By the completion of this lecture you should be able to:

1. Describe and analyse penalty methods for solving constrained optimization.
2. Describe and analyse augmented Lagrange method for solving constrained optimization.

Reference:

- *Chapter 17, Jorge Nocedal and Stephen J. Wright, 'Numerical Optimization'.*

1.1 Introduction

Generally, in optimization problems, constraints complicate the algorithmic solution and limit the range of available algorithms. For this reason, it is natural to try to eliminate constraints. The main idea in the methods to be examined in this lecture is to approximate the constrained minimization problem by a considerably easier problem to solve. Naturally, we can only expect to obtain an approximate solution to the original problem by solving an approximate problem. However, suppose we can construct a sequence of approximate problems that converge in a well-defined sense to the original problem. In that case, hopefully, the corresponding sequence of approximate solutions will yield the solution of the actual problem.

It may appear odd at first sight that we would prefer solving a sequence of minimization problems rather than a single problem. However, in practice, only a finite number of approximate problems need to be solved to obtain an acceptable approximate solution of the original problem. Furthermore, each approximate problem need not be solved itself exactly but rather only approximately. In addition, one may efficiently utilize information obtained from each approximate problem to search for a solution to the next approximate problem.

We will discuss penalty and augmented Lagrangian methods. We will state the convergence theorems without proofs. You are not required to know the proofs. If you are interested in proof; refer to the reference.

1.2 The Quadratic Penalty Method

The basic idea in penalty methods is to eliminate the constraints and add to the objective function a penalty term that prescribes a high cost to infeasible points. Associated with these methods is a positive parameter μ, which determines the severity of the penalty and the extent to which the resulting unconstrained problem approximates the original constrained problem. By making μ large, we penalize the constraint violations more severely, whereby forcing the minimizer of the penalty function closer to the feasible region for the constrained problem.

The most straightforward penalty function of this type is the *quadratic penalty function*, in which the penalty terms are the squares of the constraint violations. We describe this approach first in the context of the equality constrained problem

$$\min_{\mathbf{x}} f(\mathbf{x}) \quad \text{subject to} \quad c_i(\mathbf{x}) = 0, \quad i \in E. \tag{1.1}$$

The quadratic penalty function $Q(\mathbf{x}, \mu)$ for this formulation is

$$Q(\mathbf{x}, \mu) = f(\mathbf{x}) + \frac{\mu}{2} \sum_{i \in E} c_i^2(\mathbf{x}), \tag{1.2}$$

where $\mu > 0$ is the penalty parameter. By driving μ to ∞, we penalize the constraint violations with increasing severity. It makes good intuitive sense to consider a sequence of values $\{\mu_k\}$ satisfying

$$0 < \mu_k < \mu_{k+1} \quad \forall k, \quad \mu_k \to \infty,$$

and seek the approximate minimizer $\mathbf{x}^k$ of $Q(\mathbf{x}, \mu_k)$ for each k. The penalty terms in the penalty objective function (1.2) are smooth; thus we can use techniques from unconstrained optimization to search for the approximate solution $\mathbf{x}^k$. We can use previous approximate solutions of $Q(\cdot, \mu)$ for smaller μ to construct an initial guess. For a suitable choice of the sequence $\{\mu_k\}$ and the initial guess, just a few steps of unconstrained minimization may be needed for each μ_k.

Example 1.1. *Consider the constrained minimization problem*

$$\min x_1 + x_2 \qquad \textit{subject to } x_1^2 + x_2^2 = 2. \tag{1.3}$$

Examine the effect of the penalty parameter μ in the quadratic penalty function (1.2).

Take note of the following deficiency of penalty methods. The penalty function may be unbounded below the given value of the penalty parameter μ even if the original constrained problem has a unique solution. This deficiency is, unfortunately, common to most penalty functions.

In the case of general constrained optimization problem, which contains equality constraints as well as inequality constraints

$$\min_{\mathbf{x}} f(\mathbf{x}) \quad \text{subject to } \begin{cases} c_i(\mathbf{x}) = 0, & i \in E, \\ c_i(\mathbf{x}) \geq 0, & i \in I, \end{cases} \tag{1.4}$$

the quadratic penalty function is defined by

$$Q(\mathbf{x}, \mu) = f(\mathbf{x}) + \frac{\mu}{2} \sum_{i \in E} c_i^2(\mathbf{x}) + \frac{\mu}{2} \sum_{i \in I} ([c_i(\mathbf{x})]^-)^2, \tag{1.5}$$

where $[y]^-$ denotes $\max(-y, 0)$. In this case, Q may be less smooth than the objective and constraints functions. For instance, if one of the inequality constraints is $x_1 \geq 0$, then the function $\min(0, x_1)^2$ has a discontinuous second derivative, so Q is no longer twice continuously differentiable.

Algorithmic framework

A general algorithmic framework based on the quadratic penalty function (1.2) is specified as follows:

Algorithm 1.1 Quadratic Penalty Method

Choose $\mu_0 > 0$, a non-negative sequence $\{\tau_k\}$ with $\tau_k \to 0$, and set the starting point $\mathbf{x}_s^0$.
for $k = 0, 1, \ldots$ **do**
    Find an approximate minimizer $\mathbf{x}^k$ of $Q(\mathbf{x}, \mu_k)$, starting at $\mathbf{x}_s^k$,
        and terminate when $\|\nabla_x Q(\mathbf{x}, \mu_k)\| \leq \tau_k$;
    if final convergence test is satisfied **then**
        Stop with approximate solution x_k;
    end if
    Choose a new penalty parameter $\mu_{k+1} > \mu_k$;
    Choose new starting point $\mathbf{x}_s^{k+1}$;
end for

The parameter sequence $\{\mu_k\}$ can be updated adaptively based on minimizing the penalty function at each iteration. When minimization of $Q(\mathbf{x}, \mu_k)$ proves to be expensive for some k, we choose μ_{k+1} to be only modestly large than μ_k; for instance $\mu_{k+1} = 1.5\mu_k$. If we find the approximate minimizer of $Q(\mathbf{x}, \mu_k)$ cheaply, we could try a more ambitious increase, for instance $\mu_{k+1} = 10\mu_k$. The convergence theory for **Algorithm 1.1** allows much freedom in the choice of the non-negative tolerance τ_k. It only requires that $\tau_k \to 0$, to ensure that the minimization is carried out more accurately as the iteration progress.

There is no guarantee that the test $\|\nabla_x Q(\mathbf{x}, \mu_k)\| \leq \tau_k$ will be satisfied. Because the iterates may move away from the feasible region when the penalty parameter is not large enough. A practical implementation must include safeguards that increase the penalty parameter when the constraint violation is not decreasing rapidly enough or when the iterate appear to be diverging.

When only equality constraints are present, $Q(\mathbf{x}, \mu_k)$ is smooth, the algorithms for unconstrained minimization problems described in this course can be used to identify the approximate solution $\mathbf{x}^k$. However, the minimization of $Q(\mathbf{x}, \mu_k)$ becomes more difficult to perform as μ_k becomes large. The Hessian $\nabla^2_{xx} Q(\mathbf{x}, \mu_k)$ becomes arbitrarily ill-conditioned near minimizer. This property alone is enough to make many unconstrained minimization algorithms such as steepest gradient, Newtons method, conjugate gradient and quasi-Newton perform poorly.

Convergence

We describe some convergence properties of the quadratic penalty method. We restrict our attention to the equality constrained problem. We assume that the penalty function $Q(\mathbf{x}, \mu_k)$ has a minimizer for each value of μ_k, the following theorem constitutes the primary convergence results for the quadratic penalty method

Theorem 1.1. *Suppose that each $\mathbf{x}^k$ is the exact global minimizer of $Q(\mathbf{x}, \mu_k)$ define by (1.2) in the **Algorithm 1.1** and $\mu_k \to \infty$. Then every limit point $\mathbf{x}^*$ of the sequence $\{\mathbf{x}^k\}$ is global solution of problem (1.4).*

Since the results require us to find the global minimizer for each subproblem, this desirable property of convergence to the global solution of (1.4) cannot be attained in general. The next result concerns convergence properties of the sequence $\{\mathbf{x}^k\}$ when we allow inexact minimization of $Q(\mathbf{x}, \mu_k)$. It shows that the sequence may be attracted to infeasible points or any KKT point rather than to a minimizer. It also shows that the quantities $\mu_k c_i(\mathbf{x^k})$ may be used as an estimate of the Lagrange multiplier λ_i^* in certain circumstances.

Theorem 1.2. *Suppose the tolerance and penalty parameters in **Algorithm 1.1** satisfy $\tau_k \to 0$ and $\mu_k \to \infty$. Then if a limit point $\mathbf{x}^*$ of the sequence $\{\mathbf{x}^k\}$ is infeasible, it is a stationary point of the function $\|c(\mathbf{x})\|^2$. On the other hand, if a limit point $\mathbf{x}^*$ is feasible and the constraint gradients $\nabla c_i(\mathbf{x}^*)$ are linearly independent, then $\mathbf{x}^*$ is a KKT point for problem (1.4). For such points, we have for any infinite subsequence K such that $\lim_{k\in K} \mathbf{x}^k = \mathbf{x}^*$ that*

$$\lim_{k\in K} -\mu_k c_i(\mathbf{x}^k) = \lambda_i^*, \quad \text{for all } i \in E, \tag{1.6}$$

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

It is reassuring that, if a limit point $\mathbf{x}^*$ is not feasible, it is at least a stationary point for the function $\|c(\mathbf{x})\|^2$.

The problem of ill conditioning

Since the penalty method is based on the solution of the subproblem (solving the quadratic penalty function), it is natural to inquire about the degree of difficulty in solving such problems. An understanding of the properties of the Hessian matrix $\nabla^2_{xx} Q(\mathbf{x}, \mu_k)$ is essential in choosing practical algorithms for the

minimization problem. The Hessian is calculated as

$$\nabla^2_{xx} Q(\mathbf{x}, \mu_k) = \nabla^2 f(\mathbf{x}) + \sum_{i \in E} \mu_k c_i(\mathbf{x}) \nabla^2 c_i(\mathbf{x}) + \mu_k J(\mathbf{x})^T J(\mathbf{x}). \tag{1.7}$$

Here, $J(\mathbf{x})^T = [\nabla c_i]_{i \in E}$ is the matrix of constraint gradients. When $\mathbf{x}$ is close to the minimizer of $Q(\mathbf{x}, \mu_k)$ and the conditions of **Theorem 1.2** are satisfied, we have from (1.6) that the sum of the first two terms of the right-hand-side of (1.7) is approximately equal to the Hessian of the Lagrangian function. To be specific, we have

$$\nabla^2_{xx} Q(\mathbf{x}, \mu_k) \approx \nabla^2_{xx} L(\mathbf{x}, \lambda^k) + \mu_k J(\mathbf{x})^T J(\mathbf{x}). \tag{1.8}$$

when $\mathbf{x}$ is close to the minimizer of $Q(\mathbf{x}, \mu_k)$. We see from this expression that $\nabla^2_{xx} Q(\mathbf{x}, \mu_k)$ is approximately equal to the sum of

- a matrix whose elements are independent of μ_k (the Lagrangian term) and
- a matrix of rank $|E|$ whose non-zero eigenvalues are of order μ_k (the second term on the right-hand side of (1.8)).

The number of constraints $|E|$ is usually smaller than n. In this case, the last term in (1.8) is singular. The overall matrix has some of its eigenvalues approaching a constant, while others are of order μ_k. Since μ_k is approaching ∞, the increasing ill-conditioning of $\nabla^2_{xx} Q(\mathbf{x}, \mu_k)$ is apparent. The corresponding unconstrained optimization problem becomes difficult to solve.

The ill-conditioning associated with unconstrained minimization problems is a basic characteristic feature of penalty methods and represents the overriding factor in determining how these methods are operated. Ill-conditioning can sometimes be overcome by using a starting point for the unconstrained minimization method, which is close to a minimizing point of $Q(\mathbf{x}, \mu_k)$. Usually, one adopts as a starting point the last point $\mathbf{x}^{k-1}$ of the previous minimization. In order for $\mathbf{x}^{k-1}$ to be near a minimizing point of $Q(\mathbf{x}, \mu_k)$, it is necessary that μ_k is close to μ_{k-1}. This in turn implies that the rate of increase of the penalty parameter μ_k should be relatively low. If the μ_k is increased at a fast rate, then the convergence of the method is fast, albeit at the expense of ill-conditioning. In practice, one must operate the method to balance the benefit of fast convergence with the irksome ill-conditioning.

1.3 Augmented Lagrangian Method

We shift our discussion to an approach known as *augmented Lagrangian method* (also known as *method of multiplier*). This algorithm is related to the quadratic penalty method, but it reduces the possibility of ill-conditioning by introducing an explicit Lagrange multiplier estimate into the function to be minimized.

We first consider the equality-constrained problem (1.4). The quadratic penalty function $Q(\mathbf{x}, \mu)$ defined by (1.2) penalizes constraint violations by squaring the infeasibilities and scaling them by $\mu/2$. As we see form **Theorem 1.2**, however, the approximate minimizers $\mathbf{x}^k$ of $Q(\mathbf{x}, \mu_k)$ do not quite satisfy the feasibility conditions $c_i(\mathbf{x}) = 0$, $i \in E$ instead, they are perturbed (see 1.6) so that

$$c_i(\mathbf{x}^k) \approx -\lambda_i^*/\mu_k, \quad \text{for all } i \in E. \tag{1.9}$$

To be sure, we have $c_i(\mathbf{x}^k) \to 0$ as $\mu_k \to \infty$, but one may ask whether we can alter the function $Q(x, \mu_k)$ to avoid this systematic perturbation−that is, to make the approximate minimizers more nearly satisfy the equality constraints $c_i(\mathbf{x}) = 0$, even for moderate values of μ_k.

The augmented Lagrangian function $L_A(\mathbf{x}, \lambda, \mu)$ achieves this goal by including an explicit estimate of the Lagrange multiplier λ, based on the estimate (1.9), in the objective. The mathematical expression of the

augmented Lagrangian function is

$$L_A(\mathbf{x}, \lambda, \mu) = f(\mathbf{x}) - \sum_{i \in E} \lambda_i c_i(\mathbf{x}) + \frac{\mu}{2} \sum_{i \in E} c_i^2(\mathbf{x}). \tag{1.10}$$

Note that the augmented Lagrangian differs from the standard Lagrangian by the presence of the squared terms. In contrast, it differs from the quadratic penalty function by the presence of the summation term involving λ. In this sense, it is a combination of the Lagrangian function and the quadratic penalty function.

We now design an algorithm that fixes the penalty parameter μ to some value $\mu_k > 0$ at its $k - th$ iteration, fixes λ at the current estimate λ^k and preform minimization with respect to $\mathbf{x}$. Using $\mathbf{x}^k$ to denote the approximate minimizer of $L_A(\mathbf{x}, \lambda, \mu_k)$, we have by the optimality conditions for unconstrained minimization that

$$\mathbf{0} \approx \nabla_x L_A(\mathbf{x}, \lambda^k, \mu) = \nabla f(\mathbf{x}^k) - \sum_{i \in E} [\lambda_i^k - \mu_k c_i(\mathbf{x})] \nabla c_i(\mathbf{x}^k). \tag{1.11}$$

By comparing this expression with the first order optimality condition of Lagrangian function, we can deduce that

$$\lambda_i^* \approx \lambda_i - \mu_k c_i(\mathbf{x}^k), \quad \text{for all } i \in E. \tag{1.12}$$

By rearranging this expression, we have that

$$c_i(\mathbf{x}^k) \approx -\frac{1}{\mu_k}(\lambda_i^* - \lambda_i^k), \quad \text{for all } i \in E,$$

so we conclude that if λ^k is close to the optimal multiplier vector λ^*, the infeasibility in $\mathbf{x}^k$ will be much smaller that $1/\mu_k$, rather that being proportional to $1/\mu_k$ as in (1.9). The relation (1.12) immediately suggest a formula for improving our current estimate λ^k of the Lagrange multiplier vector, using the approximate minimizer $\mathbf{x}^k$ just calculated: We can set

$$\lambda_i^{k+1} = \lambda_i^k - \mu_k c_i(\mathbf{x}^k), \quad \text{for all } i \in E. \tag{1.13}$$

This discussion motivates the following algorithmic framework.

Algorithm 1.2 Augmeted Lagrangian Method-Equality Constraints

Choose $\mu_0 > 0$, tolerance $\tau_0 > 0$, and set the starting point $\mathbf{x}_s^0$ and λ^0.
for $k = 0, 1, \ldots$ **do**
    Find an approximate minimizer $\mathbf{x}^k$ of $L_A(\mathbf{x}, \lambda^k, \mu_k)$, starting at $\mathbf{x}_s^k$,
        and terminate when $\|\nabla_x L_A(\mathbf{x}, \lambda^k, \mu_k)\| \leq \tau_k$;
    if final convergence test is satisfied **then**
        Stop with approximate solution x_k;
    end if
    Update the Lagrange multiplier using (1.13) to obtain λ^{k+1}
    Choose a new penalty parameter $\mu_{k+1} > \mu_k$;
    Choose new starting point $\mathbf{x}_s^{k+1}$;
    Select the tolerance τ_{k+1}.
end for

Convergence of this method can be assured without increasing μ indefinitely. Ill conditioning is therefore less of a problem than in the quadratic penalty function, so the choice of starting point $\mathbf{x}_s^{k+1}$ in Algorithm 1.2 is less critical. We can simply start the search at the new iteration from the previous approximate

minimizer $\mathbf{x}^k$. The tolerance τ_k could be chosen to depend on the infeasibility $\sum_{i\in E}|c_i(\mathbf{x}^k)|$, and the penalty parameter μ may be increased if the reduction in this infeasibility measure is insufficient at the present iteration.

Properties of the Augmented Lagrangian

We state convergence theorems for the augmented Lagrangian method, and the results justify using the augmented Lagrangian function for solving equality constrained problems. The first results validate the approach of **Algorithm 1.2** by showing that when we have knowledge of the exact Lagrange multiplier λ^*, the solution $\mathbf{x}^*$ of the equality constrained optimization problem (1.4) is a strict minimizer of $L_A(\mathbf{x}, \lambda^*, \mu)$ for all μ sufficient large. Although we do not know λ^* exactly in practice, the result suggest that we can obtain a good estimate of $\mathbf{x}^*$ by minimizing $L_A(\mathbf{x}, \lambda, \mu)$ even then μ is not particularly large, provided that λ is a reasonably good estimate of λ^*.

Theorem 1.3. *Let $\mathbf{x}^*$ be a local solution of (1.4) at which the LICQ is satisfied, and the second order sufficient conditions are satisfied for $\lambda = \lambda^*$. Then there is a threshold value $\bar{\mu}$ such that for all $\mu > \bar{\mu}$, $\mathbf{x}^*$ is a strict local minimizer of $L_A(\mathbf{x}, \lambda^*, \mu)$.*

The second result, describes the more realistic situation of $\lambda \neq \lambda^*$. It gives conditions under which there is a minimizer of $L_A(\mathbf{x}, \lambda, \mu)$ that lies close to $\mathbf{x}^*$ and gives error bounds on both $\mathbf{x}^k$ and the updated multiplier estimate λ^{k+1} obtained form solving the subproblem at iteration k.

Theorem 1.4. *Suppose that the assumptions of **Theorem 1.3** are satisfied at $\mathbf{x}^*$ and λ^* and let $\bar{\mu}$ be chosen as in that theorem. Then there exist positive scalars δ, ε, and M such that the following claims hold:*

(a) For all λ^k and μ_k satisfying

$$\|\lambda^k - \lambda\| \leq \mu_k \delta, \quad \mu_k \geq \bar{\mu} \tag{1.14}$$

them the problem

$$\min_{\mathbf{x}} L_A(\mathbf{x}, \lambda^k, \mu_k) \quad \textit{subject to } \|\mathbf{x} - \mathbf{x}^*\| \leq \varepsilon$$

has a unique solutions $\mathbf{x}^k$. Moreover, we have

$$\|\mathbf{x} - \mathbf{x}^*\| \leq M\|\lambda^k - \lambda^*\|/\mu. \tag{1.15}$$

(b) For all λ^k and μ_k that satisfy (1.14), we have

$$\|\lambda^{k+1} - \lambda^*\| \leq M\|\lambda^k - \lambda^*\|/\mu. \tag{1.16}$$

(c) For all λ^k and μ_k that satisfy, the matrix $\nabla^2_{xx} L_A(\mathbf{x}^k, \lambda^k, \mu_k)$ is positive definite and the constraint gradients $\nabla c_i(\mathbf{x}^k)$, $i \in E$, are linearly independent.

This theorem illustrates some salient properties of the augmented Lagrangian approach. The bound (1.15) show that $\mathbf{x}^k$ will be close to $\mathbf{x}^*$ if λ^k is accurate or if the penalty parameter μ_k is large. Hence this approach gives us two ways to improving the accuracy of $\mathbf{x}^k$, whereas the quadratic penalty approach gives us only one option: increase μ_k. The bound (1.16) state that, locally, we can ensure an improvement in the accuracy of the multipliers by choosing a sufficiently large value μ_k. The final observation of the theorem shows that second-order sufficient conditions for unconstrained minimization are also satisfied for the k-th subproblem under the given conditions, so one can expect a good performance by applying standard unconstrained minimization techniques.

PROBLEM SET I

1. Consider the following constrained optimization problem

$$\min -5x_1^2 + x_2^2, \quad \text{subject to } x_1 = 1.$$

Draw the contour lines of the quadratic penalty function Q corresponding to $\mu = 1$. Find the stationary points of Q.

2. Consider the constrained minimization problem

$$\min x_1 + x_2, \quad \text{subject to } x_1^2 + x_2^2 = 0.$$

Construct the quadratic penalty function for this problem. Minimize the quadratic penalty function for $\mu_k = 1, 10, 100, 100$ using an unconstrained minimization algorithm. Set $\tau_k = \frac{1}{\mu_k}$ in **Algorithm 1.1**, and chose the staring point $\mathbf{x}_s^{k+1}$ for each minimization to be the solution for the previous value of the penalty parameter. Report the approximate solution of each penalty function.

3. Consider the minimization problem

$$\min \frac{1}{2}\left[x_1^2 + \frac{1}{3}x_2^2\right] \quad \text{subject to } x_1 + x_2 = 1.$$

The optimal solution for this problem is $\mathbf{x}^* = [0.25, 0.75]$, and the corresponding Lagrange multiplier is $\lambda^* = -0.25$. Compare the quadratic penalty method and augmented Lagrange method for solving this problem. Use the initial Lagrange multiplier $\lambda_0 = 0$, experiment with the following sequence of the penalty parameter (1) $\mu_k = 0.1 \times 2^k$, (2) $\mu_k = 0.1 \times 4^k$, and (3) $\mu_k = 0.1 \times 8^k$. Compare and comment on the results.

4. Consider the problem

$$\min x_1 x_2 \quad \text{subject to } x_1 - 2x_2 = 3.$$

   (a) For which values of the penalty parameter does the quadratic penalty function have a minimum? Calculate the minimum point as a function of the penalty parameter.

   (b) For sufficient large value of the penalty parameter, analyse the operation of the augmented Lagrangian method from $\lambda = 0$.