Complementary Notes for Nonlinear Optimization (Class 11)

Marcos Raydan*

May 20, 2020

Penalization methods: starting at Slide 125

One of the main difficulties to propose and study algorithmic ideas to solve constrained minimization problems is the conflicting nature that we need to face. Unconstrained minimization is easy in the sense that we have only one goal: to minimize the objective function. This is not true in the constrained case because there is now a conflict of requirements: the already mentioned objective minimization but at the same time a requirement of feasibility of the solution. It would be very convenient if we could propose a single merit function that takes into account both conflicting requirements, and then apply unconstrained minimization techniques to solve the original constrained problem. We dedicate this class to that goal.

Penalty methods transform constrained optimization problems into a sequence of unconstrained subproblems, whose solutions ideally converge to a solution of the original optimization problem. In that sense, under this convenient practical approach, we can solve constrained problems using our favorite method for unconstrained optimization. We will start by describing the classical and straightforward penalty approach for solving the equality constrained minimization problem:

$$\min_{h(x)=0} f(x), \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable and its gradient $\nabla f(x)$ is available, $h_i : \mathbb{R}^n \to \mathbb{R}$, $1 \leq i \leq m$, and $h_i \in C^1$ for all i.

The main idea is to add to the objective function a positive continuous function (or a term) $p(x)$ that involves the constraints in such a way that it enforces feasibility, i.e., $p(x) = 0$ if and only if x is a feasible point for problem (1), and it will be strictly positive otherwise. Associated with these methods there is also a positive penalty parameter μ that multiplies the additional term, and that will be increased in a dynamically convenient way to penalize constraint violations. Under some mild assumptions, and some specific choices of the penalty term and the sequence of penalty parameters, it can be established that the sequence of solutions of the unconstrained optimization problems converges to a solution of (1) when the penalty parameter μ goes to infinity. However, in practice, when the penalty parameter reaches very high values the convergence might be extremely slow and moreover the subproblems become ill-conditioned.

*Centro de Matemática e Aplicações (CMA), FCT, UNL, 2829-516 Caparica, Portugal (m.raydan@fct.unl.pt).

To circumvent these difficulties it is convenient to bring into play an additional term related to the Lagrange multipliers. The use of $p(x)$ together with this additional term produce the so-called Augmented Lagrangian penalization methods, for which more powerful theoretical and practical results can be obtained. In here, we will discuss the details associated with the theoretical and practical aspects of penalization method and also with augmented Lagrangian methods for solving (1).

Slides 126 and 127: Problem (1) can be reduced to a sequence of unconstrained minimization problems. For that, let us consider for a given constant $\mu > 0$, the minimization of the following function:

$$P(x, \mu) = f(x) + \mu\, p(x) = f(x) + \frac{\mu}{2} h(x)^T h(x) = f(x) + \frac{\mu}{2} \sum_{i=1}^{n} h_i(x)^2,$$

where p is a continuous nonnegative function on $\mathbb{R}^n$, such that $p(x) = 0$ if and only if $h(x) = 0$. In here, the additional term $\mu\, p(x) = \frac{\mu}{2} h(x)^T h(x)$ can be seen as a penalization when the constraints are not satisfied, i.e., when x is not feasible. We call $\mu > 0$ the penalty parameter. Note that the function $P(x, \mu)$ can be seen as a merit function associated with (1), that measures in a combined way optimality and feasibility.

In practice we start with $\mu_0 > 0$ and iteratively solve an unconstrained problem at each k, minimizing $P(x, \mu_k)$, choosing $\mu_{k+1} > \mu_k > 0$, and x_0 at iteration $k + 1$ as the minimizer of the subproblem at iteration k. In Slide 127, this model algorithm is presented. We note that the same role played in Slide 126 by $p(x) = \frac{1}{2}\|h(x)\|_2^2$, can also be played by $p(x) = \|h(x)\|_2$, and by $p(x) = \|h(x)\|_1$, as well as by some other options, as long as p satisfies the fundamental properties: continuous on $\mathbb{R}^n$, $p(x) \geq 0$ for all x, and $p(x) = 0$ if and only if $h(x) = 0$. The choice of $p(x) = \frac{1}{2}\|h(x)\|_2^2$ is quite standard, but depending on the characteristics of some specific applications, some other choices are more convenient.

A key practical issue, that needs to be considered, is that the parameter $\mu_{k+1} > \mu_k$ cannot be increased too drastically and too fast because in that case we can end up simply finding a feasible point which is not taking into account the function $f(x)$, i.e., not taking into account the optimality of the problem. Similarly, increasing the penalty parameter too slowly could avoid taking into account the feasibility of the problem. As a consequence, in practice, the way of increasing the penalty parameter is a delicate issue.

Slides 128 and 129: From a theoretical point of view, the model algorithm presented in Slide 127 has some fundamental properties that we need to establish before presenting the main convergence result.

Lemma 1. *If at every k, x_k is the global minimizer of $P(x, \mu_k) = f(x) + \mu_k\, p(x)$, where $\mu_{k+1} > \mu_k$ and p is a continuous nonnegative function on $\mathbb{R}^n$ such that $p(x) = 0$ if and only if $h(x) = 0$, then:*

$$(i) \quad P(x_k, \mu_k) \leq P(x_{k+1}, \mu_{k+1})$$

$$(ii) \quad p(x_{k+1}) \leq p(x_k)$$

$$(iii) \quad f(x_k) \leq f(x_{k+1}).$$

Moreover, if x^* is the global minimizer of the original problem, it follows that for all k:

$$f(x_k) \leq P(x_k, \mu_k) \leq f(x^*).$$

Proof. (i) Since x_k is the global minimizer of $P(x, \mu_k)$, then $P(x_k, \mu_k) \leq P(z, \mu_k)$ for all $z \in \mathbb{R}^n$. Hence, using that $\mu_{k+1} > \mu_k$, and $p(x) \geq 0$ for all x, we obtain:

$$\begin{aligned} P(x_k, \mu_k) &= f(x_k) + \mu_k\, p(x_k) \leq P(x_{k+1}, \mu_k) = f(x_{k+1}) + \mu_k\, p(x_{k+1}) \\ &\leq f(x_{k+1}) + \mu_{k+1}\, p(x_{k+1}) = P(x_{k+1}, \mu_{k+1}). \end{aligned}$$

For (ii), note that since x_k is the global minimizer of $P(x, \mu_k)$ and x_{k+1} is the global minimizer of $P(x, \mu_{k+1})$, we have

$$P(x_k, \mu_k) = f(x_k) + \mu_k\, p(x_k) \leq f(x_{k+1}) + \mu_k\, p(x_{k+1}),$$

and also

$$P(x_{k+1}, \mu_{k+1}) = f(x_{k+1}) + \mu_{k+1}\, P(x_{k+1}) \leq f(x_k) + \mu_{k+1}\, p(x_k).$$

Adding both inequalities we get

$$(\mu_{k+1} - \mu_k)\, p(x_{k+1}) \leq (\mu_{k+1} - \mu_k)\, p(x_k),$$

and since $\mu_{k+1} > \mu_k$, it follows that $p(x_{k+1}) \leq p(x_k)$.

Concerning (iii), using now (ii) we obtain

$$f(x_k) + \mu_k\, p(x_k) \leq f(x_{k+1}) + \mu_k\, p(x_{k+1}) \leq f(x_{k+1}) + \mu_k\, p(x_k),$$

and hence $f(x_k) \leq f(x_{k+1})$.

Finally, if x^* is the global minimizer of the original problem then $p(x^*) = 0$, and since x_k is the global minimizer of $P(x, \mu_k)$, we have that

$$f(x^*) = f(x^*) + \mu_k p(x^*) \geq P(x_k, \mu_k) = f(x_k) + \mu_k\, p(x_k) \geq f(x_k),$$

and the result is established. □

Theorem 2. *If at every k, x_k is the global minimizer of $P(x, \mu_k)$, and if $\mu_k \to \infty$, then every limit point of the sequence $\{x_k\}$ is a solution of the original problem. Moreover, if $p(x) = \frac{1}{2}\|h(x)\|_2^2$ and we assume regularity of the constraints, then for $1 \leq i \leq m$*

$$\lim_{k\to\infty} \mu_k\, h_i(x_k) = \lambda_i^* \qquad \textit{(Lagrange multipliers)}\ .$$

Proof. Let us assume that $\bar{x} \in \mathbb{R}^n$ is a limit point of $\{x_k\}$, i.e., let $K \subset \mathbb{N}$ be an infinite set such that $\lim_{k\in K} x_k = \bar{x}$. By the continuity of f we have

$$\lim_{k\in K} f(x_k) = f(\bar{x}). \tag{2}$$

Let f^* be the optimal value associated with problem (1). From Lemma 1 the sequence $\{P(x_k, \mu_k)\}$ is nondecreasing and bounded from above by f^*. Therefore,

$$\lim_{k\in K} P(x_k, \mu_k) = \lim_{k\in K} [f(x_k) + \mu_k\, p(x_k)] = p^* \leq f^*. \tag{3}$$

Subtracting (2) from (3), we obtain:

$$\lim_{k\in K} \mu_k\, p(x_k) \;=\; p^* - f(\bar{x}) < \infty. \tag{4}$$

Since $p(x_k) \geq 0$ and $\mu_k \to \infty$, then (4) holds only if

$$\lim_{k\in K} p(x_k) \;=\; 0.$$

By the continuity of p, $p(\bar{x}) = 0$, which means that $h(\bar{x}) = 0$ and we conclude that $\bar{x}$ is feasible. To show that $\bar{x}$ is optimal, it is enough to note that from Lemma 1, $f(x_k) \leq f^*$ and

$$f(\bar{x}) \;=\; \lim_{k\in K} f(x_k) \;\leq\; f^*,$$

and so the optimality of $\bar{x}$ is established.

For the second part of the theorem, let us focus on the unconstrained minimization of

$$P(x,\mu) = f(x) + \frac{\mu}{2} h(x)^T h(x).$$

At iteration k, defining $\lambda_k = \mu_k\, h(x_k)$ and forcing the gradient to be zero, we get

$$\nabla f(x_k) + J_h(x_k)\lambda_k \;=\; 0. \tag{5}$$

Since x^* is a regular solution of (1), then forcing the first order necessary conditions on (1) we obtain that there exists a unique $\lambda^* \in \mathbb{R}^m$ such that

$$\nabla f(x^*) + J_h(x^*)\lambda^* \;=\; 0.$$

Thus, using the pseudo-inverse of $J_h(x^*)$ we have

$$\lambda^* \;=\; -(J_h(x^*))^\dagger \nabla f(x^*), \tag{6}$$

where $(J_h(x^*))^\dagger = (J_h(x^*)^T J_h(x^*))^{-1} J_h(x^*)^T$. Now, since $h \in C^1$, for all k sufficiently large $J_h(x_k)$ is a full rank matrix (rank $= m$), and so from (5), we get

$$\mu_k h(x_k) \;=\; -(J_h(x_k))^\dagger \nabla f(x_k). \tag{7}$$

Hence, taking limits in (7), by the continuity of $[J_h(x)]^\dagger$ near x^*, from (6) it follows that

$$\lim_{k\to\infty} \lambda_k \;=\; \lim_{k\to\infty} \mu_k h(x_k) \;=\; \lambda^*,$$

and the theorem is established. □

Slides 130–133: In Slide 130 we describe a simple two-variable example to illustrate the numerical difficulty that appears when using the standard penalization methods for large values of $\mu > 0$. For that, let us consider the problem

$$\min_{x_1+x_2=2} (x_1^2 + x_2^2).$$

In that case, $P(x,\mu) = x_1^2 + x_2^2 + \frac{\mu}{2}(x_1 + x_2 - 2)^2$, whose gradient and Hessian are given by

$$\nabla_x P(x,\mu) = [2x_1 + \mu(x_1 + x_2 - 2),\; 2x_2 + \mu(x_1 + x_2 - 2)]^T,$$

$$\nabla_x^2 P(x,\mu) = \begin{bmatrix} 2+\mu & \mu \\ \mu & 2+\mu \end{bmatrix}.$$

Notice that the eigenvalues of $\nabla_x^2 P(x,\mu)$ are 2 and $2+2\mu$. Therefore, the condition number $\kappa(\nabla_x^2 P(x,\mu)) = (1+\mu)$, which tends to ∞ if μ tends to ∞.

In Slides 131, 132, and 133, we can see the level curves of $P(x,\mu) = f(x) + \frac{\mu}{2}h(x)^T h(x)$ for $\mu = 1, 4, 10$, respectively. Even though this is a simple example with only two variables, we can see the effect that increasing the penalty parameter μ produces in the elongation of the ellipsoids (level curves of $P(x,\mu)$) which in turn produces an increase in the condition number of the Hessian of $P(x,\mu)$. As we mentioned above, to overcome this negative effect, a more advanced and convenient approach is to use the augmented Lagrangian penalization methods.

Slides 134 and 135: As we notice before, the main difficulty with the standard penalization method is that the penalization parameter $\mu > 0$ must increase up to infinity producing very ill-conditioned Hessian matrices of the unconstrained functions to be minimized, which in turn tends to produce severe numerical instabilities during the convergence process. The augmented Lagrangian methods, to be presented now, includes explicit Lagrange multiplier estimates to avoid the ill-conditioning that is inherent in the penalization method. In that sense, the penalization method can be considered as a precursor to the augmented Lagrangian method. A clear advantage to be discussed, is that it reduces the possibility of ill conditioning of the subproblems that are generated by introducing explicit Lagrange multiplier estimates at each step into the function to be minimized.

For solving (1), let us consider for a given $\mu > 0$, the minimization of:

$$L(x,\lambda,\mu) = f(x) + \lambda^T h(x) + \frac{\mu}{2}h(x)^T h(x).$$

Note that in here we are using the convenient choice $p(x) = \frac{1}{2}\|h(x)\|_2^2$. As before, the additional term $\frac{\mu}{2}h(x)^T h(x)$ is a penalization when $h(x) = 0$ is not satisfied. The Lagrangian term $\lambda^T h(x)$ avoids the use of "very" large values of μ. In here, the vector λ is used as a parameter. Once again, $\mu > 0$ is the penalty parameter, and $L(x,\lambda,\mu)$ can also be seen as a merit function.

As in the penalization method, in practice, we start with $\mu_0 > 0$ and iteratively solve an unconstrained problem at each k, choosing $\mu_{k+1} > \mu_k$, and x_0 at $k+1$ as the minimizer of the subproblem at k. The new difficulty, is that now we also need an estimate of the Lagrangian vector λ_k. In Slide 135 this model algorithm is presented.

Slides 136 and 137: Before discussing how to choose an estimate of the Lagrangian vector λ_k, let us focus on the theoretical properties of the model algorithm related to the augmented Lagrangian method, paying special attention to the specific one that motivates the use of this approach over the standard penalization methods. We need a preliminary lemma from linear algebra.

Lemma 3. *Let $B = B^T \in \mathbb{R}^{n\times n}$ such that $z^T Bz > 0$ for all $z \in null(A)$, $z \neq 0$, and $A \in \mathbb{R}^{m\times n}$. Then, there exists $\bar{\lambda} \geq 0$ such that $B + \lambda A^T A$ is positive definite (PD) for all $\lambda \geq \bar{\lambda}$.*

Proof. (by way of contradiction): Let us suppose that for all $k \in \mathbb{N}$ there exists $x_k \in \mathbb{R}^n$ such that $\|x_k\|_2 = 1$ and

$$x_k^T(B + kA^TA)x_k \leq 0. \tag{8}$$

Clearly, $\{x_k\}$ lives in a compact set (closed and bounded in $\mathbb{R}^n$), and hence there exists $K \subset \mathbb{N}$ (infinite) such that $\lim_{k\in K} = \bar{x}$, where $\|\bar{x}\|_2 = 1$. On the other hand, since A^TA is positive semi-definite, for all k, $x_k^TA^TAx_k \geq 0$; which combined with (8) implies that $\bar{x}^TA^TA\bar{x} = 0$, and we have that $\bar{x} \in null(A)$. Therefore, taking the limit when $k \in K$ goes to infinity in (8), we get $\bar{x}^TB\bar{x} \leq 0$, which is a contradiction. □

Lemma 3 plays a key role to establish the numerical advantage offered by the augmented Lagrangian penalization scheme: while solving the sequence of unconstrained penalized problems, the penalization parameter μ_k does not need to go up to infinity to converge to a solution x^* of the original equality constrained problem.

Theorem 4. *If x^* satisfies the sufficient optimality conditions for the problem $\min_{h(x)=0} f(x)$, and $\lambda^* \in \mathbb{R}^m$ is the vector of Lagrange multipliers at x^*, then there exists $\bar{\mu} \geq 0$ such that the function*

$$L(x) = f(x) + (\lambda^*)^Th(x) + \frac{\mu}{2}h(x)^Th(x)$$

has at x^* a local strict minimizer for all $\mu \geq \bar{\mu}$.

Proof. Notice that

$$\nabla L(x) = \nabla f(x) + J_h(x)\,\lambda^* + \mu J_h(x)\,h(x) = \nabla f(x) + J_h(x)(\lambda^* + \mu\, h(x)). \tag{9}$$

Therefore, $\nabla L(x^*) = 0$, i.e., x^* is also a critical point of $L(x)$. Now, considering the Hessian:

$$\nabla^2 L(x) = \nabla^2 f(x) + \sum_{i=1}^{m} \lambda_i^* \nabla^2 h_i(x) + \mu(J_h(x)J_h(x)^T + \sum_{i=1}^{m} h_i(x)\nabla^2 h_i(x)).$$

Consequently, $\nabla^2 L(x^*) = \nabla^2 \ell(x^*) + \mu(J_h(x^*)J_h(x^*)^T)$, and the result follows by recalling the second order sufficient optimality conditions of nonlinear equality constrained problems, and applying Lemma 3 where $B = \nabla^2 \ell(x^*)$ and $A = J_h(x^*)^T$. □

We note from Theorem 4 that the there is a finite $\bar{\mu} > 0$ (sufficiently large) such that choosing any penalty parameter $\mu \geq \bar{\mu}$ then the obtained unconstrained solution of $L(x)$ is a solution of the original constrained problem. Notice also that it requires the knowledge of the vector of Lagrange multipliers$\lambda^* \in \mathbb{R}^m$ at x^*.

Slides 138, 139, and 140: Unfortunately, we do not have access (in advance) to the Lagrange multipliers $\lambda^* \in \mathbb{R}^m$. Nevertheless, we can estimate them in an iterative way during the convergence process of the augmented Lagrangian penalization scheme. A motivation can be obtained by the development in the proof of Theorem 4.

Notice, from the first order necessary condition, that solving $\min_{h(x)=0} f(x)$ is equivalent to solving

$$\min_{h(x)=0} f(x) + \lambda^T h(x)$$

for any vector λ (e.g., any Lagrange multipliers approximation). If we now add a penalization term (as if we were using the classical penalization approach on this last problem), we obtain

$$\min[f(x) + \lambda^T h(x) + \frac{\mu}{2} h(x)^T h(x)]$$

that for each vector λ is a different problem. Forcing the gradient to zero we have that

$$\nabla f(x) + J_h(x)(\lambda + \mu h(x)) = 0.$$

Recalling the first order conditions, and recalling the expression (9) obtained in the proof of Theorem 4, it seems only natural to use $(\lambda + \mu h(x))$ to estimate the multipliers. In Slide 140 this model algorithm is presented, including the discussed estimation of the Lagrange multipliers at each k.

Notice that in this last model algorithm, instead of increasing the penalty parameter μ_k at every k, it can be kept equal to the previous one, which is a positive consequence of Theorem 4. The classical rule of thumb is that μ_k is increased when at some k we have observed a significant reduction in the value of $L(x, \lambda_k, \mu_k)$. Finally, we note that in addition to the proposed way of estimating the vector of Lagrange multipliers at each k, another option (also inspired by the development in the proof of Theorem 4) is to solve $J_h(x_{k+1})\lambda = -\nabla f(x_{k+1})$ in the least-squares sense to compute the estimation of the Lagrange multipliers at iteration $k + 1$.