

Chapter 10

Penalty, barrier and augmented Lagrangian methods

The material of this chapter is mostly contained in Chapter 17 of Nocedal & Wright [26]. The “old bible” on penalty and barrier methods is Fiacco & McCormick [14].

In this chapter we consider methods for solving a general *constrained optimization problem*

$$\begin{aligned} \min_{\mathbf{x} \in \Omega} \quad & f(\mathbf{x}) \\ \Omega = \{ \mathbf{x} \in \mathbb{R}^n \mid & c_i(\mathbf{x}) = 0, \quad i \in \mathcal{E}, \quad c_i(\mathbf{x}) \geq 0, \quad i \in \mathcal{I} \}. \end{aligned}$$

We consider penalty methods for problems with equality constraints (Section 10.1) and barrier methods for problems with inequality constraints (Section 10.2). These are easy to combine. In Section 10.3 we consider the related, better and more complex class of augmented Lagrangian methods.

10.1 Penalty method

In this section we consider the problem

$$\min \quad f(\mathbf{x}) \tag{10.1a}$$

$$s.t. \quad c_i(\mathbf{x}) = 0, \quad i = 1, 2, \dots, m. \tag{10.1b}$$

Define

$$A^T = [\nabla c_1(\mathbf{x}), \dots, \nabla c_m(\mathbf{x})] \tag{10.1c}$$

and assume that A has full row rank $m \leq n$.

Recall that the Lagrangian is defined by

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) = f(x) - \sum_{i=1}^m \lambda_i c_i(\mathbf{x}) \quad (10.1d)$$

and that the KKT conditions require, in addition to (10.1b), that at $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$,

$$\mathcal{L}_{\mathbf{x}} = \nabla f(\mathbf{x}) - \sum_{i=1}^m \lambda_i \nabla c_i(\mathbf{x}) = \mathbf{0}.$$

In the *quadratic penalty method* we solve a sequence of unconstrained minimization problems of the form

$$\min \phi(\mathbf{x}; \mu) = f(\mathbf{x}) + \frac{1}{2\mu} \sum_{i=1}^m c_i^2(\mathbf{x}) \quad (10.2)$$

for a sequence of values $\mu = \mu_k \downarrow 0$. We can use, for instance, the solution $\mathbf{x}^*(\mu_{k-1})$ of the $(k-1)$ st unconstrained problem as an initial guess for the unconstrained problem (10.2) with $\mu = \mu_k$. This is a simple *continuation technique*.

It is hard to imagine anything simpler to intuit. Unfortunately, however, the problem (10.2) becomes ill-conditioned as μ gets small. Both BFGS and CG methods become severely affected by this. Even Newton's method, which is much less affected by ill-conditioning provided the linear system at each iteration is solved directly, is affected by having its domain of attraction shrink as $\mu \downarrow 0$. Hence the importance of the continuation technique.

An **algorithmic framework** for such a method reads:

Given $\mu_0 > 0$, \mathbf{x}_0 and a final tolerance *tol*,

For $k = 0, 1, 2, \dots$

Starting with \mathbf{x}_k solve (10.2) for $\mu = \mu_k$, terminating when

$$\|\nabla \phi(\mathbf{x}; \mu_k)\| \leq \tau_k$$

where $\tau_k \downarrow 0$. Call the result \mathbf{x}_k^* .

If final convergence test holds (e.g. $\tau_k \leq \text{tol}$) exit

Else

Choose $\mu_{k+1} \in (0, \mu_k)$

Choose a new starting point \mathbf{x}_{k+1} , e.g. $\mathbf{x}_{k+1} = \mathbf{x}_k^*$.

End

End

The choice of how to decrease μ can depend on how difficult it has been to solve the previous subproblem, e.g.,

$$\mu_{k+1} = \begin{cases} 0.7\mu_k & \text{if } \phi(\mathbf{x}; \mu_k) \text{ was hard} \\ 0.1\mu_k & \text{if } \phi(\mathbf{x}; \mu_k) \text{ was easy} \end{cases}$$

When comparing the gradients of the unconstrained objective function $\phi(\mathbf{x}, \mu)$ of (10.2) and the Lagrangian $\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda})$ of (10.1d) it appears that $-\frac{c_i}{\mu}$ has replaced λ_i . Indeed, it can be shown that if $\tau_k \downarrow 0$ then $\mathbf{x}_k^* \rightarrow \mathbf{x}^*$ and

$$-\frac{c_i(\mathbf{x}_k^*)}{\mu_k} \rightarrow \lambda_i^*, \quad i = 1, 2, \dots, m. \quad (10.3)$$

Example 10.1

Let us consider the minimization of the objective functions of Examples 4.1 and 4.2 under the constraint

$$\mathbf{x}^T \mathbf{x} = 1.$$

For each value of penalty parameter μ encountered, we define the function ϕ of (10.2) and call the routine `newtgl`, given in Section 4.3, with the line search option and tolerance $\tau_k = \text{tol} = 1.e - 6$, to solve the subproblem of minimizing ϕ . If more than 9 iterations are needed then the update is $\mu \leftarrow 0.7\mu$, otherwise it is $\mu \leftarrow 0.1\mu$.

For the quadratic 4-variable objective function of Example 4.1 we start with the unconstrained minimizer, $\mathbf{x}_0 = (1, 0, -1, 2)^T$, and obtain convergence after a total of 44 damped Newton iterations to

$$\mathbf{x}^* \approx (-0.02477, 0.31073, -0.78876, 0.52980)^T.$$

The objective value increases from the unconstrained (and infeasible) $f(\mathbf{x}_0) \approx -167.28$ to $f(\mathbf{x}^*) \approx -133.56$. The resulting penalty parameter sequence was

$$\mu = 1, .1, .01, \dots, 1.e - 8,$$

i.e., all subproblems encountered were deemed “easy”, even though for one subproblem a damped Newton step (i.e. step size < 1) was needed. The final approximation for the Lagrange multiplier is

$$(1 - \mathbf{x}^T \mathbf{x})/10^{-8} \approx -40.94.$$

For the non-quadratic objective function of Example 4.2,

$$f(\mathbf{x}) = [1.5 - x_1(1 - x_2)]^2 + [2.25 - x_1(1 - x_2^2)]^2 + [2.625 - x_1(1 - x_2^3)]^2,$$

we start with $\mathbf{x}_0 = \frac{\sqrt{2}}{2}(1, 1)^T$, which satisfies the constraint. Convergence to $\mathbf{x}^* \approx (0.99700, -0.07744)^T$ is reached after a total of 39 iterations. The objective value is $f(\mathbf{x}^*) \approx 4.42$, up from 0 for the unconstrained minimum. The penalty parameter sequence was

$$\mu = 1, .1, .01, .001, 7.e - 4, 7.e - 5, \dots, 7.e - 8,$$

but the path to convergence was more tortuous than these numbers indicate, as solution of the 3rd and 4th subproblems failed. The final approximation for the Lagrange multiplier is

$$(1 - \mathbf{x}^T \mathbf{x}) / (7 \times 10^{-8}) \approx -3.35.$$

◆

To understand the nature of the ill-conditioning better, note that the Hessian of ϕ of (10.2) is

$$\begin{aligned} \nabla^2 \phi(\mathbf{x}; \mu_k) &= \nabla^2 f(\mathbf{x}) + \frac{1}{\mu_k} A^T(\mathbf{x}) A(\mathbf{x}) + \frac{1}{\mu_k} \sum_{i=1}^m c_i(\mathbf{x}) \nabla^2 c_i(\mathbf{x}) \\ &\approx \nabla^2 \mathcal{L} + \frac{1}{\mu_k} A^T A. \end{aligned}$$

The matrix $\frac{1}{\mu_k} A^T A$ has $n - m$ zero eigenvalues as well as m eigenvalues with size $\mathcal{O}(\mu_k^{-1})$. So, we have an unholy mixture of very large and zero eigenvalues. This could give trouble even for Newton's method.

Fortunately, for Newton's iteration, to find the next direction \mathbf{p} we can write the linear system in augmented form (verify!),

$$\begin{pmatrix} \nabla^2 f(\mathbf{x}) + \sum_{i=1}^m \frac{c_i(\mathbf{x})}{\mu_k} \nabla^2 c_i(\mathbf{x}) & A^T(\mathbf{x}) \\ A(\mathbf{x}) & -\mu_k I \end{pmatrix} \begin{pmatrix} \mathbf{p} \\ \boldsymbol{\xi} \end{pmatrix} = \begin{pmatrix} -\nabla_{\mathbf{x}} \phi(\mathbf{x}; \mu_k) \\ \mathbf{0} \end{pmatrix}.$$

This matrix tends towards the KKT matrix and all is well in the limit.

Finally, we note for later purposes that instead of the quadratic penalty function (10.2) the function

$$\phi_1(\mathbf{x}; \mu) = f(\mathbf{x}) + \frac{1}{\mu} \sum_{i=1}^m |c_i(\mathbf{x})| \quad (10.4)$$

could be considered. This is an *exact penalty function*: for sufficiently small $\mu > 0$ one minimization yields the optimal solution.

Unfortunately, that one unconstrained minimization problem turns out in general practice to be harder to solve than applying the continuation method presented before with the quadratic penalty function (10.2). (Why are we not surprised?) But the function (10.4) also has other uses, namely, as a *merit function*. Thus, (10.4) can be used to assess the quality of iterates obtained by some other method for constrained optimization.

10.2 Barrier method

We now consider only inequality constraints,

$$\min \quad f(\mathbf{x}) \quad (10.5a)$$

$$s.t. \quad c_i(\mathbf{x}) \geq 0, \quad i = 1, 2, \dots, m. \quad (10.5b)$$

We may or may not have $m > n$, but we use the same notation $A(\mathbf{x})$ as in (10.1c), dropping the requirement of a full row rank.

In the *log barrier method* we solve a sequence of unconstrained minimization problems of the form

$$\min \psi(\mathbf{x}; \mu) = f(\mathbf{x}) - \mu \sum_{i=1}^m \log c_i(\mathbf{x}) \quad (10.6)$$

for a sequence of values $\mu = \mu_k \downarrow 0$.

Starting with a feasible \mathbf{x}_0 in the interior of Ω we always stay strictly in the interior of Ω , so this is an *interior point method*. This feasibility is a valuable property if we stop before reaching optimum.

Example 10.2

Here is a very simple problem:

$$\min_x \quad x, \quad x \geq 0.$$

By (10.6),

$$\psi(x; \mu) = x - \mu \log x.$$

Setting $\mu = \mu_k$ we consider $0 = \frac{d\psi}{dx} = 1 - \mu/x$, from which we get $x_k = \mu_k \rightarrow 0 = x^*$.



Clearly, as $\mu_k \downarrow 0$ we expect numerical trouble, just like for the penalty method in Section 10.1. The algorithmic framework is also the same as for the penalty method, with ψ replacing ϕ there.

Note that

$$\nabla_{\mathbf{x}}\psi(\mathbf{x}; \mu) = \nabla f(\mathbf{x}) - \sum_{i=1}^m \frac{\mu}{c_i(\mathbf{x})} \nabla c_i(\mathbf{x}).$$

Comparing this to the gradient of the Lagrangian (10.1d) we expect that, as $\mu_k \downarrow 0$,

$$\frac{\mu_k}{c_i(\mathbf{x}_k^*)} \rightarrow \lambda_i^*, \quad i = 1, 2, \dots, m. \quad (10.7)$$

For the strictly inactive constraints ($c_i > 0$) we get $\lambda_i^* = 0$ in (10.7), as we should.

It is easy to see that, sufficiently close to the optimal solution,

$$\begin{aligned} \nabla^2\psi(\mathbf{x}; \mu_k) &\approx \nabla^2\mathcal{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*) + \sum_{i=1}^m \frac{(\lambda_i^*)^2}{\mu_k} \nabla c_i(\mathbf{x}) \nabla c_i(\mathbf{x})^T \\ &= \nabla^2\mathcal{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*) + \sum_{i \in \mathcal{A}(\mathbf{x}^*)} \frac{(\lambda_i^*)^2}{\mu_k} \nabla c_i(\mathbf{x}) \nabla c_i(\mathbf{x})^T. \end{aligned}$$

This expresses ill-conditioning exactly as in the penalty case (unless there are no active constraints at all).

Let us denote by $\mathbf{x}(\mu)$ the minimizer of $\psi(\mathbf{x}; \mu)$, and let (for $\mu > 0$)

$$\lambda_i(\mu) = \frac{\mu}{c_i(\mathbf{x}(\mu))}, \quad i = 1, \dots, m. \quad (10.8)$$

Then $\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}(\mu), \boldsymbol{\lambda}(\mu)) = \nabla_{\mathbf{x}}\psi = \mathbf{0}$. Also, $\mathbf{c}(\mathbf{x}) > \mathbf{0}$, $\boldsymbol{\lambda} > \mathbf{0}$. So, all KKT conditions hold except for complementarity. In place of complementarity we have

$$c_i(\mathbf{x}(\mu))\lambda_i(\mu) = \mu > 0, \quad i = 1, \dots, m.$$

We are therefore on the *center path* defined artificially by (10.8),

$$\mathcal{C}_{pd} = \{(\mathbf{x}(\mu), \boldsymbol{\lambda}(\mu), \mathbf{s}(\mu)) \mid \mu > 0\}, \quad (10.9)$$

where $\mathbf{s} = \mathbf{c}(\mathbf{x})$ are the slack variables.

For the primal-dual formulation we can therefore write the above as

$$\nabla f(\mathbf{x}) - A(\mathbf{x})^T \boldsymbol{\lambda} = \mathbf{0}, \quad (10.10a)$$

$$\mathbf{c}(\mathbf{x}) - \mathbf{s} = \mathbf{0}, \quad (10.10b)$$

$$\Lambda \mathbf{S} \mathbf{e} = \mu \mathbf{e}, \quad (10.10c)$$

$$\boldsymbol{\lambda} > \mathbf{0}, \mathbf{s} > \mathbf{0}, \quad (10.10d)$$

where, as in Section 8.2, Λ and S are diagonal matrices with entries λ_i and s_i , respectively, and $\mathbf{e} = (1, 1, \dots, 1)^T$.

The setup is therefore that of *primal-dual interior point methods*. A *modified* Newton iteration for the equalities in (10.10) reads

$$\begin{pmatrix} \nabla_x^2 \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) & -A^T(\mathbf{x}) & 0 \\ A(\mathbf{x}) & 0 & -I \\ 0 & S & \Lambda \end{pmatrix} \begin{pmatrix} \delta \mathbf{x} \\ \delta \boldsymbol{\lambda} \\ \delta \mathbf{s} \end{pmatrix} = \begin{pmatrix} -\nabla f + A^T \boldsymbol{\lambda} \\ \mathbf{s} - \mathbf{c} \\ \mu \mathbf{e} - \Lambda S \mathbf{e} + \mathbf{r}_{\lambda, s} \end{pmatrix}. \quad (10.11)$$

The modification term $\mathbf{r}_{\lambda, s}$ (which does not come out of Newton's methodology at all!) turns out to be crucial for both theory and practice.

Upon solving (10.11) we update

$$\mathbf{x} \leftarrow \mathbf{x} + \alpha \delta \mathbf{x}, \quad \boldsymbol{\lambda} \leftarrow \boldsymbol{\lambda} + \alpha \delta \boldsymbol{\lambda}, \quad \mathbf{s} \leftarrow \mathbf{s} + \alpha \delta \mathbf{s},$$

choosing α so that the inequalities (10.10d) are obeyed.

A lot of recent effort has gone into research of this type of methods and many people have become disappointed. There are open problems here. See the survey [17].

10.3 Augmented Lagrangian method

Consider the problem with only equality constraints, (10.1). The basic difficulty with the quadratic penalty method has been that elusive limit of dividing 0 by 0. Let us therefore consider instead adding the same penalty term to the Lagrangian, rather than to the objective function.

Thus, define the *augmented Lagrangian*

$$\begin{aligned} \mathcal{L}_A(\mathbf{x}, \boldsymbol{\lambda}; \mu) &= \mathcal{L}(\mathbf{x}) + \frac{1}{2\mu} \sum_{i=1}^m c_i^2(\mathbf{x}), \\ &= f(\mathbf{x}) - \sum_{i=1}^m \lambda_i c_i(\mathbf{x}) + \frac{1}{2\mu} \sum_{i=1}^m c_i^2(\mathbf{x}). \end{aligned} \quad (10.12)$$

The KKT conditions require, to recall, that $\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*) = 0$ and $\mathbf{c}(\mathbf{x}^*) = \mathbf{0}$, so at the optimum the augmented Lagrangian coincides with the Lagrangian, and μ no longer need be small!

In fact, at some non-critical point,

$$\nabla_{\mathbf{x}} \mathcal{L}_A(\mathbf{x}, \boldsymbol{\lambda}; \mu) = \nabla f(\mathbf{x}) - \sum_{i=1}^m \left[\lambda_i - \frac{c_i(\mathbf{x})}{\mu} \right] \nabla c_i(\mathbf{x}),$$

hence we expect near a critical point that

$$\lambda_i - \frac{c_i(\mathbf{x})}{\mu} \approx \lambda_i^*, \quad i = 1, \dots, m. \quad (10.13)$$

We can choose some $\mu > 0$ not very small. The minimization of the augmented Lagrangian (10.12) therefore yields a *stabilization*, replacing $\nabla^2 f$ by $\nabla^2 f + \frac{1}{\mu} A^T A$. Thus, the Hessian matrix (wrt to \mathbf{x}) of the augmented Lagrangian, $\nabla_x^2 \mathcal{L}_A$, is s.p.d. provided that the reduced Hessian matrix of the Lagrangian, $Z^T (\nabla_x^2 \mathcal{L}) Z$, is s.p.d. It can further be shown that for μ small enough the minimization of (10.12) with respect to \mathbf{x} at $\boldsymbol{\lambda} = \boldsymbol{\lambda}^*$ yields \mathbf{x}^* .

Moreover, the formula (10.13) suggests a way to update λ_i in a penalty-like sequence of iterates:

$$\lambda_i^{k+1} = \lambda_i^k - \frac{c_i(\mathbf{x}_k^*)}{\mu_k}.$$

In fact, one asks next, why not update also μ while updating λ ? This then leads to the following **algorithmic framework**:

Given $\mu_0 > 0$, \mathbf{x}_0 , $\boldsymbol{\lambda}_0$ and a final tolerance tol ,

For $k = 0, 1, 2, \dots$

Starting with \mathbf{x}_k minimize $\mathcal{L}_A(\mathbf{x}, \boldsymbol{\lambda}_k; \mu_k)$, terminating when

$$\|\nabla \mathcal{L}_A(\mathbf{x}, \boldsymbol{\lambda}_k; \mu_k)\| \leq \tau_k$$

Call the result \mathbf{x}_k^* .

If final convergence test holds (e.g. $\tau_k \leq tol$) exit

Else

$$\boldsymbol{\lambda}_{k+1} = \boldsymbol{\lambda}_k - \mu_k^{-1} \mathbf{c}(\mathbf{x}_k^*)$$

Choose $\mu_{k+1} \in (0, \mu_k)$

Choose a new starting point \mathbf{x}_{k+1} , e.g. $\mathbf{x}_{k+1} = \mathbf{x}_k^*$.

End

End

This allows a gentler decrease of μ : both primal and dual variables participate in the iteration. The constraints satisfy

$$\frac{c_i(\mathbf{x}_k^*)}{\mu_k} = \lambda_i^k - \lambda_i^{k+1} \rightarrow 0, \quad 1 \leq i \leq m,$$

a clear improvement over the expression (10.3) relevant for the quadratic penalty method.

Example 10.3

Let us repeat the experiments of Example 10.1 using the Augmented Lagrangian method. We use the same parameters and starting points, with $\lambda_0 = 0$.

For the quadratic 4-variable objective function of Example 4.1 we obtain convergence after a total of 29 Newton iterations. No damping was needed. The resulting penalty parameter sequence is

$$\mu = 1, .1, .01, \dots, 1.e - 5,$$

and the corresponding Lagrange multiplier estimates are

$$\lambda = 0, -2.09, -12.78, -32.85, -40.59, -40.94.$$

For the non-quadratic objective function of Example 4.2, we obtain convergence after a total of 28 iterations. The penalty parameter sequence is

$$\mu = 1, .1, .01, .001, 1.e - 4,$$

and the corresponding Lagrange multiplier estimates are

$$\lambda = 0, 9.7e - 7, -2.63, -3.33, -3.35.$$

The smallest values of μ required here are much larger than in Example 10.1, and no difficulty is encountered in the path to convergence for the augmented Lagrangian method: the advantage over the penalty method of Example 10.1 is more than the iteration counts alone indicate. ◆

It is possible to extend the augmented Lagrangian method directly for inequality constraints, see [26]. But instead we can use slack variables. Thus, for a given constraint

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

we write

$$c_i(\mathbf{x}) - s_i = 0, \quad s_i \geq 0.$$

For the general problem (7.1) this yields the problem with equality constraints plus nonnegativity constraints

$$\min_{\mathbf{x}, \mathbf{s}} \quad f(\mathbf{x}), \tag{10.14a}$$

$$c_i(\mathbf{x}) = 0, \quad i \in \mathcal{E}, \tag{10.14b}$$

$$c_i(\mathbf{x}) - s_i = 0, \quad i \in \mathcal{I}, \tag{10.14c}$$

$$\mathbf{s} \geq \mathbf{0}. \tag{10.14d}$$

For the latter problem we can utilize a mix of the augmented Lagrangian method applied for the equality constraints and the gradient projection method, as described in Section 9.3, applied for the nonnegativity constraints.

This is the approach taken by the highly successful general-purpose code LANCELOT by Conn, Gould and Toint [11]. In the algorithmic framework presented earlier we now have the subproblem

$$\begin{aligned} \min_{\mathbf{x}, \mathbf{s}} \quad & \mathcal{L}_A(\mathbf{x}, \mathbf{s}, \boldsymbol{\lambda}; \mu) \\ \text{s.t.} \quad & \mathbf{s} \geq \mathbf{0}, \end{aligned} \tag{10.15}$$

where $\boldsymbol{\lambda}$ and μ are held fixed when (10.15) is solved. Since this is within an outer iteration it makes sense to derive a quadratic model for the augmented Lagrangian in (10.15) and solve the resulting QP using the method described in Section 9.3.