{r_0, Ar_0, \ldots, A^{k-1}r_0\} \\ &= \{P(A)r_0 \mid P \text{ polynomial with degree}(P) < k\}\end{aligned}$$

where any point $x$ can be written as follows:

$$\begin{aligned}x &= c_0 r_0 + c_1 A r_0 + \cdots + c_{k-1}A^{k-1}r_0 \\ &= (c_0 A^0 + c_1 A^1 + \cdots + c_{k-1}A^{k-1})r_0\end{aligned}$$

**Interpretation and Consequences**
CG can be viewed as generating a sequence of points $x_k$ for $k \geq 1$ such that

$$x_k = \text{argmin}_{x \in x_0 + \mathcal{K}_k(r_0)}\phi(x)$$

and, as a consequence:

- $\phi(x_{k+1}) \leq \phi(x_k)$ (although $\|r_k\|$ may increase)
- $x_k = x_0 + P_k(A)r_0$, where $P_k$ is a polynomial and $\deg(P_k) < k$

**Definition 3.9.3.** The **characteristic polynomial** of a matrix $A$ is given by
$\chi(\lambda) = \det(\lambda I - A) = \lambda^n + c_1\lambda^{n-1} + c_2\lambda^{n-2} + \cdots + c_n\lambda^0$

**Cayley-Hamilton Theorem**
If we recall the characteristic polynomial of a matrix $A$ we can use the *Cayley-Hamilton Theorem* and replace $\lambda$ in the above expression by $A$, so we get:

$$\chi(A) = A^n + c_1A^{n-1} + c_2A^{n-2} + \cdots + c_nA^0 = 0 \Leftrightarrow A^{-1} = \frac{-1}{c_n}A^{n-1} - \frac{c_1}{c_n}A^{n-2} - \cdots - \frac{c_{n-1}}{c_n}I$$
$$\text{so } x^* \in A^{-1}b \in x_0 + \mathcal{K}_n(r_0)$$

## 3.9.2 Convergence Rate of CG

To study the convergence rate of the conjugate gradient method, we will make some adjustments. Let $A = Q\Lambda Q^T$, where $Q$ is an orthogonal matrix ($QQ^T = I$) and $\Lambda = \text{diag}\{\lambda_1, \ldots, \lambda_n\}$. We also define $y = Q^Tx$, $\bar{b} = Q^Tb$ and $y^* = Q^Tx^*$. We then express the "quadratic" objective function in terms of $y, \bar{b}$ and $\Lambda$, which gives:

$$\begin{aligned}\phi(x) &= \frac{1}{2}x^TAx - b^Tx = \frac{1}{2}x^TQ\Lambda Q^Tx - b^TQQ^Tx \\ &= \frac{1}{2}y^T\Lambda y - \bar{b}^Ty \equiv \bar{\phi}(y) \\ &= \sum_{i=1}^{n}\left(\frac{1}{2}\lambda_i y_i^2 - \bar{b}_i y_i\right)\end{aligned}$$

so $y_i^* = \bar{b}_i/\lambda_i$ and $\bar{\phi}(y^*) = \frac{-1}{2}\sum_{i=1}^{n}(\bar{b}_i^2/\lambda_i)$.
We'll make some further simplifications; for example $x_0 = 0$ so that $r_0 = -b$
Applying this "change of basis" $Q$ gives $y_0 = 0$ and $\bar{r}_0 = -Q^Tb = -\bar{b}$ thus we have

$$y_k = \text{argmin}_{y \in \overline{\mathcal{K}}_k(\bar{b})}\bar{\phi}(y) \qquad \text{where } \overline{\mathcal{K}}_k(\bar{b}) = \{\bar{b}, \Lambda\bar{b}, \ldots, \Lambda^{k-1}\bar{b}\}$$

Since $y_k \in \overline{\mathcal{K}}_k(\bar{b})$, we have that $y_k = P_k(\Lambda)\bar{b}$ and thus we have that $[y_k]_i = P_k(\lambda_i)\bar{b}_i$ where $\deg(P_k) < k$ and

$$P_k = \text{argmin}_{P|\deg(P)<k}\sum_{i=1}^{n}\left(\frac{1}{2}\lambda_i[y_k]_i^2 - \bar{b}_i[y_k]_i\right) = \text{argmin}_{P|\deg(P)<k}\sum_{i=1}^{n}\bar{b}_i^2\left(\frac{1}{2}\lambda_i P(\lambda_i)^2 - P(\lambda_i)\right)$$

which will help with finding the convergence rate. We then have the following expression:

$$\begin{aligned}
\phi(x_k) - \phi(x^*) &= \bar{\phi}(y_k) - \bar{\phi}(y^*) \\
&= \min_{P|\deg(P)<k} \sum_{i=1}^{n} \bar{b}_i^2 \left( \frac{1}{2}\lambda_i P(\lambda_i)^2 - P(\lambda_i) \right) + \frac{1}{2}\sum_{i=1}^{n} \frac{\bar{b}_i^2}{\lambda_i} \\
&= \min_{P|\deg(P)<k} \frac{1}{2}\sum_{i=1}^{n} \bar{b}_i^2 \frac{(\lambda_i P(\lambda_i) - 1)^2}{\lambda_i} \\
&= \min_{P|\deg(P)<k} \frac{1}{2}\sum_{i=1}^{n} (y_i^*)^2 \lambda_i (\lambda_i P(\lambda_i) - 1)^2 \\
&= \min_{q|\deg(q)\leq k, q(0)=-1} \frac{1}{2}\sum_{i=1}^{n} (y_i^*)^2 \lambda_i q(\lambda_i)^2
\end{aligned}$$

and recall that we were working with respect to a "relative" rate of convergence, so we then have that

$$\begin{aligned}
\frac{\phi(x_k) - \phi(x^*)}{\phi(0) - \phi(x^*)} &= \frac{\min_{q|\deg(q)\leq k, q(0)=-1} \frac{1}{2}\sum_{i=1}^{n} (y_i^*)^2 \lambda_i q(\lambda_i)^2}{\sum_{i=1}^{n} (y_i^*)^2 \lambda_i} \\
&\leq \min_{q|\deg(q)\leq k, q(0)=-1} \left( \max_{i=1,\ldots,n} q(\lambda_i)^2 \right)
\end{aligned}$$

(because if we just take the largest element of the polynomial $q(\lambda_i)^2$, then whatever is left will cancel out and this is surely greater than or equal to the other side)

**Remarks**

- If there is a polynomial $q$ of degree $k$, with $q(0) = -1$, that is *small* on the spectrum of $A$, then $\phi(x_k) - \phi(x^*)$ is small too
- If $A$ has $m < n$ distinct eigenvalues, then CG converges in at most $m$ iterations
- If the eigenvalues of $A$ are clustered into $m < n$ groups, then $x_m$ is a good approximate solution
- If $x^*$ is approximately a linear combination of $m$ eigenvectors, then $x_m$ is a good approximate solution

**Numerical Considerations**

- CG can be implemented so that each iteration requires a few inner products of $n$-dimensional vectors and one matrix-vector multiplication $Ap$
- For *dense* matrices $A$, the matrix-vector multiplication is $O(n^2)$, so the total cost is $O(n^3)$, which is the same as the standard method
- For *sparse* matrices $A$ (or other structures), the matrix-vector multiplication can be much faster; this is where we save computationally
- Because of round-off error that occurs, CG can work poorly (or not at all); if arithmetic was exact, it would converge in at most $n$ steps
- But for favorable $A$ and $b$, we can get good approximate solutions in $\ll O(n)$ iterations; thus the overall complexity is $\ll O(n^3)$

**Preconditioned Conjugate Gradient**

- We apply CG after performing a linear change of coordinates $x = Ty$; where $\det(T) \neq 0$
- Use CG to solve $T^T ATy = C^T b$, then set $x^* = Ty$
- $T$ or $M = TT^T$ is called the **preconditioner**
- In practice, $T$ is chosen very heuristically (Cluster eigenvalues, make $T^T AT$ more diagonal, better scaling/conditioning)
- This is all done in order to improve performance

# Chapter 4

# Constrained Minimization

A constrained optimization problem follows the following structure:

$$\min_{x \in \mathbb{R}^n} \; f(x) \tag{4.1}$$

where $x \in \mathbb{R}^n$ and $f$ is a function $f : \mathbb{R}^n \to \mathbb{R}$ subject to

$$c_i(x) = 0 \quad i \in \mathcal{E}$$

$$c_i(x) \geq 0 \quad i \in \mathcal{I}$$

where we call the **Feasible Set** $\Omega$ the set of $x \in \mathbb{R}^n$ where the constraints are satisfied

**Example 4.0.1.**

$$\min \quad f(x) = \frac{1}{2}\|Ax - b\|_2^2$$

subject to

$$\|x\|_1 \leq 1 \Rightarrow c_1(x) = 1 - \|x\|_1 \qquad \mathcal{I} = \{1\},\ \mathcal{E} = \emptyset$$

If we make the additional assumption that $x \in \mathbb{R}^2$, then $\|x\|_1 = |x_1| + |x_2| \leq 1$, so we get the "unit diamond" centered around zero; which becomes the feasible region $\Omega$. We can re-write $c_1(x)$ as four separate equations that essentially give the diamond

**Definition 4.0.1.** A point $x^*$ is a **local solution** if there exists a neighborhood $\mathcal{N}$ of $x^*$ such that $f(x^*) \leq f(x)$ for all $x \in \mathcal{N} \cup \Omega$. Likewise, a point is a **strict local solution** if there exists a ... $f(x^*) < f(x)$ ...

**Example 4.0.2.**

$$\min_{x \in \mathbb{R}^2} \quad f(x) = \max\{x^2, x\}$$

which is technically a non-smooth optimization problem. Some additional terminolgy will help us turn this into a smooth optimization problem with constraints.

**Definition 4.0.2.** The **graph** of $f : \mathbb{R}^n \to \mathbb{R}$ is $\{(x, t) \in \mathbb{R}^n \text{ x } \mathbb{R} \mid f(x) = t\}$

**Definition 4.0.3.** The **epigraph** of $f : \mathbb{R}^n \to \mathbb{R}$ is $\{(x, t) \in \mathbb{R}^n \text{ x } \mathbb{R} \mid f(x) \leq t\}$

**Example 4.0.3.**

$$\min_{x \in \mathbb{R}^2} \quad f(x) = \max\{x^2, x\}$$

can be re-written as:

$$\min_{t,x} t$$

such that $t \geq x^2$ and $t \geq x$

**Example 4.0.4.**

$$\min_{x\in\mathbb{R}^2} \quad f(x) = \frac{1}{2}\|Ax-b\|_2^2 + \gamma\|x\|_1$$

where we recall that $\|x\|_1 = \sum_{i=1}^n |x_i|$. Additionally, if we let $y, z \in \mathbb{R}^n$ such that $x = y - z$, then

$$\|x\|_1 = \|y-z\|_1 \leq \|y\|_1 + \|z\|_1 = (y+z)^T\mathbb{1}$$

so the optimization problem can be written as:

$$\min_{y,z\in\mathbb{R}^2} \quad f(x) = \frac{1}{2}\|A(y-z)-b\|_2^2 + \gamma(y+z)^T\mathbb{1}$$

such that $y \geq 0$ and $z \geq 0$

**Example 4.0.5.**

$$\min_{x\in\mathbb{R}^2} \quad f(x) = x_1 + x_2$$

subject to

$$c_1(x) = x_1^2 + x_2^2 - 2 = 0 \quad \mathcal{E} = \{1\},\ \mathcal{I} = \emptyset$$

where the solutions are basically exact points on the circle of radius $\sqrt{2}$

**Remark**
Suppose we have $x \in \Omega$. Can we take a small step to $x + s$ while staying in $\Omega$ and decreasing $f$? i.e. $x + s \in \Omega$ and $f(x + s) < f(x)$. If this is not possible, then $x$ is a local solution.
We note the first order Taylor expansion:

$$f(x+s) \simeq f(x) + s^T\nabla f(x) \Rightarrow s^T\nabla f(x) < 0 \ \text{ for a decrease in } f(x)$$

and, in the constraint:

$$c_1(x+s) \simeq c_1(x) + s^T\nabla c_1(x) = 0 \Rightarrow s^T\nabla c_1(x) = 0 \ \text{ for } x+s \in \Omega$$

If $\nabla c_1(x) \neq 0$ and $\nabla f(x) \neq 0$ and $\nabla f(x) = \lambda_1 \nabla c_1(x)$ ($\lambda_1 \neq 0$), then there exists no such direction that we can move.

**Definition 4.0.4. Lagrangian functions** are given by $\mathcal{L}(x, \lambda_1) = f(x) - \lambda_1 c_1(x)$. We also note that

$$\nabla_x \mathcal{L}(x, \lambda_1) = \nabla f(x) - \lambda_1 \nabla c_1(x) = 0$$

which gets us back to our previous result and if $(x^*, \lambda_1^*)$ is a local solution, then $\nabla_x \mathcal{L}(x^*, \lambda_1^*) = 0$, where necessarily $x^* \in \Omega$. Additionally we have $\lambda_1^* \nabla c_1(x^*) = 0$ and $\lambda_1^* \geq 0$

**Example 4.0.6.**

$$\min_{x\in\mathbb{R}^2} \quad f(x) = x_1 + x_2$$

subject to

$$c_1(x) = x_1^2 + x_2^2 - 2 = 0 \quad \mathcal{E} = \{1\},\ \mathcal{I} = \emptyset$$

so the respective gradients are (PLOT THESE):

$$\nabla f(x) = (1,1)^T \qquad \nabla c_1(x) = (-2x_1, -2x_2)^T$$

**Remark**

1. The above is necessary but not a sufficient condition
2. $\lambda_1$ can be either greater than or less than 0 (or even exactly 0 if $\nabla_x f(x^*) = 0$)

**Example 4.0.7.**

$$\min_{x \in \mathbb{R}^2} \quad f(x) = x_1 + x_2$$

subject to

$$c_1(x) = -x_1^2 - x_2^2 + 2 \geq 0 \quad \mathcal{E} = \emptyset,\ \mathcal{I} = \{1\}$$

so the respective gradients are (PLOT THESE):

$$\nabla f(x) = (1, 1)^T \qquad \nabla c_1(x) = (-2x_1, -2x_2)^T$$

so now the feasible set is the set of points INSIDE the circle of radius $\sqrt{2}$, instead of just around the ball. To find the right point to minimize everything, we want $x \to x + s$, so to have $f(x + s) < f(x)$, we need $s^T \nabla f(x) < 0$ but in order to stay feasible;

$$c_1(x + s) \simeq c_1(x) + s^T \nabla c_1(x) \geq 0$$

and so if $c_1(x) > 0$, then we are not on the boundary and the constraint is "inactive" and any direction is okay for a small enough step-size $\|s\|$; thus a local solution is one in which $\nabla f(x^*) = 0$. If $c_1(x) = 0$, then the boundary is "active", and thus we need $s^T \nabla c_1(x) \geq 0$ in order to stay feasible (See Daniel Banh's paper). The steps that remain feasible and provide a decrease becomes **empty** if $\nabla f(x) = \lambda_1 \nabla c_1(x)$ with $\lambda_1 > 0$

**Example 4.0.8.**

$$\min_{x \in \mathbb{R}^2} \quad f(x) = x_1 + x_2$$

subject to

$$c_1(x) = -x_1^2 - x_2^2 + 2 \geq 0 \quad c_2(x) = x_2 \geq 0 \quad \mathcal{E} = \emptyset,\ \mathcal{I} = \{1, 2\}$$

so the Lagrangian for this problem is:

$$\mathcal{L}(x, \lambda) = f(x) - \lambda_1 c_1(x) - \lambda_2 c_2(x)$$

and $(x^*, \lambda^*)$ is a local solution then we need: $\nabla_x \mathcal{L}(x^*, \lambda^*) = 0$ with $\lambda_i^* \geq 0$ and $\lambda_i^* c_i(x^*) = 0$ for $i = \{1, 2\}$. We have that the local solution is the left-corner of the circle, since it is the point that provides the lowest $f(x)$ that is still within the feasible set

To stay feasible, we said we need the following:

$$c_i(x + s) \simeq c_i(x) + s^T \nabla c_i(x)$$

where, in addition:

$$\text{For } i \in \mathcal{E} : s^T \nabla c_i(x) = 0$$

and for "active" $i \in \mathcal{I}$ (i.e. $c_i(x) = 0$), we need $s^T \nabla c_i(x) \geq 0$
For $x \in \Omega$, we have that $z_1, z_2, \ldots$ is a sequence approaching $x$ if $z_i \to x$ and $z_k \in \Omega$ for large enough $k$

**Definition 4.0.5.** A vector $d$ is tangent to $\Omega$ at $x \in \Omega$ if there exists $z_k \to x$ for $z_k \in \Omega$ and positive scalars $t_k \to 0$ such that

$$\lim_{k \to \infty} \frac{z_k - x}{t_k} = d$$

**Definition 4.0.6.** The set of all tangent directions to $\Omega$ at $x$ is called the **Tangent Cone**, denoted $T_\Omega(x)$; it is the set of directions we can move in and still remain feasible (double check)

**Definition 4.0.7.** A set $C \in \mathbb{R}^n$ is a cone if for all $x \in C$, $ax \in C$, where $a \in \mathbb{R}^+$

**Example 4.0.9.**

$$\min_{x \in \mathbb{R}^2} \quad f(x) = x_1 + x_2$$

subject to

$$c_1(x) = -x_1^2 - x_2^2 + 2 \geq 0 \quad c_2(x) = x_2 \geq 0 \quad \mathcal{E} = \emptyset,\ \mathcal{I} = \{1, 2\}$$

where we know that the local solution is $x^* = [-\sqrt{2}, 0]^T$. Declare the following:

$$z_k = \left[ -\sqrt{2 - \frac{1}{k^2}}, -\frac{1}{k} \right]^T$$

and let $t_k = \|z_k - x\|_2 \to 0$. Then we'll get that:

$$\lim_{k \to \infty} \frac{z_k - x}{t_k} = [0, -1]^T$$

**Definition 4.0.8.** The set of linearized feasible directions at $x \in \Omega$ is defined as:

$$\mathcal{F}(x) = \{d \in \mathbb{R}^n \mid d^T \nabla c_i(x) = 0 \ \forall i \in \mathcal{E} \text{ and } d^T \nabla c_i(x) \geq 0 \ \forall i \in \mathcal{I} \cap \mathcal{A}(x)\}$$

where $\mathcal{A}(x) = \{i \in \mathcal{I} \mid c_i(x) = 0\} \cup \mathcal{E}$. We also note that $\mathcal{F}(x)$ is also a cone!

**Definition 4.0.9.** We say that a linear independence constraint qualifications (LICQ) hold at $x \in \Omega$ if

$$\{\nabla c_i(x) \mid i \in \mathcal{A}(x)\}$$

is linearly independent

## 4.1 First-Order Optimality Conditions

Let $S \subset \mathbb{R}^N$, where $S \neq \emptyset$ and $S$ is closed. Let $\bar{x} \in S$.

**Definition 4.1.1.** Recall the definition of the **Tangent Cone**

$$T_S(\bar{x}) = \left\{ d \in \mathbb{R}^N \mid \exists \{x_n\} \in S \text{ s.t } x_n \to \bar{x}, \{t_n\} \to 0 \ : \ \frac{x_n - \bar{x}}{t_n} \to d \right\}$$

where this is closed and convex *if* $S$ is convex

**Proposition 4.1.1.** (Basic First Order Necessary Condition) Let $S \subset \mathbb{R}^N$ be a closed set and let $\bar{x} \in S$ be a local minimizer of $\min f(x)$ such that $x \in S$ with $f : \mathbb{R}^N \to \mathbb{R}$ being continuously differentiable (i.e. smooth) then the following hold:

1. $\nabla f(\bar{x})^T d \geq 0$ where $d \in T_S(\bar{x})$
2. If $S$ is convex, then $\nabla f(\bar{x})^T (x - \bar{x}) \geq 0$ for $x \in S$

*Proof.*
(1) Let $d \in T_S(\bar{x})$ i.e. let there be a sequence of points $\{x_n\} \in S$ such that they converge to $\bar{x}$ and there exists another sequences $\{t_n\}$ such that decreases to zero, thus $\frac{x_n - \bar{x}}{t_n} \to d$. Since $x_n \in S$ and $x_n \to \bar{x}$ AND $\bar{x}$ is a local minimum of $f$ over $S$, then we have that $0 \leq f(x_k) - f(\bar{x})$ for sufficiently large $k$, basically due to continuity of $f$ ($\star$). Moveover, by the Mean-Value Theorem (MVT), we have that for all $n \geq \mathbb{N}, \exists \theta_k \in [x_k, \bar{x}] : f(x_k) - f(x) = \nabla f(\theta_k)^T (x_k - \bar{x})$, we'll call this ($\star\star$). We also note that since $t_k > 0$, then

$$0 \leq \frac{f(x_k) - f(\bar{x})}{t_k} = \underbrace{\nabla f(\underbrace{\theta_k}_{\to \bar{x}})^T}_{\to \nabla f(\bar{x})} \underbrace{\left( \frac{x_k - \bar{x}}{t_k} \right)}_{\to d} \to \nabla f(\bar{x})^T d$$

(2) Assume that there exists a $\hat{x} \in S$ such that $\nabla f(\bar{x})^T(\hat{x} - \bar{x}) < 0$, so essentially suppose not. Since $S$ is convex, we can take a line between $\hat{x}$ and $\bar{x}$ and still remain in $S$, i.e. $\bar{x}+\lambda(\hat{x}-\bar{x}) = \lambda\hat{x}+(1-\lambda)\bar{x} \in S$. Once again, we have by the MVT that, for all $\lambda \in (0,1)$, there exists $\theta_\lambda \in [\hat{x}, \bar{x}] : f(\bar{x} + \lambda(\hat{x} - \bar{x})) - f(\bar{x}) = \lambda \nabla f(\theta_\lambda)^T(\hat{x} - \bar{x}) < 0$ due to the initial assumption and for very small $\lambda$. Thus we have that $f(\bar{x}) > f(\bar{x} + \lambda(\hat{x}) - \bar{x})$, which is a contradiction! □

**Lemma 4.1.0.1.** *Let $\bar{x} \in \Omega$, then $T_\Omega(\bar{x}) \subseteq F(\bar{x})$ where we recall*

$$\mathcal{F}(\bar{x}) = \{d \mid \nabla c_i(x)^T d = 0 \ \forall i \in \mathcal{E}, \nabla c_j(x)^T d \geq 0 \ \forall j \in \mathcal{I} \cup \mathcal{A}(\bar{x})\} \tag{4.2}$$

*Proof.* Take $d \in T_\Omega(\bar{x})$ and also let $i \in \mathcal{A}(\bar{x})$, thus $c_i(\bar{x}) = 0$. By MVT, we have that for all $k \in \mathbb{N}$, there exists $\theta_k \in [x_k, \bar{x}] : c_i(x_k) - c_i(\bar{x}) = \nabla c_i(\theta_k)^T(x_k - \bar{x})$ but recall that since $t_k > 0$ we have that

$$\frac{c_i(x_k) - c_i(\bar{x})}{t_k} = \nabla c_i(\theta_k)^T \left( \frac{x_k - \bar{x}}{t_k} \right) \to \nabla c_i(\bar{x})^T d$$

(see notes) □

**Definition 4.1.2.** (ACQ) Let $\bar{x} \in \Omega$ be feasible for the minimization problem. Then we say that the **Abadie Constraint Qualification** holds at $\bar{x}$ if $T_\Omega(\bar{x}) = F(\bar{x})$; so due to the previous lemma, all this implies is that we need $F(\bar{x}) \subseteq T_\Omega(\bar{x})$ for this definition to hold.

**Proposition 4.1.2.** (Farkas' Lemma) Let $B \in \mathbb{R}^{lxn}$ and $h \in \mathbb{R}^n$ then TFAE

1. The system "$B^T x = h, \vec{x} \geq 0''$ has a solution (where this means that all $x_i \geq 0$
2. $h^T d \geq 0$ (where $d : Bd \geq 0$)

*Proof.*
(1) $\Rightarrow$ (2) Let $x \geq 0$ such that $B^T x = h$. Now take $d$ such that $Bd \geq 0$. Thus we have that $h^T d = (B^T x)^T d = x^T(Bd) \geq 0$, since both $x^T$ and $Bd$ are "greater then or equal to" than 0 □
(2) $\Rightarrow$ (1) (Proof by Contrapositive); define $\mathcal{K} := \{B^T x \mid x \geq 0\}$. Then we have that $\mathcal{K}$ is non-empty, closed and convex. We assume that (1) is false; thus $h \notin \mathcal{K}$. Now we consider the following optimization problem

$$\min_s f(s) := \frac{1}{2}\|s - h\|^2 \quad \text{such that } s \in \mathcal{K} \quad (\star)$$

where the solution to the above optimization problem is the *projection* of $h$ onto the set $\mathcal{K}$! Now, we use the previous proposition (in the convex case) that $\nabla f(\bar{s})^T(s - \bar{s}) \geq 0$ where : $s \in \mathcal{K}$ and $\nabla f(\bar{s}) = \bar{s} - h = \bar{d}$. Putting this all together, we have that $\bar{d}^T(s - \bar{s}) \geq 0$. (Following part makes no sense) Inserting $s = 0$ and $s = 2\bar{s}$, we have that $\bar{d}^T\bar{s} = 0$ thus $\bar{d}^T s \geq 0$ for all $s \in \mathcal{K}$. By the definition of $\mathcal{K}$, we have that $\bar{d}^T(B^T x) \geq 0$, where $x \geq 0$; using transpose-properties, we have that $(B\bar{d})^T x \geq 0$. We now let $x = \hat{e}_i$, thus $[Bd]_i \geq 0$ for all $i \in [1, l]$. Thus, $B\bar{b} \geq 0$ since each of the components are. But we also have the following

$$h^T\bar{d} = (\bar{s} - \bar{d})^T\bar{d} = \bar{s}^T\bar{d} - \bar{d}^T\bar{d} = -\|d\|_2 < 0$$

Thus, $\neg(1)$ implies $\neg(2)$, so we're done! □

**Theorem 4.1.1.** *(KKT Conditions) Let $\bar{x} \in \Omega$ be a local minimizer to an optimization problem such that **ACQ** holds at $\bar{x}$, then there exists a $\lambda_i \in \mathbb{R}$ such that*

1. $\nabla f(\bar{x}) - \sum_{i \in \mathcal{A}(\bar{x})} \bar{\lambda}_i \nabla c_i(\bar{x}) = 0$
2. $c_i(\bar{x}) = 0$ *for* $i \in \mathcal{E}$ *(Feasibility)*
3. $c_i(\bar{x}) \geq 0$ *for* $i \in \mathcal{I}$ *(Feasibility)*
4. $\bar{\lambda}_i \geq 0$
5. $\bar{\lambda}_i c_i(\bar{x}) = 0$ *for* $i \in \mathcal{E} \cup \mathcal{I}$

*Proof.* Done in class; I'd use what the book provides □

**Recall: LICQ** This is said to hold at $\bar{x} \in \Omega$ if $\{\nabla c_i(\bar{x}) \mid i \in \mathcal{A}(\bar{x})\}$ is a set of linearly independent vectors

**Theorem 4.1.2.** *(Implicit Function Theorem) Let $F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ be a function such that:*

1. *$F(z^\star, 0) = 0$ for some $z^\star \mathbb{R}^n$*
2. *The function $F()$ is continuously differentiable in some neighborhood of $(z^\star, 0)$*
3. *$\nabla_z F(z,t)$ is non-singular at the point $(z,t) = (z^\star, 0)$*

*Then there exists open sets $\mathcal{N}_z \subset \mathbb{R}^n$ and $\mathcal{N}_t \subset \mathbb{R}^m$ containing $z^\star$ and $0$, respectively, and a continuous function $z : \mathcal{N}_t \to \mathcal{N}_z$ such that $z^\star = z(0)$ and $h(z(t), t) = 0$ for all $t \in \mathcal{N}_t$. Further, we have that $z(t)$ is uniquely defined. Finally, if $h$ is $p$ times continuously differentiable with respect to both its arguments for some $p > 0$, then $z(t)$ is also $p$ times continously differentiable with respect to $t$, and we have*

$$\nabla z(t) = -\nabla_t h(z(t), t)[\nabla_z h(z(t), t)]^{-1}$$

*for all $t \in \mathcal{N}_t$*

**Lemma 4.1.2.1.** *LICQ = ACQ at $\bar{x} \in \Omega$*

*Proof.* Look in textbook (Nocedal and Wright). □

**Corollary 4.1.2.1.** *Let $\bar{x} \in \Omega$ be a local min of $(P)$ such that LICQ holds at $\bar{x}$. Then there exists a unique! $\lambda \in \mathbb{R}^{|\mathcal{E}|+|\mathcal{I}|}$ such that $(\bar{x}, \bar{\lambda})$ satisfies the KKT conditions*

*Proof.* This follows fom Lemma 2 (the one just above) and Theorem 1 (Proposition 4.1.1?), which show existence. Then LICQ determines the uniqueness (or something) □

(There was an example done in class)

**Example 4.1.1.** Given $\alpha_1, \ldots, \alpha_n$

$$\min -\sum_{i=1}^{n} \log(\alpha_i + x_i) \qquad \text{s.t. } \sum_{i=1}^{n} x_i = 1 \text{ and } x_i \geq 0 \ \forall i \in [1, n]$$

so we can set up the Lagrangian for the problem

$$\mathcal{L}(x, \lambda) = -\sum_{i=1}^{n} \log(\alpha_i - x_i) - \sum_{i=1}^{n} \lambda_i x_i - \lambda_l \Big( \sum_{i=1}^{n} x_i - 1 \Big)$$

so if $x^\star$ is a local solution, then there exists a $\lambda^\star$ such that

$$\frac{\partial \mathcal{L}(x^\star, \lambda^\star)}{\partial x_j} = \frac{-1}{\alpha_j - x_j^\star} - \lambda_j^\star - \lambda_l^\star = 0 \quad \forall j = 1, \ldots, n$$
$$\Rightarrow 0 \leq \lambda_j^\star = \frac{-1}{\alpha_j - x_j^\star} - \lambda_l^\star$$

and for complementary conditions, we need

$$x_i^\star \big( \frac{-1}{\alpha_i + x_i^\star} + \lambda_l^\star \big) \quad \forall j = 1, \ldots, n$$

so finally, we note that $\lambda_l^\star \geq 1/(\alpha_i + x_i^\star)$.

- If $\lambda_l^\star < 1/(\alpha_i)$ then $x_i^\star > 0$. Using the complementary conditions, we have that $x_i = 1/(\lambda_l^\star) - \alpha_i$
- If $\lambda_l^\star \geq 1/\alpha_i$, then $\lambda_l^\star$, then $\lambda_l^\star - 1/(\alpha_i + x_i^\star) \neq 0$ for any $x_i^\star \geq 0$, thus $x_i^\star = 0$

So to summarize these two cases, we have $x_i^\star = \max\{0, \frac{1}{\lambda_i} - \alpha_i\}$ and then we also need that $\sum_{i=1}^n \max\{0, \frac{1}{\lambda_i} - \alpha_i\} = 1$

Let $x^\star$ and $\lambda^\star$ satisfy the KKT conditions and assume that $LICQ$ holds at $x^\star$. Recall the set of linearized feasible directions $\mathcal{F}(x^\star)$. For $w \in \mathcal{F}(x^\star)$, then $w^T \nabla f(x^\star) > 0$ or $w^T \nabla f(x^\star) = 0$.

**Definition 4.1.3.** Define the **critical cone** at $x^\star, \lambda^\star$ to be

$$\mathcal{C}(x^\star, \lambda^\star) = \{w \in \mathcal{F}(x^\star) \mid w^T \nabla c_i(x^\star) = 0 \ \forall i \in \mathcal{I} \cap \mathcal{A}(x^\star) \text{ with } \lambda_i^\star > 0\}$$

so for $w \in \mathcal{C}(x^\star, \lambda^\star)$ then TFAE

1. $w^T \nabla c_i(x^\star) = 0$ for $i \in \mathcal{E}$
2. $w^T \nabla c_i(x^\star) = 0$ for $i \in \mathcal{I} \cap \mathcal{A}(x^\star)$ with $\lambda_i^\star > 0$
3. $x^T \nabla c_i(x^\star) \geq 0$ for $i \in \mathcal{I} \cap \mathcal{A}(x^\star)$ with $\lambda_i^\star = 0$

From the following KKT condition

$$\nabla_x \mathcal{L}(x^\star, \lambda^\star) = \nabla f(x^\star) - \sum_{i \in \mathcal{E} \cup \mathcal{I}} \lambda_i^\star \nabla c_i(x^\star) = 0$$

and if $w \in \mathcal{C}(x^\star, \lambda^\star)$ then we have that

$$w^T \nabla_x \mathcal{L}(x^\star, \lambda^\star) = w^T \nabla f(x^\star) - \underbrace{\sum_{i \in \mathcal{E} \cup \mathcal{I}} \lambda_i^\star w^T \nabla c_i(x^\star)}_{=0} = 0 \Rightarrow w^T \nabla f(x^\star) = 0$$

**Second Order Necessary Conditions**

If $x^\star$ is a local solution and LICQ holds and $(x^\star, \lambda^\star)$ satisfy KKT, then $w^T \nabla_{xx} \mathcal{L}(x^\star, \lambda^\star) w \geq 0$ for all $w \in \mathcal{C}(x^\star, \lambda^\star)$

**Second Order Sufficient Conditions**

Suppose $x^\star$ is feasible and $(x^\star, \lambda^\star)$ satisfy the KKT conditions. If $w^T \nabla_{xx} \mathcal{L}(x^\star, \lambda^\star) w > 0$ for all $w \in \mathcal{C}(x^\star, \lambda^\star)$, $w \neq 0$, then $x^\star$ is a strict local solution.

**Remark** Suppose $f$ is strictly convex ($\nabla^2 f \succ 0$) and $-c_i(x)$ is convex for all $i \in \mathcal{E} \cup \mathcal{I}$ ???

**Remark** If $\mathcal{A}(x^\star) \cap \mathcal{I} = \emptyset$ then $\mathcal{C}(x^\star, \lambda^\star) = \mathcal{F}(x^\star)$

**Example 4.1.2.** Consider the "perturbed" problem

$$\min_{x \in \mathbb{R}} x^2 \qquad \text{such that } x - 1 \geq -\epsilon$$

For the non-perturbed problem, the answer is obviously $x^\star = 1$ with $\lambda^\star = 2$, where the constraints are "strongly active". Now, we say that $x^\star(\epsilon) = 1 - \epsilon$. We have that $f(x^\star(\epsilon)) = (1 - \epsilon)^2 \simeq 1 - 2\epsilon$, thus

$$\frac{df(x^\star(\epsilon))}{d\epsilon} \simeq -2$$

**Example 4.1.3.** Consider the "perturbed" problem

$$\min_{x \in \mathbb{R}} x^2 \qquad \text{such that } x \geq 0$$

then $x^\star = 0 = \lambda_1^\star$, thus this is "weakly active" since $\lambda_1^\star = 0$

**Example 4.1.4.** Consider the "perturbed" problem

$$\min_{x\in\mathbb{R}} x^2 \qquad \text{such that } x \geq -\epsilon$$

then if $\epsilon > 0$, then $x^\star = 0$ and the constraint is no longer active. For $\epsilon < 0$, then $x^\star = -\epsilon$

So what happens in most of these scenarios is that we changed $c_i(x) \geq 0$ to $c_i(x) \geq -\epsilon\|\nabla c_i(x^\star)\|$. We assume that $i \in \mathcal{A}(x^\star)$ and $\mathcal{A}(x^\star(\epsilon)) \equiv \mathcal{A}(x^\star)$. What is $f(x^\star(\epsilon)) - f(x^\star) \simeq \nabla f(x^\star)^T(x^\star(\epsilon) - x^\star)$? From KKT1, there exists a $\lambda^\star$ such that

$$\nabla_x \mathcal{L}(x^\star, \lambda^\star) = \nabla f(x^\star) - \sum_{j\in\mathcal{E}\cup\mathcal{I}} \lambda_j^\star \nabla c_j(x^\star) = 0$$

for $j \notin \mathcal{A}(x^\star)$, we have by complementary conditions that $\lambda_j^\star = 0$ and, for $j \in \mathcal{A}(x^\star)$

$$\begin{aligned} c_j(x^\star(\epsilon)) &\simeq c_j(x^\star) + \nabla c_j(x^\star)^T(x^\star(\epsilon) - x^\star) \\ \Rightarrow \nabla c_j(x^\star(\epsilon))^T(x^\star(\epsilon) - x^\star) &\sim c_j(x^\star(\epsilon)) - c_j(x^\star) \qquad = -\epsilon\|\nabla c_i(x^\star)\| \text{ if } j = i \\ &\qquad\qquad\qquad\qquad\qquad\qquad = 0 \text{ if } j \in \mathcal{A}(x^\star), j \neq i \end{aligned}$$

so

$$\begin{aligned} \frac{1}{\epsilon}(f(x^\star(\epsilon)) - f(x^\star)) &\simeq \sum_{i\in\mathcal{A}(x^\star)} \lambda_i^\star c_j(x^\star)^T(x^\star(\epsilon) - x^\star) \\ &= -\lambda_i^\star \|\nabla c_i(x^\star)\| \\ \Rightarrow \frac{df(x^\star(\epsilon))}{d\epsilon} &= -\lambda_i^\star \|\nabla c_i(x^\star)\| \end{aligned}$$

## 4.2 Duals

Let's say we have the following minimization problem

$$\min_{x\in\mathbb{R}^n} f(x) \qquad \text{such that } c_i(x) = 0, i \in \mathcal{E} \text{ and } c_i(x) \geq 0, j \in \mathcal{I}$$

then if we define $c(x) = [c_1(x), c_2(x), \ldots, c_{|\mathcal{E}|+|\mathcal{I}|}(x)]$ then we have that the Lagrangian can just be written as $\mathcal{L}(x, \lambda) = f(x) - \lambda^T c(x)$.

**Definition 4.2.1.** We have that the Lagrangian **dual function**

$$q(\lambda) = \inf_{x\in\mathbb{R}^n} \mathcal{L}(x, \lambda)$$

where the domain of $q(\lambda)$ is

$$D = \{\lambda \mid q(\lambda) > -\infty\}$$

**Claim 4.** *$q$ is concave and $D$ is always convex*

*Proof.* Let $\lambda^0$ and $\lambda^1$ be given and let $\alpha \in (0, 1)$. Then

$$\begin{aligned}\mathcal{L}(x, \alpha\lambda^0 + (1-\alpha)\lambda^1) &= \alpha f(x) + (1-\alpha)f(x) - [\alpha\lambda^0 + (1-\alpha)\lambda^1]^T c(x) \\ &= \alpha\mathcal{L}(x, \lambda^0) + (1-\alpha)\mathcal{L}(x, \lambda^1)\end{aligned}$$

then we have that

$$\begin{aligned}q(\alpha\lambda^0 + (1-\alpha)\lambda^1) &= \inf_{x\in\mathbb{R}^n} \Big(\alpha\mathcal{L}(x, \lambda^0) + (1-\alpha)\mathcal{L}(x, \lambda^1)\Big) \\ &\geq \inf_{x\in\mathbb{R}^n} \Big(\alpha\mathcal{L}(x, \lambda^0)\Big) + \inf_{x\in\mathbb{R}^n} \Big((1-\alpha)\mathcal{L}(x, \lambda^1)\Big) \\ &= \alpha q(x, \lambda^0) + (1-\alpha)q(x, \lambda^1)\end{aligned}$$

and also if $\lambda^0, \lambda^1 \in D$, then $\alpha\lambda^0 + (1-\alpha)\lambda^1 \in D$ for any $\alpha \in [0, 1]$ □

**Definition 4.2.2.** The **Lagrangian Dual Problem** is

$$\max_{\lambda} q(\lambda) \qquad \text{such that } \lambda_i \geq 0 \quad i \in \mathcal{I} \tag{4.3}$$

Recall that for any feasible point $\bar{x}$ and any $\lambda$ such that $\lambda_i \geq 0$ for $i \in \mathcal{I}$, we have that

$$\mathcal{L}(\bar{x}, \lambda) = f(\bar{x}) \underbrace{- \sum_{i\in\mathcal{I}} \underbrace{\lambda_i}_{\geq 0} \underbrace{c_i(\bar{x})}_{\geq 0}}_{\leq 0} \leq f(\bar{x})$$

Next,

$$q(\lambda) = \inf_x \mathcal{L}(x, \lambda) \leq \mathcal{L}(\bar{x}, \lambda) \leq f(\bar{x})$$

and since this is true for any $\bar{x}$ (feasible), it is true for $x^\star$, so $q(\lambda) \leq f(\bar{x})$

**Definition 4.2.3.** When $q(\lambda^\star) \leq f(x^\star)$ is called **weak duality**

Let $d^\star$ be the optimal value of the dual problem, and $p^\star$ be the optimal value of the primal problem. Weak duality says that $d^\star \leq p^\star$

**Example 4.2.1.** (**Two Way Partitioning**) Split $\{1, 2, \ldots, n\}$ into two groups. Let $W_{ij} = W_{ji}$ (symmetric matrix). Problem: Pay $W_{ij}$ if items $i$ and $j$ are in the same groupp and pay $-W_{ij}$ if the items are in different groups. The optimization variables are $x_i = \{-1, +1\}$, where $i = 1, \ldots, n$. This optimization problem can be written as

$$\min x_{\mathbb{R}^n} \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} W_{ij} x_i x_j \Leftrightarrow \frac{1}{2} x^T W x \qquad \text{such that } x_i^2 = 1$$

so we can write our Lagrangian as follows

$$\begin{aligned} \mathcal{L}(x, \lambda) &= \frac{1}{2} x^T W x - \sum_{i=1}^{n} \lambda_i (x_i^2 - 1) = \lambda^T \mathbb{1} + \frac{1}{2} x^T W x - x^T \text{diag}(\lambda_i) x \\ &= \lambda^T \mathbb{1} + x^T \left( \frac{1}{2} W - \text{diag}(\lambda) \right) x \end{aligned}$$

so we can define the dual problem as follows

$$\begin{aligned} q(\lambda) &= -\infty \text{ if } \frac{1}{2} W - \text{diag}(\lambda) \text{ is not positive semidefinite} \\ &= \lambda^T \mathbb{1} \text{ if } \frac{1}{2} W - \text{diag}(\lambda) \succeq 0 \end{aligned}$$

Suppose that $\frac{1}{2} W$ has eigenvalues $\xi_1, \xi_2, \ldots, \xi_n$. Consider setting $\lambda_i = \xi_1$ (setting it to the smallest one). Then $\text{diag}(\lambda) = \xi_1 \cdot I$; then $\frac{1}{2} W - \xi_1 I \succeq 0 \Rightarrow q(\lambda) = n\xi_1 \leq p$

# 4.3 Strong Duality

Suppose $f$ is convex, $c_i$ are concave for $i \in \mathcal{I}$ and $c_i(x) = a_i^T x + b_i$ for $i \in \mathcal{E}$. Further, suppose that $x^\star$ is a solution of the primal problem and $(x^\star, \lambda^\star)$ satisfy KKT conditions (i.e. LICQ holds at $(x^\star, \lambda^\star)$. Then $\lambda^\star$ is a solution of the dual problem and $q(\lambda^\star) = f(x^\star$. We then have that for any feasible $x$

$$\mathcal{L}(x, \lambda^\star) = f(x) - \sum_{i=\mathcal{E}+\mathcal{I}} \lambda_i^\star c_i(x)$$

is just a convex function of $x$ (since the sum of convex functions is convex). Thus we have that

$$\begin{aligned} \mathcal{L}(x, \lambda^\star) &\geq \mathcal{L}(x^\star, \lambda^\star) + \nabla_x \mathcal{L}(x^\star, \lambda^\star)^T (x - x^\star) \\ &= \mathcal{L}(x^\star, \lambda^\star) \end{aligned}$$

Then

$$q(\lambda^\star) = \inf_x \mathcal{L}(x, \lambda^\star) = \mathcal{L}(x^\star, \lambda^\star) = f(x^\star)$$

## 4.3.1 Equivalency?

Consider the following simpler problem

$$\min_x f(x) \qquad \text{such that } c_i(x) \geq 0 \text{ for } i \in \mathcal{I}$$

(note that any problem with equality constraints can be decomposed into more equality constraints). For feasible $x$, we have that $\sup_{\lambda \geq 0} \mathcal{L}(x, \lambda) = f(x)$ thus we have that

$$\text{Primal: } \inf_x \sup_{\lambda \geq 0} \mathcal{L}(x, \lambda)$$

$$\text{Dual: } \sup_{\lambda \geq 0} \inf_x \mathcal{L}(x, \lambda)$$

and thus we say that we have strong duality when

$$\sup_{\lambda \geq 0} \inf_x \mathcal{L}(x, \lambda) = \inf_x \sup_{\lambda \geq 0} \mathcal{L}(x, \lambda)$$

**Definition 4.3.1.** A KKT point $(x^\star, \lambda^\star)$ is called a **saddle point** when

$$\mathcal{L}(x^\star, \lambda)\ (\forall\ \lambda \geq 0) \leq \mathcal{L}(x^\star, \lambda^\star) \leq \mathcal{L}(x, \lambda^\star)\ (\forall\ x, \text{definitely feasible but not necessarily } \textbf{all})$$

**Example 4.3.1.** (**Linear Program**) The problem is outlined as follows

$$\min_{x \in \mathbb{R}^n} c^T x \qquad \text{such that } Ax - b \geq 0$$

where $c$, $A$ and $b$ are given, where $A$ is an $m \times n$ matrix with $m < n$ and $\text{rank}(A) = m$ (so picture short-fat matrix); so the Lagrangian is the following

$$\mathcal{L}(x, \lambda) = c^x - \lambda^T(Ax - b) = (c^T - \lambda^T A)x + \lambda^T b$$

thus

$$q(\lambda) = \inf_x \mathcal{L}(x, \lambda) = \begin{cases} \lambda^T b & \text{if } A^T\lambda = c, \\ -\infty & \text{otherwise.} \end{cases}$$

thus the dual problem is $\max q(\lambda)$ such that $\lambda \geq 0$. We have that the equivalent dual problem is written as

$$\max_\lambda b^T \lambda \qquad \text{such that } \lambda \geq 0, \quad A^T\lambda = c$$

# 4.4 Quadratic Programming

Quadrtic programs have the following form

$$\min_{x \in \mathbb{R}^2} \frac{1}{2} x^T G x + c^T x \quad \text{subject to: } a_i^T x = b_i\ \imath \in \mathcal{E} \text{ and } a_i^T x \geq b_i\ \imath \in \mathcal{I}$$

**Example 4.4.1.** (**Markowitz Portfolio Optimizaion**) The problem is described as follows: there are $n$ assets with returns $r_i \sim N(\mu_i, \sigma_i^2)$, with $i = 1, \ldots, n$. Let $x_i$ be a fraction that we invest in asset $i$ (thus, $x_i \geq 0$) such that $\sum_{i=1}^n x_i = 1$. We also note the following formulas:

- Correlations: $\rho_{ij} = \mathbb{E}[(r_i - \mu_i)(r_j - \mu_k)]/\sigma_i\sigma_j$
- Expected return, given $x$: $\mathbb{E}\big[\sum_{i=1}^n x_i r_i\big] = x^T\mu$
- Risk (or variance) of $x$: $\mathbb{E}\big[\bigcap(\sum_{i=1}^n x_i r_i - \mu^T x)^2\big] = \sum_{i=1}^n \sum_{j=1}^n x_i x_j \underbrace{\sigma_i\sigma_j\rho_{ij}}_{G_{ij}} = x^T G x$

thus the problem can be formalized as

$$\max_{x \in \mathbb{R}^n} x^T\mu - k x^T G x \quad \text{subject to: } x \geq 0, x^T\mathbb{1} = \sum_{i=1}^n x_i = 1$$

where $k \geq 0$ is called the risk tolerance.

## 4.4.1 Only Equality Constraints

If $\mathcal{I} = \emptyset$, then the quadratic programming problem becomes

$$\min_{x \in \mathbb{R}^2} \frac{1}{2} x^T G x + c^T x \quad \text{subject to: } Ax = b$$

where $G, c, A, b$ are given. We have that $G = G^T \succeq 0$ and $A \in \mathbb{R}^{m \times n}$ (where $m < n$) and $\text{rank}(A) = m$, the Lagrangian is as follows

$$\mathcal{L}(x, \lambda) = \frac{1}{2} x^T G x + c^T x - \lambda^T(Ax - b)$$

where $\lambda \in \mathbb{R}^n$; using KKT, we have that $\nabla_x \mathcal{L}(x, \lambda) = 0 \Rightarrow Gx - A^T\lambda = -c$ and, in addition to the constraint equations, we have the following system

$$\begin{bmatrix} G & -A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} x \\ \lambda \end{bmatrix} = \begin{bmatrix} -c \\ b \end{bmatrix}$$

### 4.4.2 Inequality and Equality Constraints - Active Set Method

If we knew $\mathcal{A}(x^\star)$, then we would simply have to solve

$$\frac{1}{2}x^T Gx + c^T x \quad \text{subject to: } a_i^T x = b_i \;\; \text{for } i \in \mathcal{A}(x^\star)$$

**Algorithm Outline**

Given a feasible point $x_0$, we choose a "working set", denoted $w_0 \subseteq \mathcal{A}(x_0)$. We perform the updating step $x_{k+1} \leftarrow x_k + \alpha_k p_k$, and update the working set $w_{k+1}$ from $w_k$ and $x_k$ such that $x_k$ is feasible for all $k = 1, \ldots$ and $w_k \subseteq A(x_k)$

**Algorithm - Indepth**

Given $x_k$ and $w_k$, where we note that $a_i^T x_k = b_i$ with $i \in w_k \subseteq A(x_k)$, we obtain $p_k$ by solving the following *subproblem*

$$\min_p \frac{1}{2}(x_k - p)^T G(x_k + p) + c^T(x_k + p) \quad \text{s.t } a_i^T(x_k + p) = b_i \;\; \text{for } i \in w_k$$

additionally, we have the simplified subproblem to be solved at the $k$th iteration, as follows

$$\min_{p \in \mathbb{R}^n} \frac{1}{2}p^T Gp + p^T Gx_k \quad \text{s.t. } a_i^T p = 0 \;\; \text{for } i \in w_k$$

And we call $(x_k, \lambda_k$ a *solution pair*, where $\lambda_k \in \mathbb{R}^{|w_k|}$; we get $\tilde{\lambda}_k \in \mathbb{R}^{|\mathcal{E}|\cup|\mathcal{I}|}$ by setting **something I can't read**.

The Active Set Algorithm (page 469 in the textbook pages, 490 in the pdf)

**if** $p_k \neq 0$ **then**
- $x_k \leftarrow x_k + \alpha_k p_k$ ;
- Choose $\alpha_k$ so that no constraints are violated ;
- **for** $i \in \mathcal{I} \setminus w_k$ **do**
  - $\alpha_i^T x_k \geq b_i$ since $x_k$ is feasible;
  - $a_i^T(x_k + \alpha_k p_k) \geq b_i$ ;
  - **if** $a_i^T p_k > 0$ **then**
    - any $\alpha_k$ is okay
  - **end**
  - **if** $a_i^T p_k < 0$ **then**
    - This is still feasible if $\alpha_k \leq (b_i - a_i^T x_k)/a_i^T p_k$
  - **end**
  - Overall, take $\alpha_k \stackrel{\text{def}}{=} \min\left(1, \min_{i \notin w_k, a_i^T p_k < 0} b_i - a_i^T x_k)/a_i^T p_k\right)$
- **end**
- $x_{k+1} \leftarrow x_k + \alpha_k p_k$ ;
- **if** $\alpha_k < 1$ **then**
  - Set $w_{k+1} \leftarrow w_k \cup \{j\}$ where $j$ is the blocking constraint;
- **end**
- **else**
  - $w_{k+1} \leftarrow w_k 0$
- **end**

**end**

**if** $p_k = 0$ **then**
- Check if $x_k$ solves the QP; $\nabla f(x_k) = \sum_{i \in w_k} \tilde{\lambda}_{ki} a_i$ by design ;
- $x_k$ is feasible by design;
- **if** $\tilde{\lambda}_{ki} \geq 0 \ \forall i \in w_k \cap \mathcal{I}$ **then**
  - **stop** with solution $x^\star = x_k$;
- **end**
- **else**
  - $j \leftarrow \arg\min_{i \in w_k \cap \mathcal{I}} \tilde{\lambda}_i$;
  - $x_{k+1} \leftarrow x_k$; $w_{k+1} \leftarrow w_k \setminus \{j\}$
- **end**

**end**

## 4.5 Newton's Method for Nonlinear Equations

Given $F : \mathbb{R}^n \to \mathbb{R}^n$, where $F = [f_1(x) : f_2(x) : \ldots : f_n(x)]$ (where the colon denotes that this is a column vector), we want to find $F(x) = 0$. We have that

$$F(x^k + p_k) \simeq F(x^k) + J(x^k)p_k$$

where $J(x^k)$ is the Jacobian of $F$ evaluated at $x^k$. So $F(x) = 0$ essentially when $J(x^k)p_k = -F(x^k)$. To implement this concept, we go back to unconstrained minimization (i.e. $\min_{x\in\mathbb{R}^n} f(x)$) . Let $F(x) := \nabla f(x)$, then $J(x) = \nabla^2 f(x)$ thus $\nabla f(x) = 0$ when $\nabla^2 f_k p_k = -\nabla f_k$

**Interior Point Methods for QP**

Recall that a QP has the form $\frac{1}{2}x^T Gx + x^T c$ with inequality constraints such that $a_i^T x - b_i \geq 0$; we have that

$$\mathcal{L}(x, \lambda) = \frac{1}{2}x^T Gx + x^T c - \sum_{i\in\mathcal{I}} \lambda_i(a_i^T x - b_i) = \frac{1}{2}x^T Gx + x^T c - \lambda^T(Ax - b)$$

and the KKT conditions for the QP are:

1. $Gx + c - A^T\lambda = 0$
2. $Ax - b \geq 0$
3. $\lambda_i(a_i^T x - b_i) = 0$ for all $i = 1, \ldots, m$
4. $\lambda_i \geq 0$ for all $i = 1, \ldots, m$

**Introduce Slack Variables**

Let $y_i = a_i^T x - b_i$ for $i = 1, \ldots, m$. Then $a_i^T x - b_i - y_i = 0$. Also, $\lambda_i y_i = 0$ and $y_i \geq 0$.

**Equivalency of Set Conditions**

- $Gx + c - A^T\lambda = 0 \qquad (n) \quad x \in \mathbb{R}^n$
- $Ax - b - y = 0 \qquad (m) \quad \lambda \in \mathbb{R}^m$
- $Y\Lambda e = 0 \qquad (m) \quad y \in \mathbb{R}^m$

where $e = (1 : 1 : \ldots : 1$, $\Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_m)$ and $Y = \text{diag}(y_1, y_2, \ldots, y_m)$, thus we can do

$$F(x, y, \lambda) = [Gx + c - A^T\lambda : Ax - b - y : Y\Lambda e] = 0$$

where $\lambda, y \geq 0$

**An equivalent problem (QP)**

We can write the constrained QP as an unconstrained problem

$$\min \frac{1}{2}x^T Gx + x^T c + \sum_{i=1}^{m} \mathbb{I}(a_i^T x - b_i)$$

where

$$\mathbb{I}(u) = \begin{cases} 0 & \text{if } a_i^T x - b_i \geq 0, \\ \infty & \text{otherwise.} \end{cases}$$

but we can approximate $\mathbb{I}$ with a **log-barrier**,

$$\mathbb{I} \simeq \frac{1}{t}\log(a_i^T x - b_i) \qquad \text{for } t > 0$$

thus for a given $t$, solve

$$\min_x f(x) - \frac{1}{t}\sum_{i=1}^m \log(a_i^T x - b_i)$$

where the solutions (we say solutions because they vary with $t$), $\bar{x}(t)$ is called the **central path**. As $t \to \infty$ then $\bar{x}(t) \to x^\star$

Assume that $f(x)$ is convex, then the subproblem is also convex. We have that

$$t\nabla f(\bar{x}(t)) - \sum_{i=1}^m \frac{a_i}{a_i^T x - b_i} = 0$$

and we take $\bar{\lambda}_i = [t(a_i^T \bar{x} - b_i)]^{-1}$. Thus, for the constrained problem,

$$\begin{aligned}
&\mathcal{L}(x,\lambda) = f(x) - \sum_{i=1}^m \lambda_i(a_i^T x - b_i) \\
&\Rightarrow \nabla_x \mathcal{L}(x,\lambda) = \nabla f(\bar{x}) - \sum_{i=1}^m \bar{\lambda}_i a_i = \nabla f(\bar{x}) - \frac{1}{t}\sum_{i=1}^m \frac{a_i}{a_i^T \bar{x} - b_i} = 0
\end{aligned}$$

so for a fixed $\bar{\lambda}$, we have that

$$\begin{aligned}
q(\bar{\lambda}) &= \inf_x \mathcal{L}(x,\bar{\lambda}) = \inf_x f(x) - \sum_{i=1}^m \frac{a_i^T x - b_i}{t(a_i^T \bar{x} - b_i)} \\
&\Rightarrow \nabla : \nabla f(x) - \sum_{i=1}^m \frac{a_i}{t(a_i^T \bar{x} - b_i)}
\end{aligned}$$

where, we recall that, $\bar{x}$ is a minimizer to the unconstrained problem from before. Thus $q(\bar{\lambda}) \leq f(x^\star)$, the solution to the constrained problem. So we have that

$$\begin{aligned}
q(\bar{\lambda}) = \mathcal{L}(\bar{x},\bar{\lambda}) &= f(\bar{x}) - \sum_{i=1}^m \frac{a_i^T \bar{x} - b_i}{t(a_i^T \bar{x} - b_i)} \\
&= f(\bar{x}) - \frac{1}{t}m \Rightarrow f(\bar{x}) - f(x^\star) \leq \frac{m}{t}
\end{aligned}$$

### 4.5.1 Barrier Method

Given $x_0 \in \Omega^o$, $t_0 > 0$, $\gamma > 1$ and $\epsilon > 0$
For $k = 1, 2, \ldots$:
Sove Centering Problem
$x_k \leftarrow \arg\min_x t_{k-1} f(x) - \sum_{i=1}^m \log(a_i^T x - b_i)$
if $m/t_{k-1} \leq \epsilon$, stop
$t_k = \gamma t_{k-1}$
end
Note that $t_k = \gamma^k t_0$, thus $\frac{m}{t_k} = \frac{m}{\gamma^k t_0} \leq \epsilon \Rightarrow \frac{m}{t_0 \epsilon} \leq \gamma^k \Rightarrow k \geq \log(m/(t_0\epsilon))/\log(\gamma)$, thus $k = \lceil \log(m/(t_0\epsilon))/\log(\gamma) \rceil$.
Barrier is a *primal* IP (interior point) method. This is because we solve for $x_k$ and we get a set of lagrange multipliers $\lambda_k$ as a side product.

## Primal Dual IP Methods

Primal Dual Interior Point methods solve jointly for $(x, \lambda)$. Recall

- $Gx + c - A^T\lambda = 0 \qquad (n) \quad x \in \mathbb{R}^n$
- $Ax - b - y = 0 \qquad (m) \quad \lambda \in \mathbb{R}^m$
- $Y\Lambda e = 0 \qquad (m) \quad y \in \mathbb{R}^m$

where $e = (1 : 1 : \ldots : 1$, $\Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_m)$ and $Y = \text{diag}(y_1, y_2, \ldots, y_m)$, thus we can do

$$F(x, y, \lambda) = [Gx + c - A^T\lambda : Ax - b - y : Y\Lambda e] = 0$$

where $\lambda, y \geq 0$.
So we want to solve $F(x, y, \lambda) = 0$ for $y \geq 0, \lambda \geq 0$ (this is equaivalent for KKT conditions). Recall the Newton Method; $J(x, y, \lambda)p = -F(x, y, \lambda)$. So given $(x_k, y_k, \lambda_k)$, we solve

$$\begin{bmatrix} G & 0 & -A^T \\ A & -I & 0 \\ 0 & \Lambda & Y \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta y \\ \Delta\lambda \end{bmatrix} = - \begin{bmatrix} Gx_k + c - A^T\lambda_k \\ Ax_k - b - y_k \\ Y_k\Lambda_k e \end{bmatrix} = \begin{bmatrix} -r_d \\ -r_p \\ Y_k\Lambda_k e \end{bmatrix}$$

where $r_p$ and $r_d$ are the primal and dual residuals, respectively. Then $(x_{k+1}, y_{k+1}, \lambda_{k+1}) = \alpha_k(x_k, y_k, \lambda_k)$ where $\alpha_k$ is chosen such that $y, \lambda \geq 0$

### Path Flowing Method

Suppose we keep $y_k, \lambda_k > 0$. How do we measure "how far" we are from reaching the complementarity conditions? We use the complementarity measure (or duality measure)!

$$\mu = \frac{1}{m} y_k^T \lambda_k$$

and we define the **central path** as follows

$$C(\tau) = \{(x, y, \lambda) \mid Gx + c - A^T\lambda = 0; \lambda_i y_i = \tau; Ax - b - y = 0; y > 0; \lambda > 0, \tau > 0\}$$

**Problem:** Hard to find all the $\lambda_i$s and the $y_i$s such that $\lambda_i y_i = \tau$ for all $i = 1, \ldots, m$. So let's deviate from the central path a little and our *Modified KKT System* is as follows

$$\begin{bmatrix} G & 0 & -A^T \\ A & -I & 0 \\ 0 & \Lambda_k & Y_k \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta y \\ \Delta\lambda \end{bmatrix} = - \begin{bmatrix} Gx_k + c - A^T\lambda_k \\ Ax_k - b - y_k \\ Y_k\Lambda_k e - \mu_k\sigma_k e \end{bmatrix}$$

where $\sigma_k \in [0, 1]$ (See textbook for more info on how to choose $\sigma_k$)

## 4.5.2 Sequential QP

Solves more general nonlinear constrained optimization problems.

$$\min_x f(x) \qquad \text{such that } c_i(x) = 0 \; i = 1, 2, \ldots, m$$

where $f : \mathbb{R}^n \to \mathbb{R}$ and $c : \mathbb{R}^n \to \mathbb{R}^m$ are smooth (so twice differentiable), then we have that

$$\mathcal{L}(x, \lambda) = f(x) - \lambda^T c(x) \Rightarrow \nabla_x \mathcal{L}(x, \lambda) = \nabla f(x) - A(x)^T\lambda = nablaf(x) - \sum_{i=1}^{m} \lambda_i \nabla c_i(x)$$

where $A^T(x) = [\nabla c_1(x), \ldots, \nabla c_n(x)]$ also

$$\nabla_{xx}\mathcal{L}(x,\lambda) = \nabla^2 f(x) - \sum_{i=1}^{m} \lambda_i \nabla^2 c_i(x)$$

thus for the KKT conditions, we have that

$$F(x,\lambda) = \begin{bmatrix} \nabla f(x) - A(x)^T\lambda \\ c(x) \end{bmatrix} = 0$$

and the Newton Step is the following

$$\begin{bmatrix} \nabla^2\mathcal{L}_k & -A_k^T \\ A_k & 0 \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta\lambda \end{bmatrix} = - \begin{bmatrix} -\nabla f_k + A_k^T\lambda_k \\ -c(x_k) \end{bmatrix}$$

**Idea of the algorithm**

At $(x_k, \lambda_k)$ solve

$$\min_p f_k + \nabla f_k^T p + \frac{1}{2} p^T \nabla^2\mathcal{L}_k p \quad \text{s.t.} c_i(x_k) + \nabla c_i(x_k)^T p = 0 \; i = 1, \ldots, m$$

and set $(x_{k+1}, \bar{\lambda}_{k+1}) = (x_k + p_k, \bar{\lambda}_{k+1})$ then

$$\bar{\mathcal{L}}(p, \bar{\lambda}) = \nabla f_k^T p + \frac{1}{2} p^T \nabla^2\mathcal{L}_k p - \sum_{i=1}^{m} \bar{\lambda}_i \big[c_i(x_k) + \nabla c_i^T(x_k) p\big]$$

and taking the derivative with respect to $p$ gives

$$\nabla f_k + \nabla^2\mathcal{L}_k p - \underbrace{\sum_{i=1}^{m} \bar{\lambda}_i \nabla c_i(x_k)}_{A(x_k)^T\bar{\lambda}} = 0$$

thus we have that

$$\begin{bmatrix} \nabla^2\mathcal{L}_k & -A_k^T \\ A_k & 0 \end{bmatrix} \begin{bmatrix} p \\ \bar{\lambda} \end{bmatrix} = - \begin{bmatrix} -\nabla f_k \\ -c(x_k) \end{bmatrix}$$

and $c(x_k) + A(x_k)^T p = 0$, which is equivalent to the Newton Step! (write out algebraically)

## 4.6 Last Class and other stuff

Consider the following problem

$$\min_x f(x) \quad \text{s.t.} \quad c_i(x) = 0, i \in \mathcal{E}$$

then the lagrangian is

$$\mathcal{L}(x, \lambda) = f(x) - \sum_{i \in \mathcal{E}} \lambda_i c_i(x)$$

thus the KKT conditions give us

$$\nabla f(x^\star) - \sum_{i \in \mathcal{E}} \lambda_i^\star \nabla c_i(x) \quad \text{and} \quad c_i(x^\star) = 0$$

**Dual Ascent**

Start with $\lambda^0$
For $k = 1, 2, \ldots$
Find $x_k = \arg\min_x \mathcal{L}(x, \lambda^{k-1})$
$\lambda_i^k = \lambda_i^{k-1} - \alpha_k c_i(x_k)$ for all $i \in \mathcal{E}$

**Recall**

$$q(\lambda) = \inf_x \mathcal{L}(x, \lambda) \Rightarrow q(\lambda^{k-1}) = \mathcal{L}(x_k, \lambda^{k-1}) = f(x_k) - \sum_{i \in \mathcal{E}} \lambda_i^{k-1} c_i(x_k)$$

thus we have that

$$\frac{\partial}{\partial \lambda_j} q(\lambda^{k-1}) = -c_j(x_k)$$

**Remark:** Similar to the gradient descent method, the dual ascent method is sensitive to scaling (of the problem?) and makes small steps in the beginning

**Augmented Lagrangian**

Let $\mu \geq 0$

$$\mathcal{L}_A(x, \lambda, \mu) = f(x) - \sum_{i \in \mathcal{E}} \lambda_i c_i(x) + \frac{\mu}{2} \sum_{i \in \mathcal{E}} c_i^2(x)$$

which we can think of $\mathcal{L}_A$ as the lagrangian for the following problem

$$\min_x f(x) + \frac{\mu}{2} \sum_{i \in \mathcal{E}} c_i^2(x) \quad \text{s.t.} \quad c_i(x) = 0 \ \forall i \in \mathcal{E}$$

**Note:** It is easy to see that the solution to the augmented Lagrangian is the same as the solution to the KKT Lagrangian. Why? For any solution to $\mathcal{L}_A$, the feasibility constraint of the $c_i(x) = 0$ implies that

$$\nabla\big(c_i^2(x)\big) = 2\nabla c_i(x) \cdot c_i(x) = 0 \Rightarrow \nabla_x \mathcal{L}_A = \nabla_x \mathcal{L}_{KKT}$$

**Note:** In genereal, the KKT Lagrangian has no "nice" properties such as convexity (unless otherwise specified — see assignment 6?)

The minimizer $x_k$ of $\mathcal{L}_A(x, \lambda^{k-1}, \mu_k)$ satisfies

$$\begin{aligned} 0 = \nabla_x \mathcal{L}_A(x, \lambda^{k-1}, \mu_k) &= \nabla f(x_k) - \sum_{i \in \mathcal{E}} \lambda_i^{k-1} \nabla c_i(x_k) + \mu_k \sum_{i \in \mathcal{E}} \nabla c_i(x) \cdot c_i(x) \\ &= \nabla f(x_k) - \sum_{i \in \mathcal{E}} \left(\lambda_i^{k-1} - \mu_k c_i(x_k)\right) \cdot \nabla c_i(x_k) \\ &\Rightarrow \lambda_i^\star \simeq \lambda_i^{k-1} - \mu_k c_i(x_k) \\ &\Rightarrow c_i(x_k) \simeq \frac{1}{\mu_k}(\lambda_i^{k-1} - \lambda_i^\star) \end{aligned}$$

Note: in the algorithm, we specifiy $\mu_k > 0$, not $\geq 0$

**Method of Multipliers**

Given $\lambda^0, \mu_0 > 0$
For $k = 1, 2, \ldots$
$\arg\min \mathcal{L}_A(x, \lambda^{k-1}, \mu_k) \to x_k$
Check stopping criterion for $(x_k, \lambda^{k-1})$
If satisfied, STOP
Otherwise
$\lambda_i^k = \lambda_i^{k-1} - \mu_k c_i(x_k)$ for all $i \in \mathcal{E}$
Choose $\mu_{k+1} \geq \mu_k$

Suppose $(x^\star, \lambda^\star)$ is a KKT point for the problem mentioned at the start of all of this. This implies that LICQ holds at $x^\star$ and $x^\star$ satisfies the second-order sufficient conditions. Then, there exists a $\bar{m}u$ such that $x^\star$ is a strict local minimizer of $\mathcal{L}_A(x, \lambda^\star, \mu)$, for all $\mu \geq \bar{\mu}$. This means we must show that $\nabla_x \mathcal{L}_A(x^\star, \lambda^\star, \mu) = 0$ and $\nabla_{xx} \mathcal{L}_A(x^\star, \lambda^\star, \mu) \succ 0$ for $\mu$ large enough

*Proof.*

$$\nabla_x \mathcal{L}_A(x^\star, \lambda^\star, \mu) = \underbrace{\nabla f(x^\star) - \sum_{i \in \mathcal{E}} \lambda_i^\star \nabla c_i(x^\star)}_{=0 \text{because KKT point}} + \mu \sum_{i \in \mathcal{E}} \underbrace{c_i(x^\star)}_{=0} \nabla c_i(x^\star) = 0$$

□

*Proof.* Suppose for contradiction, we have an increasing sequence of $\mu_k$, with $\mu_1 < \mu_2 < \cdots$ with $\mu_k \to \infty$ such that $\nabla_{xx} mathcal L_A(x^\star, \lambda^\star, \mu)$ is not positive definite. Then we can find $w_k$ (not! working set) such that $\|w_k\| = 1$ and

$$w_k^T \nabla_{xx} \mathcal{L}_A(x^\star, \lambda^\star, \mu_k) w_k \leq 0$$

Thus, by linearity we have that

$$w_k^T \nabla_{xx} \mathcal{L}(x^\star, \lambda^\star) w_k \leq -\mu_k \|A w_k\|_2^2 \leq 0$$

But we assumes by the second order sufficient condition that $\nabla_{xx} \mathcal{L}(x^\star, \lambda^\star)$ is positive definite, which is a contradiction.

Alternatively, we have an accumulation point $w$. Then if $w \in \mathcal{F}(x^\star)$ then $w^T \nabla_{xx} \mathcal{L}(x^\star, \lambda^\star) w < 0$ which is a contradiction. For $w \notin \mathcal{F}(x^\star)$ then we have that $\|Aw\| > 0$, also leads to a contradiction □

**Aside:** Can also show that there exists $M > 0$ such that

$$\|x_k - x^\star\|_2 \leq \frac{M}{\mu_k}\|\lambda^k - \lambda^\star\|_2$$

and

$$\|\lambda^{k+1} - \lambda^\star\|_2 \leq \frac{M}{\mu_k}\|\lambda^k - \lambda^\star\|_2$$

which can be found in "D.P. Bertsekas, Nonlinear Programming"