Optimisation: Questions from past exam papers

September 10, 2016

Part I
Disclaimer

As it appears that exam questions only span through a limited number of possibilities, I tried to cover them all thoroughly in the hope of gaining a decent mark. Now, as I also wanted to improve my non-existing ability to work with LyX, I decided to create this document. However, even though my endevour has always been that of being as accurate as possible, I must warn you of two different issues. First of all, this notes are based on the slides prepared by Prof. Amaldi (which can be found here: http://home.deib.polimi.it/amaldi/teaching/) and on a book named *Numerical Optimization*, by J. Nocedal and S. Wright, Springer Edition. I did not attend the vast majorities of the classes, mainly because they were in the morning or after lunch, so I cannot guarantee that the way things are covered here matches the one followed by Prof. Amaldi. I followed both the book and the slides according to my taste. The second issue, perhaps even more relevant, ~~is that there might be mistakes scattered all over~~ the place despite my best efforts. This notes have not been tested yet, as I still have not taken the exam.

Having said that, I hope they will somehow prove to be useful.

Andrea Mascaretti

Part II
Gradient Method

Discuss the method of gradient for unconstrained optimization problems, illustrating it with an example. Discuss the convergence properties together with the advantages and disadvantages of this method. Briefly explain alternative approaches that were developed to avoid them.

The gradient method, or steepest descent, is a 1-D exact search method for unconstrained nonlinear minimisation (and, of course, maximisation) problems. In other words, we aim at solving the following:

$$\min_{\underline{x}\epsilon\mathbb{R}^n} f(\underline{x}),\ f:\mathbb{R}^n\to\mathbb{R}\ and\ f\,\epsilon\, C^1(\mathbb{R}^n)$$

The goal is to generate a sequence

$$\{x_k\}_{k\geq 0}\subset\mathbb{R}^n\ s.t.\ \underline{x}_k\to\underline{x}^*\ and\ \underline{x}^*\epsilon\,\Omega$$

where

$$\Omega=\{\underline{x}\epsilon\mathbb{R}^n\ s.t.\ \nabla f(\underline{x})=0\}$$

i.e. a sequence converging in Ω, the set of stationary points.

It is clear that, as we are dealing with a sequence, three different problems arise. The first one is connected with the *initialisation* of the sequence: we must pick some $\underline{x}_0$to begin with. The second one is about *iterations*: how do we connect $\underline{x}_k$to $\underline{x}_{k+1}$? And the third one is about the *termination* of the algorithm: despite hypothising convergence in the first place, we cannot wait forever for the sequence to converge. We shall require some criterion to decide when enough is enough and our approximate solution will be fit for the job.

Moving from the third problem back to the first, we now propose three commonly-used termination criteria. The first one takes into account the distance between points in the sequence: when points begin to get too close to each other, it is sensible to say that we are making computations without actually moving much and we can therefore consider ourselves satisfied with our results. This, in mathematical terms, translates as follows. We set $\bar{k}$ to be the index at wich our sequence is truncated. We fix $\varepsilon>0$, we will have that

$$\bar{k}\geq 0\ is\ such\ that\ \|\underline{x}_{\bar{k}+1}-\underline{x}_{\bar{k}}\|\leq\varepsilon,\ \underline{x}^*\epsilon\Omega$$

In a similar fashion, we can control the variation in the output function: when it gets too narrow from one iteration to the other, it is probably high time we stopped making calculations. In other words, again fix $\varepsilon>0$,

$$\bar{k}\geq 0\ is\ such\ that\ |f(\underline{x}_{\bar{k}})-f(\underline{x}_{\bar{k}+1})|\leq\varepsilon$$

and we are done.

The last one takes into account the gradient of the function: when it is close to zero, we are close to finding a stationary point (keep in mind we are working with continuous functions in an unbounded set). Fix $\varepsilon>0$,

$$\bar{k}\geq 0\ is\ such\ that\ \|\nabla f(\underline{x}_{\bar{k}})\|\leq\varepsilon$$

We now consider the second problem: how do we build our sequence? The big thing about 1-D exact search methods is that one tries to find a "best"

direction and then minimises the function along that direction, finding the best pace. In other words, given $\underline{x}_k$,

$$\underline{x}_{k+1} := \underline{x}_k + \alpha_k \underline{d}_k$$

We begin our search looking for the "best" direction. In this case, as we have asked for our function to be differentiable with continuous derivatives, we will try to exploit the information this gives us. To do so, we first notice that the function we are trying to minimise can be approximated with a linear function

$$f(\underline{x}_k + \underline{d}) \approx f(\underline{x}_k) + \nabla^t f(\underline{x}_k)\underline{d} = L_k(\underline{d})$$

We wish to find the directions along which our function is minimised. We therefore solve the following minimisation problem:

$$\min L_k(\underline{d}) = \min f(\underline{x}_k) + \nabla^t f(\underline{x}_k)\underline{d} = \min \nabla^t f(\underline{x}_k)\underline{d}$$

adding the constraint

$$\|\underline{d}\| = \|\nabla f(\underline{x}_k)\|$$

so that we can be sure that the minimising direction is not arbitrarily big. We notice what follows:

$$\nabla^t f(\underline{x}_k)\underline{d} = \|\nabla f(\underline{x}_k)\|\|\underline{d}\| \cos\theta$$

as we have a scalar product between to n-dimensional vectors. We only need to consider the angle to obtain that

$$\nabla^t f(\underline{x}_k)\underline{d} \geq -\|\nabla f(\underline{x}_k)\|\|\underline{d}\|$$

and, taking into account the constraint,

$$\nabla^t f(\underline{x}_k)\underline{d} \geq -\|\nabla f(\underline{x}_k)\|\|\underline{d}\| = -\|\nabla f(\underline{x}_k)\|^2 = -\nabla^t f(\underline{x}_k)\nabla f(\underline{x}_k)$$

so that we have that the direction we wish to find is given by $d = -\nabla f(\underline{x}_k)$.

We will know take a look at the second part of the issue: how do we find the exact pace? We will have to solve the minimisation problem:

$$\min_{\alpha \geq 0} \Phi(\alpha) = \min_{\alpha \geq 0} f(\underline{x}_k + \alpha_k \underline{d}_k) = \min_{\alpha \geq 0} f(\underline{x}_k - \alpha_k \nabla f(\underline{x}_k))$$

However, before doing that, we notice that at each iteration the gradient method takes a direction that is orthogonal to the previous one. We can easily prove this.

Let α^* be such that $\alpha^* = \arg\min_{\alpha \geq 0} \Phi(\alpha)$, then $\Phi^{'}(\alpha^*) = 0$. We have

$$\Phi'(\alpha) = -\nabla^t f(\underline{x}_k - \alpha_k \nabla f(\underline{x}_k))\nabla f(\underline{x}_k) = -\nabla^t f(\underline{x}_{k+1})\nabla f(\underline{x}_k) = -\underline{d}_{k+1}^t \underline{d}_k^t = 0$$

and we see that the directions are orthogonal.

As for the exact search, please note that whilst it is more than desirable to have an exact α (it guarantess the method is descent, that is non-growing, and this property implies that our solution cannot worsen if we add iterations; it strengthen convergence properties), on the other hand, to spend too much time looking for the perfect number can turn out to be rather computationally expensive. We shall limit ourselves to the study of strictly convex quadratic functions, so that we can work on a simple case that is also somewhat representative of what happens when we are close to the optimum.

We have

$$f(x) = \frac{1}{2}\underline{x}^t Q \underline{x} - \underline{b}^t \underline{x}$$

with Q positive definite.

We know that the minimum is attained when

$$\nabla f(\underline{x}) = Q\underline{x} - \underline{b} = 0 \Longrightarrow \underline{x} = Q^{-1}\underline{b}$$

We know wish to find the optimum α.

$$\Phi(\alpha) = \frac{1}{2}\left(\underline{x}_k - \alpha_k \nabla f(\underline{x}_k)\right)^t Q \left(\underline{x}_k - \alpha_k \nabla f(\underline{x}_k)\right) - \underline{b}^t \left(\underline{x}_k - \alpha_k \nabla f(\underline{x}_k)\right)$$

and

$$\Phi^{'}(\alpha) = -\nabla^t f(\underline{x}_k) Q \left(\underline{x}_k - \alpha_k \nabla f(\underline{x}_k)\right) + \nabla^t f(\underline{x}_k)\underline{b} =$$
$$= -\nabla^t f(\underline{x}_k) Q \underline{x}_k + \alpha_k \nabla^t f(\underline{x}_k) Q \nabla f(\underline{x}_k) + \nabla^t f(\underline{x}_k)\underline{b}$$

We observe that

$$\nabla f(\underline{x}_k) = Q\underline{x}_k - \underline{b} \Longrightarrow \underline{b} = Q\underline{x}_k - \nabla f(\underline{x}_k)$$

so we have

$$\Phi^{'}(\alpha) = -\nabla^t f(\underline{x}_k) Q \underline{x}_k + \alpha_k \nabla^t f(\underline{x}_k) Q \nabla f(\underline{x}_k) + \nabla^t f(\underline{x}_k)\underline{b} =$$
$$= -\nabla^t f(\underline{x}_k) Q \underline{x}_k + \alpha_k \nabla^t f(\underline{x}_k) Q \nabla f(\underline{x}_k) + \nabla^t f(\underline{x}_k)\left(Q\underline{x}_k - \nabla f(\underline{x}_k)\right) =$$
$$= \alpha_k \nabla^t f(\underline{x}_k) Q \nabla f(\underline{x}_k) - \nabla^t f(\underline{x}_k)\nabla f(\underline{x}_k) = 0$$

to finally find that

$$\alpha_k = \frac{\nabla^t f(\underline{x}_k)\nabla f(\underline{x}_k)}{\nabla^t f(\underline{x}_k) Q \nabla f(\underline{x}_k)}$$

which, if we define

$$\|\underline{y}\|_Q = \sqrt{\underline{y}^t Q \underline{y}}$$

becomes

$$\alpha_k = \frac{\nabla^t f(\underline{x}_k)\nabla f(\underline{x}_k)}{\nabla^t f(\underline{x}_k) Q \nabla f(\underline{x}_k)} = \frac{\|\nabla f(\underline{x}_k)\|^2}{\|\nabla f(\underline{x}_k)\|_Q^2}$$

We now wish to understand more about the convergence properties of this method. First of all, it is usual to consider the convergence of $f(\underline{x}_k) - f(x^*)$ rather than that of $\|\underline{x}_k - \underline{x}^*\|$, when $k \to +\text{inf}$. Moreover, if $f \in C^1(\mathbb{R}^n) \cap C^2(\mathbb{R}^n) = C^2(\mathbb{R}^n)$ and $\nabla^2 f(\underline{x}^*)$ is positive definite, $\|\underline{x}_k - \underline{x}^*\| \to 0$ (super)linearly if and only if $|f(\underline{x}_k) - f(\underline{x}^*)| \to 0$, if we are in a neighbourhood $\mathcal{N}(\underline{x}^*)$ of our stationary point. This is easily demonstrated by considering the Taylor expansion of the function. (Please remember that $\nabla f(\underline{x}^*) = \underline{0}$)

$$f(\underline{x}) \approx f(\underline{x}^*) + \nabla^t f(\underline{x}^*)(\underline{x} - \underline{x}^*) + \frac{1}{2}(\underline{x} - \underline{x}^*)^t \nabla^2 f(\underline{x}^*)(\underline{x} - \underline{x}^*) =$$

$$= f(\underline{x}^*) + \frac{1}{2}(\underline{x} - \underline{x}^*)^t \nabla^2 f(\underline{x}^*)(\underline{x} - \underline{x}^*)$$

so

$$\exists\, m,\ M \in \mathbb{R}^+ \ s.t.\ m\|\underline{x} - \underline{x}^*\| \leq |f(x) - f(x^*)| \leq M\|\underline{x} - \underline{x}^*\|$$

This is even more evident if we consider quadratic strictly convex functions. Knowing that $Q\underline{x}^* = \underline{b}$, we have that

$$\frac{1}{2}\|\underline{x} - \underline{x}^*\|_Q^2 = \frac{1}{2}(\underline{x} - \underline{x}^*)^t Q(\underline{x} - \underline{x}^*) = \frac{1}{2}\left[\underline{x}^t Q\underline{x} - (\underline{x}^*)^t Q\underline{x} - \underline{x}^t Q\underline{x}^* + (\underline{x}^*)^t Q\underline{x}^*\right] =$$

$$= \frac{1}{2}\underline{x}^t Q\underline{x} - (\underline{x}^*)^t Q\underline{x} + \frac{1}{2}(\underline{x}^*)^t Q\underline{x}^* = \frac{1}{2}\underline{x}^t Q\underline{x} - \underline{b}^t\underline{x} - \frac{1}{2}(\underline{x}^*)^t Q\underline{x}^* + (\underline{x}^*)^t Q\underline{x}^* =$$

$$= \frac{1}{2}\underline{x}^t Q\underline{x} - \underline{b}^t\underline{x} - \frac{1}{2}(\underline{x}^*)^t Q\underline{x}^* + \underline{b}^t\underline{x}^* = f(\underline{x}) - f(\underline{x}^*)$$

We now state an important theorem about convergence.

Let

$$f(\underline{x}) = \frac{1}{2}\underline{x}^t Q\underline{x} - \underline{b}^t\underline{x}$$

be a quadratic strictly convex function (i.e. Q is positive definite). Then, if the gradient method with exact 1-D search is applied to find the stationary point we have that

$$\forall \underline{x}_0 \in \mathbb{R}^n,\ \lim_{k \to +\text{inf}} \underline{x}_k = \underline{x}^*$$

and

$$\|\underline{x}_{k+1} - \underline{x}^*\|_Q^2 \leq \left(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}\right)^2 \|\underline{x}_k - \underline{x}^*\|_Q^2$$

where $\Lambda = \{\lambda_1, \ldots, \lambda_n\}$is the set of eigenvalues of Q. The proof of this theorem is based on Zoutendijk's theorem and on the Kantorovich inequality.

We can now formulate some considerations on what we have here. Let $\kappa = \frac{\lambda_n}{\lambda_1}$ be the *condition number* of the matrix. We have that

$$\|\underline{x}_{k+1} - \underline{x}^*\|_Q^2 \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^2 \|\underline{x}_k - \underline{x}^*\|_Q^2$$

and we can see that if the eigenvectors are very different, $\kappa \uparrow \implies \left(\frac{\kappa-1}{\kappa+1}\right) \approx 1$, wheareas if they are close $\kappa \approx 1$, $\left(\frac{\kappa-1}{\kappa+1}\right) \ll 1$ and we have better convergence.

Now, if we consider a generic function $f \in C^2(\mathbb{R}^n)$ such that we have a stationary point $\underline{x}^*$ at which $\nabla^2 f(\underline{x}^*)$ is positive definite, we have the following

$$f(\underline{x}^*) - f(\underline{x}_{k+1}) \leq \left(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}\right)^2 [f(\underline{x}^*) - f(\underline{x}_k)]$$

where $\Lambda = \{\lambda_1, \ldots, \lambda_n\}$ here is the set of eigenvalues of $\nabla^2 f(\underline{x}^*)$ and we can easily see how convergence properties of this method are heavily affected by the structure of the Hessian matrix at the stationary point, at least when we are approaching it. We will know try to see if a different approach could help us deal with this issue: we will briefly review the conjugate gradient method. The main idea beyond this result is simple. Let us consider a strictly convex quadratic function with a diagonal matrix. If we take a look at its level lines, we will notice we have concentric ellipses. The algorithm will proceed along the canonical directions $\hat{e}_i$ to reach the minimum (we have that $\mathbb{R}^n = span(\{\hat{e}_1, \ldots, \hat{e}_n\})$), that will be attained in n steps. In a two-dimensional settings, at first one coordinate of the ellipse will be found along the x-axis (or the y-axis) and then the second will be reached taking into account the y-axis (or the x-axis). We want to transform a generic symmetric positive definite matrix into something that is diagonal and go along some analogous directions. To achieve that, we first have to introduce a number of preliminary concepts.

Let A be a symmetric positive definite matrix of dimensions $n \times n$ and let $P = \{\underline{p}_1, \ldots, \underline{p}_n\}$ be a set of vectors. Then we say that the vectors are *conjugate* with respect to A if the following holds:

$$\underline{p}_i^t A \underline{p}_j = 0,\ i \neq j$$

and it is easy to see that if the vectors are conjugate, they are also independent and therefore we have that $\mathbb{R}^n = span(P)$.

We want to build a method in which we have

$$\underline{x}_{k+1} := \underline{x}_k + \alpha_k \underline{p}_k$$

and we have already prove that, if we set $\underline{r}_k = Q\underline{x}_k - \underline{b} = \nabla f(\underline{x}_k)$, the optimum α_k is obtained when we have

$$\alpha_k = \frac{\underline{r}_k^t \underline{p}_k}{\underline{p}_k^t Q \underline{p}_k}$$

We want to show we can have convergence in n iterations.

As $\mathbb{R}^n = span(P)$, we can write

$$\underline{x}^* - \underline{x}_0 = \sum_{i=0}^{n-1} \sigma_i \underline{p}_i$$

and if we pre-multiply by $\underline{p}_k^t A$, we have

$$\underline{p}_k^t A\,(\underline{x}^* - \underline{x}_0) = \underline{p}_k^t A \left(\sum_{i=0}^{n-1} \sigma_i \underline{p}_i\right) = \sigma_k \underline{p}_k^t A \underline{p}_k$$

so that

$$\sigma_k = \frac{\underline{p}_k^t A\,(\underline{x}^* - \underline{x}_0)}{\underline{p}_k^t A \underline{p}_k}$$

We aim at demonstrating that σ_k and α_k coincides for each k. To do so, we consider $\underline{x}_k$ obtained at the k-th iteration of the algorithm. We have

$$\underline{x}_k = \underline{x}_0 + \alpha_1 \underline{p}_1 + \ldots + \alpha_{k-1} \underline{p}_{k-1}$$

and therefore

$$\underline{p}_k^t A\,(\underline{x}_k - \underline{x}_0) = \underline{p}_k^t A \left(\alpha_0 \underline{p}_0 + \ldots + \alpha_{k-1} \underline{p}_{k-1}\right) = 0$$

as they are all conjugate directions. We can now observe that

$$\underline{p}_k^t A\,(\underline{x}^* - \underline{x}_0) = \underline{p}_k^t A\,(\underline{x}^* - \underline{x}_k + \underline{x}_k - \underline{x}_0) = \underline{p}_k^t A\,(\underline{x}^* - \underline{x}_k) + \underline{p}_k^t A\,(\underline{x}_k - \underline{x}_0) =$$

$$= \underline{p}_k^t A\,(\underline{x}^* - \underline{x}_k) = \underline{p}_k^t\,(\underline{b} - A\underline{x}_k) = \underline{p}_k^t \underline{r}_k$$

and it is now evident that

$$\sigma_k = \frac{\underline{p}_k^t A\,(\underline{x}^* - \underline{x}_0)}{\underline{p}_k^t A \underline{p}_k} = \frac{\underline{p}_k^t \underline{r}_k}{\underline{p}_k^t A \underline{p}_k} = \alpha_k$$

and therefore the algorithm converges in n steps.

To fix the idea, we say that we define a new set of variables

$$\hat{\underline{x}} = S^{-1}\underline{x}$$

where $S = [\underline{p}_0 \ \ldots \ \underline{p}_n]$ and we have

$$\hat{f}(\hat{\underline{x}}) = \frac{1}{2}\hat{\underline{x}}^t (S^t Q S)\hat{\underline{x}} - (S^t \underline{b})^t \hat{\underline{x}}$$

and by the conjugacy property the matrix $S^t Q S$ is diagonal.

We now state one final theorem about subspaces. We have noticed that in the case of diagonal matrices, not only does the gradient method manage to move from solving some n-dimensional optimisation problem to n singled dimensional problems, but it also build up the correct coordinates step by step. We want to see whether this also happens whenever we use conjugate directions for our matrix Q.

Let $\underline{x}_0 \in \mathbb{R}^n$ be a generic starting point and suppose that we have a sequence that is being generated with the conjugate gradient algorithm. Then we have that

$$for\, i = 0, \ldots, k-1, \quad \underline{r}_k^t \underline{p}_i = 0$$

and $\underline{x}_k$ is the minimiser of $f(\underline{x})$ over the set

$$\left\{\underline{x} \,|\, \underline{x} = \underline{x}_0 + span(\{\underline{p}_0, \ldots, \underline{p}_{k-1}\})\right\}$$

One example of conjugate directions, that are also orthogonal, is the set of eigenvectors of Q.

The conjugate gradient method can compute the new direction $\underline{p}_k$ without knowing all the previous ones. Each direction is chosen moving from a linear combination of the negative residuals $-\underline{r}_k$ and the previous one $\underline{p}_{k-1}$. In other words, we have

$$\underline{p}_k = -\underline{r}_k + \beta_k \underline{p}_{k-1}$$

so, we pre-multiply by $\underline{p}_{k-1}^t A$ to have

$$\underline{p}_{k-1}^t A \underline{p}_k = 0 \implies \underline{p}_{k-1}^t A \left(-\underline{r}_k + \beta_k \underline{p}_{k-1}\right) = 0$$

and

$$\beta_k = \frac{\underline{p}_{k-1}^t A \underline{r}_k}{\underline{p}_{k-1}^t A \underline{p}_{k-1}}$$

Part III
Newton Method

Discuss the Newton method for unconstrained optimization problems, illustrating it with an example. Discuss the convergence properties together with the advantages and disadvantages of this method. Briefly explain alternative approaches that were developed to avoid them.

We are again considering a 1-D search method. We shall therefore skip the same preliminar observations already formulated for the gradient method.

Let $f \in C^2(\mathbb{R}^n)$, we can write the Taylor expansion of it as follows:

$$f(\underline{x} - \underline{x}_k) \approx f(\underline{x}_k) + \nabla^t f(\underline{x}_k)(\underline{x} - \underline{x}_k) + \frac{1}{2}(\underline{x} - \underline{x}_k)^t H(\underline{x}_k)(\underline{x} - \underline{x}_k) := q_k(\underline{x})$$

where $H(\underline{x}_k) = \nabla^2 f(\underline{x}_k)$. We want to get to a stationary point.

$$\nabla_{\underline{x}} q_k(\underline{x}) = \nabla f(\underline{x}_k) + H(\underline{x}_k)(\underline{x} - \underline{x}_k) = \underline{0}$$

so that we get

$$\underline{x}_{k+1} := \underline{x}_k - [H(\underline{x}_k)]^{-1} \nabla f(\underline{x}_k)$$

whenever the Hessian is positive definite. It is, in fact, this hypotheis together with that on differential properties of the function that implies that the inverse of the Hessian matrix is also positive definite and therefore that the direction is well-defined in a neighbourhood $\mathcal{N}(\underline{x}_k)$. The pure Newton method impose $\alpha_k = 1$ for every iteration. Note that the method is invariant with respect to affine (and non singular) coordinate changes. Moreover, it is important to point out that we do not have global convergence: for the method to converge we ask that $\underline{x}_0 \epsilon \mathcal{N}(\underline{x}^*)$. With a strictly convex quadratic function, we have convergence in just one iteration. We now state a theorem on convergence.

Let $f \epsilon C^2(\mathbb{R}^n)$, $\underline{x}_0 \epsilon \mathcal{N}(\underline{x}^*)$, with $\underline{x}^* \epsilon \Omega$ (i.e. $\underline{x}^* s.t. \nabla f(\underline{x}^*) = \underline{0}$) and $H(\underline{x}^*)$ positive definite. What is more, we hypothise that

$$\exists L > 0 s.t. \|H(\underline{x}) - H(\underline{y})\| \leq L\|\underline{x} - \underline{y}\|, \quad \forall \underline{x}, \underline{y} \epsilon \mathcal{N}(\underline{x}^*)$$

then we have that, for $k \to +\inf$,

1. $\{\underline{x}_k\} \to \underline{x}^*$ with a quadratic convergence order
2. $\{\|\nabla f(\underline{x}_k)\|\} \to 0$ quadratically

Despite being rather efficient when the starting point is close to the stationary one, this methods does come with some disadvantages:

- if the Hessian matrix is not well defined at the k-th iteration, we might not have a direction
- if the Hessian matrix is not definite positive at the k-th iteration, we might not be able to univocally find the successive direction
- we are not certain the method is descent as α might violate Wolfe (or alternative conditions)
- the computational cost is high: $O(n^3)$
- the algorithm might well converge to local maxima

Some variations have been proposed to eliminate various problems. As far as the α is concerned, it is possible to shift to inexact 1-D search. Moreover, it is possible to combine the gradient method with the Newton method so that we get that

$$\underline{d}_k = -D_k \nabla f(\underline{x}_k)$$

and we set

$$D_k = [\varepsilon I + H(\underline{x}_k)]^{-1}$$

with $\varepsilon s.t. eig_i(D_k) > 0$. A third modification introduces a new way of considering the approximation and it is called *trust region method*. The approximation

$q(\underline{x}_k)$ is minimised over a region determined by setting some $r > 0$ and considering, for example, the ball $B(r, \underline{x}_k)$. However, the most relevant modifications are connected with the Hessian matrix: many methods were developed in order to create some iterative procedures that could approximate it.

The idea is to consider a quadratic function of the form

$$m_k(\underline{p}) = f_k + \nabla^t f_k \underline{p} + \frac{1}{2} \underline{p}^t B_k \underline{p}$$

where $f(\underline{x}_k) = f_k$ and $\nabla f(\underline{x}_k) = \nabla f_k$. We want to find understand how to set B_k to obtain a good approximation of the Hessian matrix at that point. First of all, we note that at $\underline{p} = \underline{0}$ the function and the gradient match with the original ones. We also know that the function is minimsed at $\underline{p}_k = -B_k^{-1} \nabla f_k$. We shall use this direction to compute each new iteration, imposing

$$\underline{x}_{k+1} := \underline{x}_k + \alpha_k \underline{p}_k$$

with α_k such that Wolfe condition are not violated. Now, instead of calculating B_k at each iteration, we want to exploit the information contained in the first derivatives to obtain something. In this case, we will ask that the gradient of the function m_{k+1} matches with the true ones at $\underline{x}_k$ and at $\underline{x}_{k+1}$. The latter case is evident: we just impose $\underline{p} = \underline{0}$ as already argued before. To have matching at $\underline{x}_k$, we will have to make some more calculations. So, we have that

$$m_{k+1}(\underline{p}) = f_{k+1} + \nabla^t f_{k+1} \underline{p} + \frac{1}{2} \underline{p}^t B_{k+1} \underline{p}$$

and we want to go back to $\underline{x}_k$, so we have

$$\nabla m_{k+1}(-\alpha_k \underline{p}_k) = \nabla f_{k+1} - \alpha_k B_{k+1} \underline{p}_k = \nabla f_k$$

so that, rearranging, we have

$$\nabla f_{k+1} - \nabla f_k = B_{k+1} \left(\alpha_k \underline{p}_k\right)$$

and if we set $\underline{\gamma}_k = \nabla f_{k+1} - \nabla f_k$ and $\underline{\delta}_k = \underline{x}_{k+1} - \underline{x}_k = \alpha_k \underline{p}_k$, we have the *secant condition*

$$\underline{\gamma}_k = B_{k+1} \underline{\delta}_k$$

There is also one other condition we have to impose: the *curvature condition*. As we must have that B_{k+1} is positive definite, we want

$$\underline{\delta}_k^t \underline{\gamma}_k > 0$$

However, those two conditions are not enough to determine the matrix as we have $\frac{n(n+1)}{2}$ degrees of freedom and n linear equations with the secant condition. So the basic idea is that we try to fix B_{k+1} by requiring it to be as close as possible to the matrix B_k. In other words, we have to solve

$$\min_B \|B - B_k\|$$

$$s.t.\ B = B^t \ and\ \underline{\gamma}_k = B\underline{\delta}_k$$

Basically, various norms lead to different quasi-Newton methods. Considering the Frobenius norm with the average Hessian we are lead to the DFP method.

In a similar fashion, if we set

$$H_{k+1}\underline{\gamma}_k = \underline{\delta}_k$$

and consider

$$\min_H \|H - H_k\|$$

$$s.t.\ H = H^t \ and\ H\underline{\gamma}_k = \underline{\delta}_k$$

With this method, each iteration can be performed at a cost of $O(n^2)$ instead of $O(n^3)$. Moreover, the algorithm is robust and its rate of convergence is more than quadratic. It is a bit slower than the Newton method and requires more storage though.

Part IV
Quadratic programming

Describe the methods for solving constrained nonlinear optimization problems with a quadratic objective function subject to linear constraints, that is, the so-called quadratic programming problems. First consider the case with only linear equality constraints and then the case with also linear inequality constraints. Discuss the properties of the methods. Describe an application that can be formulated as a quadratic program.

1 Review: the Karush-Kuhn-Tucker conditions

We consider the following problem

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

$$subject\ to$$

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

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

or, in other words, setting $\Omega = \{\underline{x} \in \mathbb{R}^n\ s.t.\ c_i(\underline{x}) = 0,\ \forall i \in \mathcal{E}\ and\ c_i(\underline{x}) \geq 0,\ \forall i \in \mathcal{I}\}$, we have

$$\min_{\underline{x}\in\Omega} f(\underline{x})$$

where $f \in C^1(\Omega)$ and $c_i \in C^1(\Omega)$, $\forall i \in \mathcal{E} \cup \mathcal{I}$.

We define the Lagrangian function

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

Suppose that $\underline{x}^*$ is a local solution of the problem and that the CQ (constraint qualification) holds at $\underline{x}^*$. Then there is a Lagrange multiplier vector $\underline{\lambda}^*$, with components λ_i, $\forall i \in \mathcal{E} \cup \mathcal{I}$, such that the following conditions are satisfied at $(\underline{x}^*, \underline{\lambda}^*)$

$$\nabla_{\underline{x}}\mathcal{L}(\underline{x}^*, \underline{\lambda}^*) = 0$$
$$c_i(\underline{x}^*) = 0, \quad \forall i \in \mathcal{E}$$
$$c_i(\underline{x}^*) \geq 0, \quad \forall i \in \mathcal{I}$$
$$\lambda_i^* \geq 0, \quad \forall i \in \mathcal{I}$$
$$\lambda_i^* c_i(\underline{x}^*) = 0, \quad \forall i \in \mathcal{E} \cup \mathcal{I}$$

2 Quadratic Programming

The general quadratic program (QP) can be stated as follows

$$\min_{\underline{x}} q(x) = \frac{1}{2}\underline{x}^t G\underline{x} + \underline{x}^t c$$

$$s.t.\ \underline{a}_i^t\underline{x} = \underline{b}_i, \quad \forall i \in \mathcal{E}$$

$$\underline{a}_i^t\underline{x} \geq \underline{b}_i, \quad \forall i \in \mathcal{I}$$

where G is a symmetric (without loss of generality) $n \times n$ matrix, $\mathcal{E}$ and $\mathcal{I}$ are finite sets of indices, and $\underline{c}$, $\underline{x}$, $\{\underline{a}_i\}$, $\forall i \in \mathcal{E} \cup \mathcal{I}$ are vectors in $\mathbb{R}^n$. Quadratic programs can always be solved (or shown to be unsolvable) in a finite number of steps. However, the difficulty can either increase or decrease depending on the nature of the matrix G. Whenever the Hessian matrix G is positive semidefinite, we have a convex QP and a relatively easy case. The situation can become a little rougher when we are facing non-convex QPs, as they can have various stationary points and local minima. We will, at first, consider only equality-constrained quadratic problems and move to the general case only in a second moment. We will therefore consider a problem of the form

$$\min_{\underline{x}} q(x) = \frac{1}{2}\underline{x}^t G\underline{x} + \underline{x}^t\underline{c} \ \ subject\ to\ \ A\underline{x} = \underline{b}$$

where A is the $m \times n$ Jacobian of constraints (with $m < n$ and $rank(A) = m$). We now want to derive the first-order optimality conditions for the problem. First of all, we can see that we can consider

$$A = \begin{bmatrix} \underline{a}_1^t \\ \vdots \\ \underline{a}_m^t \end{bmatrix}$$

and we have *for* $i = 1, \dots, m, \quad \underline{a}_i^t \underline{x} = b_i$. We now make some calculations, knowing that

$$\mathcal{L}(\underline{x}, \underline{\lambda}) := q(\underline{x}) - \sum_{i=1}^{m} \lambda_i \left(\underline{a}_i^t \underline{x} - b_i\right)$$

we have

$$\nabla_{\underline{x}} \mathcal{L}(\underline{x}, \underline{\lambda}) = \nabla_{\underline{x}} \left(q(\underline{x}) - \sum_{i=1}^{m} \lambda_i \left(\underline{a}_i^t \underline{x} - b_i\right) \right) = \nabla_{\underline{x}} q(\underline{x}) - \nabla_{\underline{x}} \left(\sum_{i=1}^{m} \lambda_i \left(\underline{a}_i^t \underline{x} - b_i\right) \right) =$$

$$= \nabla_{\underline{x}} q(\underline{x}) - \sum_{i=1}^{m} \lambda_i \underline{a}_i = G\underline{x} + \underline{c} - \sum_{i=1}^{m} \lambda_i \underline{a}_i = 0$$

and, considering the costraints, we can rewrite the KKT conditions as follows

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

as we have

$$\sum_{i=1}^{m} \lambda_i \underline{a}_i = \begin{bmatrix} \underline{a}_1 \\ \vdots \\ \underline{a}_m \end{bmatrix} \begin{bmatrix} \lambda_1 & \dots & \lambda_m \end{bmatrix} = A^t \underline{\lambda}$$

However, the system can be rewritten in a form that is more suitable for calculations. That is, we set $\underline{x}^* = \underline{x} + \underline{p}$ to get that

$$\begin{cases} G\left(\underline{x} + \underline{p}\right) - A^t \underline{\lambda}^* = -\underline{c} \\ A\left(\underline{x} + \underline{p}\right) = \underline{b} \end{cases}$$

$$\begin{cases} G(-\underline{p}) + A^t \underline{\lambda}^* = \underline{c} + G\underline{x} \\ A(-\underline{p}) = A\underline{x} + \underline{b} \end{cases}$$

which translates into the following system

$$\begin{bmatrix} G & A^t \\ A & 0 \end{bmatrix} \begin{bmatrix} -\underline{p} \\ \underline{\lambda}^* \end{bmatrix} = \begin{bmatrix} \underline{g} \\ \underline{h} \end{bmatrix}$$

where

$$\begin{bmatrix} \underline{g} \\ \underline{h} \end{bmatrix} = \begin{bmatrix} \underline{c} + G\underline{x} \\ A\underline{x} + \underline{b} \end{bmatrix}$$

and the matrix $\begin{bmatrix} G & A^t \\ A & 0 \end{bmatrix}$ is called the Karush-Kuhn-Tucker matrix. We know want to provide some condition for the KKT to be nonsingular. Let Z denote the $n \times (n - m)$ matrix whose columns are a basis for the null space of A. In other words, Z is full rank and $AZ = 0$. We have that if A has got full row-rank and the reduced Hessian $Z^t G Z$ is positive definite, then the KKT matrix is non singular and we have a unique solution.

2.1 The Null-Space method

The idea is to exploit the block structure in the KKT system to decouple it into two smaller systems. First of all, we decompose $\underline{p}$ into two components as follows

$$\underline{p} = Y\underline{p}_Y + Z\underline{p}_Z$$

where Z is the $n \times (n - m)$ null space matrix, Y is any $n \times m$ matrix such that $det\,[Y|Z] \neq 0$, $\underline{p}_Y$ is an m-vector and $\underline{p}_Z$ is an $(n - m)$-vector. We substitue $\underline{p}$ in the second equation to get

$$A\underline{p} = A\left(Y\underline{p}_Y + Z\underline{p}_Z\right) = A\left(Y\underline{p}_Y\right) + A\left(Z\underline{p}_Z\right) = A\left(Y\underline{p}_Y\right) + 0 = -\underline{h}$$

As AY is such that $rank(AY) = m$, we have that $\underline{p}_Y$ is well determined.

We also substitue the expression into the first equation and have

$$-GY\underline{p}_Y - GZ\underline{p}_Z + A^t\underline{\lambda}^* = g$$

and we pre-multiply by Z^t to get

$$-Z^tGY\underline{p}_Y - Z^tGZ\underline{p}_Z + Z^tA^t\underline{\lambda}^* = Z^tg$$

$$(Z^tGZ)\underline{p}_Z = -Z^tGY\underline{p}_Y - Z^tg$$

and the system can be solved by performing a Cholesky factorisation of the reduced Hessian to get $\underline{p}_Z$. We now have $\underline{p} = Y\underline{p}_Y + Z\underline{p}_Z$ but we are still missing $\underline{\lambda}^*$. To find it, we pre-multiply the first equation by Y^t to obtain

$$(AY)^t\underline{\lambda}^* = Y^t(G\underline{p} + \underline{g})$$

which can be solved for $\underline{\lambda}^*$.

2.2 Active set method for Convex QPs

We now consider problems that present both equalities and inequalities. However, we shall limit ourselves to the case of convex matrices (easier). We notice that if we knew the active set $\mathcal{A}(\underline{x}^*)$ at the optimum (that is the set of active vincles, i.e. inequalities active as equalities plus all the equalities), we could easily apply the null-space method to obtain the solution. Yet, as it is obvious, we do not have any prior knowledge of such set. What we can do, though, is to try and work of various active sets and solve the quadratic subproblems with the method hereby described. We will refer to this set of vincles as to the *working set* and shall denote it with $\mathcal{W}_k$ at the k-th iterate $\underline{x}_k$. Notice that we require all the gradients of the active constraints to be linearly indepent for each working set, even if they would not be so if we considered the whole set $\mathcal{A}(\underline{x}_k)$. So, given an iterate $\underline{x}_k$ and the working set $\mathcal{W}_k$, we first check whether $\underline{x}_k$ minimises the subproblem in the subspace considered. If it is not the case,

we compute a step $\underline{p}$ by solving an equality-constrained QP subproblem. So, let us begin by making some calculations.

$$x = \underline{x}_k + \underline{p}$$

and we also set

$$\underline{g}_k = G\underline{x}_k + \underline{c}$$

We get

$$\frac{1}{2}\left(\underline{x}_k + \underline{p}\right)^t G\left(\underline{x}_k + \underline{p}\right) + \left(\underline{x}_k + \underline{p}\right)^t \underline{c} =$$

$$= \frac{1}{2}\underline{x}_k^t G\underline{x}_k + \underline{p}^t G\underline{x}_k + \frac{1}{2}\underline{p}^t G\underline{p} + \underline{x}_k^t \underline{c} + \underline{p}^t \underline{c} =$$

$$= \frac{1}{2}\underline{p}^t G\underline{p} + \underline{p}^t\left(G\underline{x}_k + \underline{c}\right) + \underline{\rho}_k = \frac{1}{2}\underline{p}^t G\underline{p} + \underline{p}^t \underline{g}_k + \underline{\rho}_k$$

and

$$\underline{\rho}_k = \frac{1}{2}\underline{x}_k^t G\underline{x}_k + \underline{x}_k^t \underline{c}$$

which we can drop from the problem as it is a number. We have

$$\begin{cases} \min\limits_{\underline{p}} \frac{1}{2}\underline{p}^t G\underline{p} + \underline{p}^t \underline{g}_k \\ \underline{a}_i^t \underline{p} = 0, \qquad \forall i \in \mathcal{W}_k \end{cases}$$

The solution of the subproblem is denoted with $\underline{p}_k$. Now, let us try and see what we have . First, we see that for each i in $\mathcal{W}_k$ the value of $\underline{a}_i^t \underline{x}$ does not change as we move along $\underline{p}_k$ as we have $\underline{a}_i^t(\underline{x}_k + \alpha \underline{p}_k) = b_i$ for all α. So the constraints in $\mathcal{W}_k$ are statisfied no matter what is the value of α. So, supposing that are direction is not null, we must decide how far we want to go in the direction we have calculated or, in other words, we need a criterion to fix α_k.

If $\underline{x}_k + \underline{p}_k$ is feasible with respect to all the constraints, we can set

$$\underline{x}_{k+1} := \underline{x}_k + \underline{p}_k$$

otherwise, we want to pick the best α_k we can. So, we know we could take any value if we considered the constraints in $\mathcal{W}_k$. So, we must see what happens for each $i \notin \mathcal{W}_k$. If we have that $\underline{a}_i^t \underline{p}_k \geqslant 0$, for some $i \notin \mathcal{W}_k$, then $\underline{a}_i^t(\underline{x}_k + \alpha_k \underline{p}_k) \geqslant \underline{a}_i^t \underline{x} \geqslant b_i$ and we will have no limit in the choice of $\alpha_k > 0$. But what if $\underline{a}_i^t \underline{p}_k < 0$? Well, in that case we have $\underline{a}_i^t(\underline{x}_k + \alpha_k \underline{p}_k) \geqslant b_i$ only if

$$\alpha_k \leqslant \frac{b_i - \underline{a}_i^t \underline{x}}{\underline{a}_i^t \underline{p}_k}$$

and we also want α_k to be as large as possible in $[0;1]$. We therefore set

$$\alpha_k := \min\left\{1; \min_{i \notin \mathcal{W}_k \,|\, \underline{a}_i^t \underline{p}_k < 0} \frac{b_i - \underline{a}_i^t \underline{x}}{\underline{a}_i^t \underline{p}_k}\right\}$$

and we call the constraint that sets it the blocking constraint.

When the algorithm provides a solution for which $\underline{p} = \underline{0}$, we consider the Lagrange multipliers. If they are all nonnegative, we have a solution. If not, we remove one of the constraints associated with the negative values and iterate.

Part V
Penalty Method

The idea of quadratic penalty method is simple: instead of solving a problem with constraints, we generate a sequence of unconstrained problems such that a penalisation is added to the objective whenever a constraint is violated. We will consider equality-constrained problem in the first place and then see how this can be extended. So, we have a problem of the form

$$\min_{\underline{x} \in \mathbb{R}^n} f(\underline{x}) \; s.t. \; c_i(\underline{x}) = 0, \; \forall i \in \mathcal{E}$$

and we define the following function:

$$Q(\underline{x}; \mu) = f(\underline{x}) + \frac{\mu}{2}\sum_{i \in \mathcal{E}} c_i^2(\underline{x})$$

which we will refer to as to the quadratic penalty function. We consider a sequence $\{\mu_k\}_{k \geq 0}$ such that $\mu_k \uparrow +\infty$ as $k \to +\infty$ and seek an approximate minimiser of $Q(\underline{x}; \mu_k)$ to find a sequence $\{\underline{x}_k\}$.

A general framework for the algorithm can be the following:

1. Define a sequence $\{\tau_k\}_{k \geq 0}$ such that $\tau_k \underset{k \to \infty}{\to} 0$. Set $\mu_0 > 0$ and select a starting point $\underline{x}_0^s$.
2. Solve a sequence of unbounded nonlinear problems until $\|\nabla Q_k\| < \tau_k$
3. If the termination condition are satisfied (e.g. $|f(\underline{x}_{k+1}) - f(\underline{x}_k)|$), then terminate. Otherwise, set $\mu_{k+1} \in (0; \mu_k)$ and starting $\underline{x}_{k+1}^s$. Also set $k := k+1$ and go back to (1)

The parameter sequence can be chosen adaptively, bearing in mind the difficulty of the problem. When the problem is tough, it is good sense to choose small increase of μ_k (e.g. $\mu_{k+1} = 1.5\mu_k$). For easier problems, we can go as far as to set, for example, $\mu_{k+1} = 10\mu_k$. Note that we have no certainty that the condition $\|\nabla Q_k\| < \tau_k$ will be respected as the iterates may move away from the feasible region when the penalty parameter is not large enough. We now analyse convergence by stating two theorems.

Theorem 1 Let $\{\underline{x}_k\}$ be the sequence of minimiser of $Q(\underline{x}; \mu_k)$. We have that $\mu_k \uparrow +\infty$. Then, every limit point of $\{\underline{x}_k\}$ is a global solution of the original problem.

Proof Let us suppose that $\underline{\bar{x}}$ is a global solution. That is

$$f(\underline{\bar{x}}) \leq f(\underline{x}) \; \forall \underline{x} \; with \; c_i(\underline{x}) = 0, \; i \in \mathcal{E}$$

We have that, as $\underline{x}_k$ is a minimiser,

$$f(\underline{x}_k) + \frac{\mu_k}{2} \sum_{i \in \mathcal{E}} c_i^2(\underline{x}_k) \leq f(\underline{\bar{x}})$$

rearranging this becomes

$$\sum_{i \in \mathcal{E}} c_i^2(\underline{x}_k) \leq \frac{2}{\mu_k} (f(\underline{\bar{x}}) - f(\underline{x}_k))$$

Now, let $\underline{x}^*$ be a limit point of the sequence $\{\underline{x}_k\}$, such that there is an infinite subsequence $\mathcal{K}$ such that $\lim_{k \in \mathcal{K}} \underline{x}_k = \underline{x}^*$. We have that

$$\sum_{i \in \mathcal{E}} c_i^2(\underline{x}^*) = \lim_{k \in \mathcal{K}} \sum_{i \in \mathcal{E}} c_i^2(\underline{x}_k) \leq \lim_{k \in \mathcal{K}} \left[\frac{2}{\mu_k} (f(\underline{\bar{x}}) - f(\underline{x}_k)) \right] = 0$$

so we see that $\underline{x}^*$ is a feasible solution of the problem. Moreover, taking the limit in the first expression and having noticed the nonnegativity of the terms, we have

$$f(\underline{x}^*) \leq f(\underline{x}^*) + \lim_{k \in \mathcal{K}} \left[\frac{\mu_k}{2} \sum_{i \in \mathcal{E}} c_i^2(\underline{x}_k) \right] \leq f(\underline{\bar{x}})$$

but then we have

$$\begin{cases} f(\underline{x}^*) \leq f(\underline{\bar{x}}) \\ f(\underline{x}^*) \geq f(\underline{\bar{x}}) \end{cases}$$

and so we have the proof.

This theorem, however, relies on the very strong assumption of finding the *global* minimiser for every subproblem and cannot therefore be applied in general. We want to find out what happens when we allow inexact (although increasingly accurate) minimisations of $Q(\cdot; \mu_k)$.

Theorem 2 Suppose that the tolerances and penalty parameters satisfy $\tau_k \to 0$ and $\mu_k \uparrow +\infty$. Then if a limit point $\underline{x}^*$of the sequence $\{\underline{x}_k\}$ is infeasible, it is a stationary point of the function $\|c(\underline{x})\|^2$. On the other hand, if a limit point $\underline{x}^*$ is feasible and the constraint gradients $\nabla c_i(\underline{x}^*)$ are linearly independent, then $\underline{x}^*$is a KKT point for the problem. For such points, we have for any infinite subsequence $\mathcal{K}$ such that $\lim_{k \in \mathcal{K}} \underline{x}_k = \underline{x}^*$ that

$$\lim_{k \in \mathcal{K}} - \mu_k c_i(\underline{x}_k) = \lambda_i^*, \;\; for \; all \; i \in \mathcal{E},$$

where λ^* is the multiplier vector that satisfies KKT conditions for the equality-constrained problem.

Proof By differentiating $Q(\underline{x}; \mu_k)$, we obtain

$$\nabla_{\underline{x}} Q(\underline{x}; \mu_k) = \nabla f(\underline{x}) + \sum_{i \in \mathcal{E}} \mu_k c_i(\underline{x}_k) \nabla c_i(\underline{x}_k)$$

and, from the termination criterion, we have that

$$\|\nabla f(\underline{x}_k) + \sum_{i \in \mathcal{E}} \mu_k c_i(\underline{x}_k) \nabla c_i(\underline{x}_k)\| \leq \tau_k$$

So, as we know that $\|a\| - \|b\| \leq \|a + b\|$, we find

$$\|\sum_{i \in \mathcal{E}} c_i(\underline{x}_k) \nabla c_i(\underline{x}_k)\| \leq \frac{1}{\mu_k} (\tau_k - \|\nabla f(\underline{x}_k)\|)$$

Now, let $\underline{x}^*$ be a limit point of $\{\underline{x}_k\}$. Then, there exists an infinte subsequence $\mathcal{K}$, such that $\lim_{k \in \mathcal{K}} \underline{x}_k = \underline{x}^*$. When we take it, we have that the part on the right of the expression goes to zero. So we have on the left-hand-side that

$$\sum_{i \in \mathcal{E}} c_i(\underline{x}_k) \nabla c_i(\underline{x}_k) = \underline{0}$$

We can have that $c_i(\underline{x}^*) \neq 0$ (if the constraint gradients $\nabla c_i(\underline{x}^*)$ are linearly dependent), but in that case we certainly have a stationary point of the function $\|c(\underline{x})\|^2$. If, on the other hand, the constraint gradients are linearly independent, then we must have that $c_i(\underline{x}^*) = 0$, *for all* $i \in \mathcal{E}$ and it is therefore feasible. Hence, the second KKT condition is satisfied. We must now check whether the first KKT condition is also satisfied and show that the limit hold.

By using $A(x)$, we denote the Jacobian (i.e. the matrix of the gradients of the constraint), that is

$$A^t(\underline{x}) = [\nabla c_i(\underline{x})]_{i \in \mathcal{E}}$$

and we have

$$\underline{\lambda}_k = -\mu_k c(\underline{x}_k)$$

We therefore have

$$\begin{cases} A^t(\underline{x}_k) \underline{\lambda}_k = \nabla f(\underline{x}_k) - \nabla_{\underline{x}} Q(\underline{x}_k; \mu_k) \\ \|\nabla_{\underline{x}} Q(\underline{x}_k; \mu_k)\| \leq \tau_k \end{cases}$$

for k sufficiently large, the matrix $A^t(\underline{x}_k)$ has full row-rank, so that $A(\underline{x}_k) A^t(\underline{x}_k)$ is nonsingular. So, by pre-multiplying by $A(\underline{x}_k)$, we get

$$\underline{\lambda}_k = \left[A(\underline{x}_k) A^t(\underline{x}_k) \right]^{-1} A(\underline{x}_k) \left[\nabla f(\underline{x}_k) - \nabla_{\underline{x}} Q(\underline{x}_k; \mu_k) \right]$$

And by taking the limit as $k \in \mathcal{K}$ goes to $+\infty$, we have that

$$\lim_{k \in \mathcal{K}} \underline{\lambda}_k = \underline{\lambda}^* = \left[A(\underline{x}^*) A^t(\underline{x}^*) \right]^{-1} A(\underline{x}^*) \left[\nabla f(\underline{x}^*) - \nabla_{\underline{x}} Q(\underline{x}^*; \mu_k) \right] =$$

$$= \left[A(\underline{x}^*)A^t(\underline{x}^*)\right]^{-1} A(\underline{x}^*)\nabla f(\underline{x}^*)$$

So, to conclude, we have that

$$\lim_{k\in\mathcal{K}} \|\nabla f(\underline{x}_k) + \sum_{i\in\mathcal{E}} \mu_k c_i(\underline{x}_k)\nabla c_i(\underline{x}_k)\| = 0$$

so that

$$\nabla f(\underline{x}^*) - A^t(\underline{x}^*)\underline{\lambda}^* = \underline{0}$$

so that $\underline{\lambda}^*$ satisfies the firt KKT condition. Hence, $\underline{x}^*$ is a KKT point for the problem, with unique Lagrange multiplier vector $\underline{\lambda}^*$.

We now see one major issue rising from this method and that is that for $\mu_k \to +\infty$, the matrix $\nabla^2_{\underline{x}\underline{x}} Q(\underline{x};\mu_k)$ becomes ill-conditioned. We see that

$$\nabla^2_{\underline{x}\underline{x}} Q(\underline{x};\mu_k) = \nabla^2 f(\underline{x}) + \mu_k \sum_{i\in\mathcal{E}} c_i(\underline{x})\nabla^2 c_i(\underline{x}) + \mu_k A(\underline{x})A^t(\underline{x})$$

When x is close to the minimiser and the hypotheses of Theorem 2 hold, we have that the first two terms on the right-hand-side of the expression are an approximation of the Hessian of the Lagrangian function, so that we get

$$\nabla^2_{\underline{x}\underline{x}} Q(\underline{x};\mu_k) \approx \nabla^2_{\underline{x}\underline{x}} \mathcal{L}(\underline{x},\underline{\lambda}^*) + \mu_k A(\underline{x})A^t(\underline{x})$$

so that we have

- a matrix whose elements are independent from μ_k
- a matrix of rank $card(\mathcal{E})$ whose nonzero values are of order μ_k

To conclude, we consider the case of problems with inequality constraint. In this case we set the function Q to be

$$Q(\underline{x};\mu) = f(\underline{x}) + \frac{\mu}{2}\sum_{i\in\mathcal{E}} c_i^2(\underline{x}) + \frac{\mu}{2}\sum_{i\in\mathcal{I}} \left([c_i(\underline{x})]^-\right)^2$$

Part VI
Augmented Lagrangian Method

The Augmented Lagrangian Method is not dissimilar from the Quadratic Penalty One. The idea is to introduce an explicit estimate of Lagrange multipliers to avoid ill-conditioning. We will indicate with λ_i^k the value of the i-th component at the k-th iteration of the estimate. We know that the Penalty method introduces a penalisation whenever a constraint is violated, considering the square of it (we are considering the equality-constrained case) and the scaling it by $\frac{\mu}{2}$, as it is clear from the function $Q(\underline{x};\mu) = f(\underline{x}) + \frac{\mu}{2}\sum_{i\in\mathcal{E}} c_i^2(\underline{x})$. However, we know that the approximate minimiser $\underline{x}_k$ of $Q(\underline{x}_k;\mu)$ does not quite satisfy

$$c_i(\underline{x}_k) \approx 0,\ \forall i \in \mathcal{E}$$

but rather

$$c_i(\underline{x}_k) \approx -\frac{\lambda_i^*}{\mu_k},\ \forall i \in \mathcal{E}$$

although we can be sure that $c_i(\underline{x}_k) \overset{\mu_k\uparrow+\infty}{\longrightarrow} 0$. We are interested in eliminating this systematic perturbation. To do so, we consider the function

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

we notice that we have something in between $Q(\underline{x};\mu)$ and the traditional $\mathcal{L}(\underline{x},\underline{\lambda})$. We now want to design an algorithm that fixes the penalty μ to some value $\mu_k > 0$ at its k-th iteration, fixes λ at the current estimate $\underline{\lambda}_k$, and performs minimisation with respect to $\underline{x}$. Using $\underline{x}_k$ to indicate the approximate minimser of $\mathcal{L}_A(\underline{x},\underline{\lambda}^k;\mu_k)$, we have, by the optimality conditions for unconstrained minimisation, that

$$0 \approx \nabla_{\underline{x}}\mathcal{L}_A(\underline{x},\underline{\lambda}^k;\mu_k) = \nabla f(\underline{x}) - \sum_{i\in\mathcal{E}} \left[\lambda_i^k - \mu_k c(\underline{x}_k)\right]\nabla c_i(\underline{x}_k)$$

and we see that

$$\lambda_i^* \approx \lambda_i^k - \mu_k c(\underline{x}_k),\ \forall i \in \mathcal{E}$$

so that by rearranging we have

$$c_i(\underline{x}_k) \approx \frac{1}{\mu_k}\left[\lambda_i^k - \lambda_i^*\right],\ \forall i \in \mathcal{E}$$

so that we have to conclude that if $\underline{\lambda}^k$ is close to the optimal multiplier $\underline{\lambda}$, the infeasibility in $\underline{x}_k$ will be much smaller than $\left(\frac{1}{\mu_k}\right)$ (and it was proportional with quadratic penalty method). The formula also suggests a way for updating the Lagrange multiplier estimate. We set

$$\lambda_i^{k+1} := \lambda_i^k - \mu_k c(\underline{x}_k),\ \forall i \in \mathcal{E}$$

We therefore have an algorithm that follows the following scheme

1. Set μ_0, $\{\tau_k\}_{k\geq 0}$ $s.t.$ $\tau_k \longrightarrow 0$ being a sequence of tolerances, $\underline{x}_0^s$ starting point, $\underline{\lambda}_0$ and set $k := 0$
2. Minimisation is performed untile we find $\|\nabla_{\underline{x}}\mathcal{L}_A(\underline{x},\underline{\lambda}^k;\mu_k)\| < \tau_k$
3. If the solution satisfies the termination criterium, we are done. Otherwise, we have $\lambda_i^{k+1} := \lambda_i^k - \mu_k c(\underline{x}_k)$, $\forall i \in \mathcal{E}$, $\mu_{k+1} \in (0;\mu_k)$, we have $\underline{x}_{k+1}^s$ and $k := k+1$

We now state two theorems that justify the use of this method. The first result show that when we have knowledge of the exact Lagrange multiplier vector $\underline{\lambda}^*$, we have that $\underline{x}^*$, global minimiser of the original problem, is a minimiser of the $\mathcal{L}_A(\underline{x},\underline{\lambda}^k;\mu_k)$ for k sufficiently large. Yes, we do not know the exact $\underline{\lambda}^*$ in real life, yet we know that getting close to it implies having a good idea of $\underline{x}^*$.

Theorem 1 Let $\underline{x}^*$ be a global solution of the problem at which the LICQ is satisfied (that is, the gradients $\nabla c_i(\underline{x}^*)$, $\forall i \in \mathcal{E}$ are linearly independent vectors), and the second-order sufficient condition are satisfied for $\underline{\lambda} = \underline{\lambda}^*$. Then there is a treshold value $\overline{\mu}$ such that for all $\mu > \overline{\mu}$, $\underline{x}^*$ is a stric minimiser of $\mathcal{L}_A(\underline{x}, \underline{\lambda}^*; \mu)$.

The second theorem describes the more realistic situation of $\underline{\lambda} \neq \underline{\lambda}^*$. It gives conditions under which there is a minimiser of $\mathcal{L}_A(\underline{x}, \underline{\lambda}; \mu)$ that lies close to $\underline{x}^*$ and gives error bound on both $\underline{x}_k$ and the updated multiplier estimate λ^{k+1} obtained from solving the subproblem at iteration k.

Theorem 2 Suppose that the assumption of Theorem 1 are satisfied at $\underline{x}^*$ and $\underline{\lambda}^*$ and let $\overline{\mu}$ be chosen as in that theorem. Then there exist positive scalars δ, ε and M such that the follow claims hold:

- For all $\underline{\lambda}^k$ and μ_k satisfying

$$\|\underline{\lambda}^k - \underline{\lambda}^*\| \leq \mu_k \delta, \ \mu_k > \overline{\mu}$$

the problem

$$\min_{\underline{x}} \mathcal{L}_A(\underline{x}, \underline{\lambda}^k; \mu_k) \ \ subject\ to \ \ \|\underline{x} - \underline{x}^*\| \leq \varepsilon$$

has unique solution $\underline{x}_k$. Moreover, we have

$$\|\underline{x}_k - \underline{x}^*\| \leq M \frac{\|\underline{\lambda}^k - \underline{\lambda}^*\|}{\mu_k}$$

- For all $\underline{\lambda}^k$ and μ_k that satisfiy the above conditions, we have

$$\|\underline{\lambda}^{k+1} - \underline{\lambda}^*\| \leq M \frac{\|\underline{\lambda}^k - \underline{\lambda}^*\|}{\mu_k}$$

- For all $\underline{\lambda}^k$ and μ_k that satisfiy the above conditions. the matrix $\nabla^2 \mathcal{L}_A(\underline{x}, \underline{\lambda}^k; \mu_k)$ is positive definite and the constraint gradients $\nabla c_i(\underline{x}_k)$, $\forall i \in \mathcal{E}$, are linearly independent.

This theorem illustrates some salient properties of the augmentend Lagrangian approach. The bound shows that $\underline{x}_k$ will be close to the optimum if $\underline{\lambda}^k$ is accurate or if μ_k is large. Hence we have two ways of improving the solution. And we can have improvement of accuracy of the multipliers by choosing a sufficiently large value of μ_k. The final observation of the theorem shows that second-order sufficient conditions for unconstrained minimisation are also satisfied for the k-th subproblem under the given conditions, so one can expect good performance by applying standard unconstrained optimisation techniques.

Part VII
Barrier Methods

We have a problem with only inequality constrains. That is, we have

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

$$s.t.\ c_i(\underline{x}) \geq 0, \ \ \forall i \in \mathcal{I}$$

and we set $\Omega = \{\underline{x} \in \mathbb{R}^n \ s.t.\ c_i(\underline{x}) \geq 0, \forall i \in \mathcal{I}\}$. We also introduce the idea of **barrier function**. A barrier function is a function that has the following properties:

$$\Phi : \mathbb{R}^n \longrightarrow \mathbb{R}, \ \Phi \in C^0(\mathring{\Omega})$$

and, let $\{\underline{x}_k\}_{k \geq 0}$ be a sequence such that $\underline{x}_k \overset{k \to +\infty}{\longrightarrow} \underline{x}^*$ and $\underline{x}^* \in \partial\Omega$, then $\Phi(\underline{x}_k) \overset{k \to +\infty}{\longrightarrow} +\infty$. Finally, we have that $\forall \underline{x} \in \mathring{\Omega}^C$, $\Phi(\underline{x}) = +\infty$.

The idea is incorporate a term to the objective function so that a sequence of unconstrained minimisation problem can be solved. In this particular case, we will consider the logarithmic barrier problem:

$$P(\underline{x}, \mu) = f(\underline{x}) - \mu \sum_{i \in \mathcal{I}} c_i(\underline{x})$$

where μ is the penalty parameter. We want to consider a positive sequence $\{\mu_k\}_{k \geq 0}$ $s.t.$ $\mu_k \longrightarrow 0$. We set

$$\underline{x}_k \approx \arg \min_{\underline{x} \in \mathbb{R}^n} P(\underline{x}, \mu_k)$$

We now consider a general structure for an algorithm:

1. Set μ_0, $\underline{x}_0^s \in \mathring{\Omega}$, a sequence of tolerances $\{\tau_k\}_{k \geq 0}$ $s.t.$ $\tau_k \longrightarrow 0$ and $k := 0$
2. Solve a sequence of UNP, until we have $\|\nabla P(\underline{x}, \mu_k)\| \leq \tau_k$
3. Apply termination criterion. If it is satisfied, we are done. Otherwise, choose $\mu_{k+1} \in (0; \mu_k)$, $\underline{x}_{k+1}^s$ and set $k := k + 1$

We now wish to see the structural analogies between KKT conditions for the original problem and stationary points of P. We have that KKT conditions are as follows

$$\begin{cases} \nabla f(\underline{x}) - \sum_{i \in \mathcal{I}} \lambda_i \nabla c_i(\underline{x}) = \underline{0} & \\ c_i(\underline{x}) \geq 0 & \forall i \in \mathcal{I} \\ \lambda_i \geq 0 & \forall i \in \mathcal{I} \\ c_i(\underline{x}) \lambda_i = 0 & \forall i \in \mathcal{I} \end{cases}$$

and we set set $\underline{x}(\mu)$ as the minimum of $P(\underline{x}, \mu)$. We see that

$$\nabla_{\underline{x}} P(\underline{x}, \mu) = \nabla f(\underline{x}) - \sum_{i \in \mathcal{I}} \frac{\mu}{c_i(\underline{x})} \nabla c_i(\underline{x}) = \underline{0}$$

and we see that by imposing

$$\tilde{\lambda}_i(\mu) := \frac{\mu}{c_i(\underline{x})}, \ \forall i \in \mathcal{I}$$

we have

$$\nabla_{\underline{x}} P(\underline{x}, \mu) = \nabla f(\underline{x}) - \sum_{i \in \mathcal{I}} \tilde{\lambda}_i(\mu) \nabla c_i(\underline{x}) = \underline{0}$$

which is equivalent to the first KKT condition. We see that the only condition that does not hold is that of complementarity, as we have

$$c_i(\underline{x}) \tilde{\lambda}_i(\mu) = \mu$$

although we have that for $\mu \longrightarrow 0$ we have that the set of equations tend to coincide. We therefore generate a sequence of point on the so-called **central path**

$$\{\underline{x}(\mu), \tilde{\lambda}(\mu) \ s.t. \ \mu > 0\}$$

We now state a theorem.

Theorem Let Ω be such that $\mathring{\Omega} \neq \emptyset$. Let $\underline{x}^* \in \mathring{\Omega}$ be a global minimiser of the problem for some $\underline{\lambda}^*$ at which LICQ holds. We also hypothise that the second-order conditions hold and we have strict complementarity:

$$\forall i \in \mathcal{I}, \ (c_i(\underline{x}) = 0) \ aut \ (\lambda_i = 0).$$

Then,

- there exists a unique continuously differentiable function $\underline{x}(\mu)$, defined for all sufficiently small values of μ by the fact that $\underline{x}(\mu)$ is a local minimiser of $P(\underline{x}, \mu)$ in some neighbourhood of $\underline{x}^*$ such that $\lim_{\mu \to 0^+} \underline{x}(\mu) = \underline{x}^*$
- For that same function, the estimate of the Lagrange multiplier converges to it
$$\lambda_i = \lim_{\mu \to 0^+} \tilde{\lambda}_i(\mu) = \lim_{\mu \to 0^+} \frac{\mu}{c_i(\underline{x})}, \ \forall i \in \mathcal{I}$$
- The Hessian matrix $\nabla_{\underline{x}\underline{x}} P$ is positive definite for all sufficiently small values of μ

To conclude, we briefly consider what happens when we also have equality constraints. In that case, we add a quadratic penalty function to the objective and solve a mixed quadratic penalty/barrier problem.