1 14th of December 2018 — A. Frangioni

In the previous lecture we addressed the problem of linear constrained optimization. Our first approach was to deal very little with constraints (projected gradient method), after a few improvements we took all of them and modify the function (Frank-Wolfe method).

1.1 Dual methods for linear constrained optimization

In this class of methods constraints are first class citizens.

Let us be given the following optimization problem:

$$\min\left\{\frac{1}{2}x^TQx+qx \ : \ Ax\leq b\right\}$$

We would like to work with dual feasibility.

$\forall$ fixed $\lambda \geq 0$, this is the lagrangian problem $\psi(\lambda) = \min\limits_x\{\frac{1}{2}x^TQx + qx + \lambda(b-Ax)\} \leq v$ and the Lagrangian dual is the following:

$$\max\{\psi(\lambda) \ : \ \lambda \geq 0\} \ (D)$$

which is equivalent to the primal problem.

The solution of the Lagrangian problem is **unique** and this is due to the fact that ψ is concave and $Q \succ 0$. The optimal solution is then $x(\lambda) = Q^{-1}(\lambda A - q)$.

The function is differentiable in the solution, because we have only one sub (or super) gradient which is $\nabla\psi(\lambda) = b - Ax(\lambda)$.

At this point we only need to solve the dual problem, which has the shape of a box constrained optimization, where the box has only one boundary and this can be solved via quasi-Newton methods.

Notice that $\psi \notin C^2$ so the Hessian is not defined.

This dual approach is advantageous in the case of a small number of constraints, because the size of the problem decreases. For example, in the case of one constraint, the Lagrangian dual becomes a problem in one variable, hence solvable through a line search.

In this method the degenerate case (more than one constraint active at a time) is not an issue.

If the quadratic function is convex we may use quasi Newton method. Othewise the global optimality is not guaranteed and may be used if we accept not to be able to solve the original problem, but only to find a lower bound.

1.1.1 Separable problems and partial dual

Let us assume that our constraints are separable, which means that it is not mandatory to work with all of them, but they can be splitted into constraints of groups of variables.

$$\min\{f(x) \ : \ Ax \leq b, \ Ex \leq d\}$$

We can decide to use a **partial dual**, writing the Lagrangian problem picking only some constraints that we chose:

$$\psi(\lambda) = \min_x \{ f(x) + \lambda(b - Ax) \ : \ Ex \leq d \}$$

The Lagrangian dual method may be better than projected gradient or worse and it depends on the instance.

In the dual approaches we can't move inside the feasible solution. We find an optimum for the dual, which surely breaks feasibility. Then, if the variable is above the upperbound it gets decreased to the upper bound, otherwise if it is below the lower bound it takes the value of the lower bound.

This way we get an upperbound for the function to be minimized.

1.2 Primal/dual methods or barrier methods

This kind of methods are designed to overcome the cons of dual approaches, namely the fact that ψ does not have the Hessian (and this creates problems to quasi Newton method) and the fact that x is not feasible until the end.

At the same time this methods keep the unconstrained property of the Lagrangian dual.

1.2.1 Barrier function and central path

The rationale behind this algorithm is to minimize a function which penalizes the value of the original function when the solution is getting closer and closer to the boundaries of the feasible set:

$$\min\{ f_\mu(x) = f(x) - \mu \sum_{i=1}^{m} \log(b_i - A_i x) \} \ (P_\mu)$$

The parameter μ is there to weight the proximity to the boundary.

Property 1.1.

- *if f is convex, f_μ is strictly convex;*
- *if $f \in \mathcal{C}^2$ then $f_\mu \in \mathcal{C}^2$, since* $\log \in \mathcal{C}^\infty$*;*
- $\forall \mu \ \exists! \ x_\mu$ *optimal of (P_μ), since* $\mu \sum_{i=1}^{m} \log(b_i - A_i x)$ *is striclty convex;*
- *as $\mu \to 0$ x_μ converges to the analytic center of the optimal face. An example of this behaviour may be seen in Figure 1.1.*

Another interesting property is that, since the barrier function is **self concordant** x^i gets "close" to $x(\mu^i)$ in very few Newton's steps.

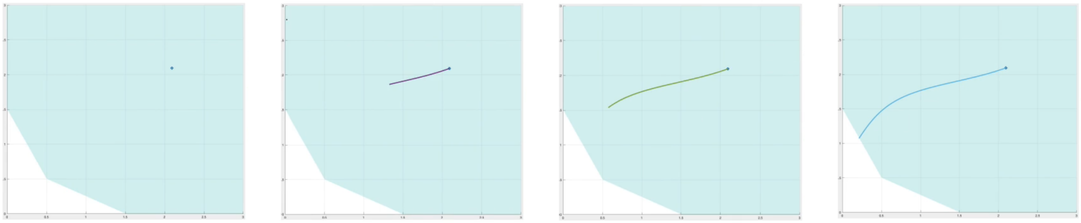

FIGURE 1.1: The trajectory converges to the optimal solution of the problem, when $\mu \to 0$.

In one step x^{i+1} is "much closer" to $x(\mu^i)$ than x^i was and x^{i+1} is "close" to $x(\mu^{i+1})$, with $\mu^{i+1} \ll \mu^i$ (more formally, $\mu^{i+1} = \tau\mu^i$, $\tau < 1$).

This behaviour may be observed in Figure 1.2

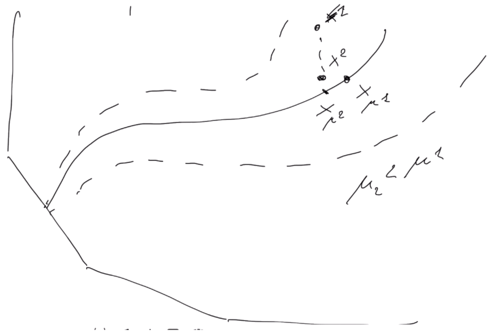

FIGURE 1.2: The dotted line represents a region where the Newton method is very efficient. We are starting from a point x_1, which belongs to that region and we want to move towards x_{μ^1}. At next iterate x_2 is closer to x_{μ^2} than the current iterate.

The convergence is exponential if τ is very small, but these iterations are very costly, because the Hessian changes at each step, hence it needs to be recomputed.

Let us focus on computing the **Newton's step**.

First, we write the Karusch-Kuhn-Tucker conditions:

PRIMAL FEASIBILITY: $Ax + s = b,\ s \geq 0$;

DUAL FEASIBILITY: $Qx + \lambda A = -q,\ \lambda \geq 0$;

COMPLEMENTARY SLACKNESS: $\lambda_i s_i = 0,\ i = 1, \dots, m$.

We can write a slackened version of KKT, imposing $\lambda_i s_i = \mu,\ i = 1, \dots, m$, where $\mu \in \mathbb{R}^n$ and should be decreased over iterations until it gets closer engough to 0.

Let us construct $\Lambda,\ S \in \text{Diag}(m, \mathbb{R})$ such that the diagonal is made of λ_i and s_i respectively.

At this point, we rewrite the problem in terms of the displacement from the fixed current point we are in:

- $x \to x + \Delta x$
- $s \to s + \Delta s$
- $\lambda \to \lambda + \Delta\lambda$

The KKT system becomes:

Primal feasibility: $Ax + A\Delta x + s + \Delta s = b,\ s \geq 0;$

Dual feasibility: $Qx + Q\Delta x + \lambda A + \Delta\lambda A = -q,\ \lambda \geq 0;$

Complementary slackness: $\lambda_i s_i + \lambda_i \Delta s_i + s_i \Delta\lambda_i + \Delta\lambda_i \Delta s_i = \mu,\ i = 1, \dots, m.$

In this new system of coordinates the first two KKT remain linear, while the third one is no longer linear $(\Delta\lambda_i \Delta s_i)$.

$$\begin{bmatrix} Q & A^T & 0 \\ A & 0 & I \\ 0 & S & \Lambda \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta\lambda \\ \Delta s \end{bmatrix} \overset{(1)}{=} \begin{bmatrix} -(Qx+q) - \lambda A \\ b - Ax - s \\ \mu u - \Lambda S u - \Delta\Lambda\Delta S u \end{bmatrix} \approx \begin{bmatrix} 0 \\ 0 \\ \mu u - \Lambda S u \end{bmatrix} \tag{1.1}$$

Where (1) holds since $\Delta\lambda A = A^T \Delta\lambda^T$, although $\Delta\lambda$ is written without the "transpose" sintax to ease notation.

Notice that $u = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \in \mathbb{R}^m$ and has the purpose of adjusting dimension:

$$S\Delta\lambda = \begin{bmatrix} s_1 \Delta\lambda_1 \\ \vdots \\ s_m \Delta\lambda_m \end{bmatrix} \in \mathbb{R}^m;\ \Lambda\Delta s = \begin{bmatrix} \lambda_1 \Delta s_1 \\ \vdots \\ \lambda_m \Delta s_m \in \mathbb{R}^m \end{bmatrix} \in \mathbb{R}^m;\ \mu u = \begin{bmatrix} \mu \\ \vdots \\ \mu \end{bmatrix} \in \mathbb{R}^m;$$

$$\Lambda S u = \begin{bmatrix} -s_1 \lambda_1 \\ \vdots \\ -s_m \lambda_m \end{bmatrix} \in \mathbb{R}^m;\ \Delta\Lambda\Delta s u = \begin{bmatrix} -\Delta\lambda_1 \Delta s_1 \\ \vdots \\ -\Delta\lambda_m \Delta s_m \in \mathbb{R}^m \end{bmatrix} \in \mathbb{R}^m$$

The Newton method can be applied if we discard the non-linear part (written in red), pretending it does not exist. Notice that vector u is the vector of all 1s.

1.3 Primal-dual interior point method

This method is based on the observation that we can solve the dual problem:

$$\max\{-\lambda b - \frac{1}{2} x^T Q x \ :\ Qx + \lambda A = -q,\ \lambda \geq 0\}\ (D)$$

thus obtaining both a lower and upper bound for the solution x.

We term **complementarity gap** $(\frac{1}{2} x^T Q x + qx) - (-\lambda b - \frac{1}{2} x^T Q x) = \lambda s = \mu$.

Once we found a solution for Equation (1.1), we perform a step and compute a new couple of primal and dual solutions and reduce the gap μ.

We are left with solving Equation (1.1). The trick is to express one between Δs and $\Delta\lambda$ as a linear combination of the other.

For example, let us take the third line of Equation (1.1):

$$\begin{aligned} 0\Delta x + S\Delta\lambda + \Lambda\Delta s &= \mu u - \Lambda S u \\ \Lambda\Delta s &= \mu u - \Lambda S u - S\Delta\lambda \\ \Delta s &= \Lambda^{-1}\mu u - \cancel{\Lambda^{-1}}\cancel{\Lambda} S u - \Lambda^{-1}S\Delta\lambda \\ \Delta s &= \Lambda^{-1}\mu u - S u - \Lambda^{-1}S\Delta\lambda \\ \Delta s &= \Lambda^{-1}(\mu u - S\Delta\lambda) - S u \\ \Delta s &= \Lambda^{-1}(\mu u - S\Delta\lambda) - s \end{aligned} \tag{1.2}$$

We obtain the modified normal equations (or KKT system)

$$\begin{bmatrix} Q & A^T \\ A & -\Lambda^{-1}S \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta\lambda \end{bmatrix} = \begin{bmatrix} 0 \\ s - \mu\Lambda^{-1}u \end{bmatrix} \tag{1.3}$$

where the last row is derived working on the second row of Equation (1.1) ($A\Delta x + 0\Delta\lambda + I\Delta s = 0 \Leftrightarrow A\Delta x = -\Delta s$), substituting Equation (1.2):

$$\begin{aligned} A\Delta x - [\Lambda^{-1}S]\Delta\lambda &= -\Delta s - [\Lambda^{-1}S]\Delta\lambda \\ &= -\Lambda^{-1}(\mu u - S\Delta\lambda) + s - \Lambda^{-1}S\Delta\lambda \\ &= -\Lambda^{-1}\mu u + \cancel{\Lambda^{-1}S\Delta\lambda} + s - \cancel{\Lambda^{-1}S\Delta\lambda} \\ &= s - \mu\Lambda^{-1}u \end{aligned} \tag{1.4}$$

With respect to normal equations of **??**, we have in position $(2,2)$ a quantity $(-\Lambda^{-1}S)$, which is not 0, but it is the opposite of a striclty positive definite matrix.

A possible approach for solving this system is making some calculations and obtain something of the shape of reduced KKT (see **??**).

$$\begin{cases} Q\Delta x + A^T\Delta\lambda = 0 \\ (Q + A^T\Lambda S^{-1}A)\Delta x = A^T(\lambda - \mu S^{-1}u) \end{cases} \tag{1.5}$$

where the first set of equations follows from the expansion of the first row of the KKT system (Equation (1.3)) and the second one is obtained taking the same row of the same system and substituting the value of $\Delta\lambda$ as follows:

$$\begin{aligned} A\Delta x - \Lambda^{-1}S\Delta\lambda &= s - \mu\Lambda^{-1}u \\ \Lambda^{-1}S\Delta\lambda &= A\Delta x - s + \mu\Lambda^{-1}u \\ \Delta\lambda &= \left(\Lambda^{-1}S\right)^{-1}A\Delta x - \left(\Lambda^{-1}S\right)^{-1}s + \left(\Lambda^{-1}S\right)^{-1}\mu\Lambda^{-1}u \\ \Delta\lambda &= S^{-1}\Lambda A\Delta x - S^{-1}\Lambda s + \mu S^{-1}\cancel{\Lambda\Lambda^{-1}}u \\ \Delta\lambda &= S^{-1}\Lambda A\Delta x - S^{-1}\Lambda s + \mu S^{-1}u \\ \Delta\lambda &= \mu S^{-1}u + \Lambda S^{-1}A\Delta x - \Lambda S^{-1}s \\ \Delta\lambda &= \mu S^{-1}u + \Lambda S^{-1}A\Delta x - \Lambda u \\ \Delta\lambda &= \mu S^{-1}u + \Lambda S^{-1}A\Delta x - \lambda \end{aligned} \tag{1.6}$$

Hence:

$$\begin{aligned} Q\Delta x + A^T\Delta\lambda &= 0 \\ Q\Delta x + A^T(\mu S^{-1}u + \Lambda S^{-1}A\Delta x - \lambda) &= 0 \\ Q\Delta x + A^T\mu S^{-1}u + A^T\Lambda S^{-1}A\Delta x - A^T\lambda &= 0 \\ Q\Delta x + A^T\Lambda S^{-1}A\Delta x &= -A^T\mu S^{-1}u + A^T\lambda \\ (Q + A^T\Lambda S^{-1}A)\Delta x &= A^T(\lambda - \mu S^{-1}u) \end{aligned} \tag{1.7}$$

We term $M = Q + A^T\Lambda S^{-1}A$ and the following holds.

Fact 1.2. *If A has full column rank (aka it is invertible), $M \succ 0$.*

At this point we need to factorize the matrix M, that changes at each iteration (since ΛS^{-1} does) and this is the bottleneck.

Cholesky factorization may be used, although its complexity is cube. Another downside of this approach is that the matrix M is much denser than A, Λ and S^{-1}.

An orthogonal approach to the reduced KKT is called **predictor-corrector** and it works computing a solution without taking into account the non linear term $\Delta\Lambda\Delta Su$, then computing it according to the approximated solution and repeat until convergence.

The bottlneck again is solving the system in Equation (1.1).

For what concerns implementation, we should start from a triplet (x, λ, s), that could be not feasible and then compute the residuals and iterate until feasibility is reached.

$r^D = -(Qx + q) - \lambda A$, $r^P = b - Ax - s$.

When dealing with the step size we need to highlight the fact that $\lambda + \Delta\lambda \geq 0$, $s + \Delta s \geq 0$ should hold.

In order to achieve this we find the maximum α that satisfies the equality and then multiply it by a constant $\bar{\alpha} = 0.995$ (or 0.9995), in order to get closer.

Let us assume that we also have a bunch of box contraints, hence our problem becomes:

$$\min\left\{\frac{1}{2}x^TQx + qx \ : \ Ax = b,\ 0 \leq x \leq u\right\} \ (P)$$

In this special case, things simplify a lot.