Constrained optimization techniques for image reconstruction
by Johnathan Matheas Bardsley

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Mathematics
Montana State University
© Copyright by Johnathan Matheas Bardsley (2002)

Abstract:
The goal of this thesis is the solution of constrained minimization problems that originate in applications from image reconstruction. These applications include microscopy, medical imaging and, our particular application, astronomical imaging. Constraints arise from the fact that the objects we are imaging are photon densities, or light intensities, and are therefore nonnegative. These minimization problems are ill-conditioned, and, since high resolution images are desired, are large-scale. Consequently, efficient numerical techniques are required. We minimize both quadratic and strictly convex functions. Existing algorithms are implemented for the quadratic minimization problem, and ideas from these algorithms are extended to the problem of minimizing the convex function. A detailed analysis and numerical study of these algorithms is presented.

CONSTRAINED OPTIMIZATION TECHNIQUES

FOR IMAGE RECONSTRUCTION

by

Johnathan Matheas Bardsley

A dissertation submitted in partial fulfillment
of the requirements for the degree

of

Doctor of Philosophy

in

Mathematics

MONTANA STATE UNIVERSITY
Bozeman, Montana

April 2002

D378
B2368

APPROVAL

of a dissertation submitted by

Johnathan Matheas Bardsley

This dissertation has been read by each member of the dissertation committee and has been found to be satisfactory regarding content, English usage, format, citations, bibliographic style, and consistency, and is ready for submission to the College of Graduate Studies.

Curtis R. Vogel Curtis R. Vogel (Signature) 4/22/02 Date

Approved for the Department of Mathematics

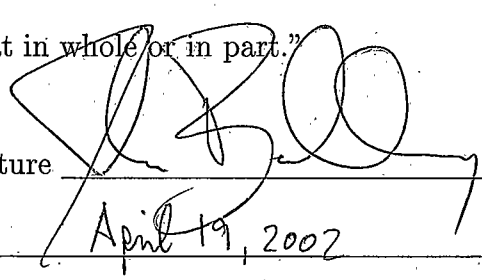
Kenneth L. Bowers Kenneth L. Bowers (Signature) 4/19/02 Date

Approved for the College of Graduate Studies

Bruce McLeod Bruce R. McLeod (Signature) 4-22-02 Date

STATEMENT OF PERMISSION TO USE

In presenting this dissertation in partial fulfillment of the requirements for a doctoral degree at Montana State University, I agree that the Library shall make it available to borrowers under rules of the Library. I further agree that copying of this dissertation is allowable only for scholarly purposes, consistent with "fair use" as prescribed in the U. S. Copyright Law. Requests for extensive copying or reproduction of this dissertation should be referred to Bell & Howell Information and Learning, 300 North Zeeb Road, Ann Arbor, Michigan 48106, to whom I have granted "the exclusive right to reproduce and distribute my dissertation in and from microform along with the non-exclusive right to reproduce and distribute my abstract in any format in whole or in part."

Signature 

Date April 19, 2002

This thesis is dedicated
to the memory of my father Scott Leavitt Bardsley, 1950-1999,
to my son Alex and daughter Ellie,
and to my wife Jennifer.

ACKNOWLEDGEMENTS

I would first like to thank Dr. Curtis R. Vogel, who has been an excellent thesis advisor in every conceivable way. Without his substantial commitment of time and energy, this thesis would not have been possible. I would also like to thank all of the other teachers, professors, and mentors who have made a difference in my life along the way.

I would like to thank my family and friends. In particular, I would like to thank my mom and dad for their endless encouragement and belief in me, my children, with whom I have found countless hours of play and unconditional love, and, most of all, my wife, whose love, support and belief in me has been a great source of strength, and whose investment of energy towards the successful completion of my PhD has been at least as great as my own.

Finally, I would like to thank James Jacklitch, who provided the LaTex template for this thesis.

TABLE OF CONTENTS

LIST OF FIGURES

ABSTRACT

The goal of this thesis is the solution of constrained minimization problems that originate in applications from image reconstruction. These applications include microscopy, medical imaging and, our particular application, astronomical imaging. Constraints arise from the fact that the objects we are imaging are photon densities, or light intensities, and are therefore nonnegative. These minimization problems are ill-conditioned, and, since high resolution images are desired, are large-scale. Consequently, efficient numerical techniques are required. We minimize both quadratic and strictly convex functions. Existing algorithms are implemented for the quadratic minimization problem, and ideas from these algorithms are extended to the problem of minimizing the convex function. A detailed analysis and numerical study of these algorithms is presented.

CHAPTER 1

INTRODUCTION

The goal of this thesis is the solution of constrained minimization problems that originate in applications from image reconstruction. These applications include microscopy, medical imaging and, our particular application, atmospheric imaging.

We want to find a vector $\mathbf{x}$ in a constraint set $\Omega \subset \mathbb{R}^N$ that minimizes a function J mapping $\mathbb{R}^N$ to $\mathbb{R}$. We express this problem mathematically as

$$\min_{\mathbf{x}\in\Omega} J(\mathbf{x}). \tag{1.1}$$

Problem 1.1 is referred to as a constrained minimization problem. We restrict our attention to a class of functions J arising in two-dimensional image reconstruction.

Two Dimensional Image Formation

Our particular imaging application arises in atmospheric optics. As light propagates through the atmosphere, light rays are bent due to random variations in the index of diffraction caused by atmospheric turbulence. This causes blurring of images of an astronomical object captured with a ground-based telescope [38].

A model for data obtained from a linear, translation-invariant optical imaging system is

$$d_{i,j} = \int\int s(x_i - x, y_j - y) f_{\text{true}}(x, y) dx dy + \eta_{i,j}, \tag{1.2}$$

$i = 1, \ldots, n_x$, $j = 1, \ldots, n_y$. Here f_{true} denotes the true image, or object. f_{true} is an energy density, or photon density, and hence, is nonnegative. s is called the point spread function (PSF) and is the response of the imaging system to a point light source. In our application, s quantifies the blurring effects of the atmosphere and is nonnegative. d is called the (discrete, noisy, blurred) image. η represents noise in the collection of the image d.

For computational purposes, we discretize the integral in (1.2) to obtain

$$d_{i,j} = \sum_{i'} \sum_{j'} s_{i-i',j-j'} [f_{\text{true}}]_{i',j'} + \eta_{i,j}. \tag{1.3}$$

With lexicographical ordering [42, p. 81] of index pairs (i, j), this yields the model

$$\mathbf{d} = S\mathbf{f}_{\text{true}} + \boldsymbol{\eta}, \tag{1.4}$$

where $\mathbf{d}$, $\mathbf{f}_{\text{true}}$, and $\boldsymbol{\eta}$ are $(n_x \cdot n_y) \times 1$ vectors, and S is the $(n_x \cdot n_y) \times (n_x \cdot n_y)$ blurring matrix. Our goal is accurate, efficient estimation of $\mathbf{f}_{\text{true}}$.

Since s is nonnegative,

$$S\mathbf{f} \geq \mathbf{0} \quad \text{for all} \quad \mathbf{f} \geq \mathbf{0}, \tag{1.5}$$

where $\mathbf{f} \geq \mathbf{0}$ means $f_i \geq 0$ for all i. S is block Toeplitz with Toeplitz blocks (BTTB) [42, Chapter 5]. In most imaging applications, S is highly ill-conditioned. Furthermore, for high resolution images, S is large, e.g., a 1024×1024 pixel image gives 1,048,576 unknowns, in which case S has approximately 10^{12} entries.

Due to the ill-conditioning of the matrix S, the pseudo inverse [42] estimate $S^{\dagger}\mathbf{d}$ for $\mathbf{f}_{\text{true}}$ is unstable with respect to perturbations in the data vector $\mathbf{d}$. We therefore

use a stabilization technique known as Tikhonov regularization [42]. Computing the regularized estimate for $\mathbf{f}_{\text{true}}$ in (1.4) requires the minimization of a function of the form

$$J_\alpha(\mathbf{f}) = \ell(S\mathbf{f}; \mathbf{d}) + \frac{\alpha}{2}\mathbf{f}^T L\mathbf{f}. \tag{1.6}$$

We incorporate the knowledge that $\mathbf{f}_{\text{true}}$ is nonnegative by minimizing J_α subject to the constraint $\mathbf{f} \geq \mathbf{0}$.

The quadratic $\frac{1}{2}\mathbf{f}^T L\mathbf{f}$, where L is a symmetric positive semidefinite matrix, is known as the regularization functional. In addition to having a stabilizing effect on the minimization of (1.6), the regularization term allows for the incorporation of prior information about the true image f_{true}. For example, if we know that f_{true} is smooth, taking L to be the negative of the discrete Laplacian operator [42, p. 84] will penalize vectors $\mathbf{f}$ with large discretized gradients. Another common choice of L is the identity matrix I, which yields the discretized L^2 regularization functional $\frac{1}{2}\mathbf{f}^T\mathbf{f}$.

The nonnegative real number α is known as the regularization parameter. Typically, as $\alpha \to 0$, the minimization of J_α becomes increasingly ill-conditioned. On the other hand, if α becomes too large, the regularization term will dominate in (1.6), and the minimization of J_α, though stable, will result in poor estimations of $\mathbf{f}_{\text{true}}$. The choice of the parameter α is therefore an important step in determining a suitable J_α. See [42, Chapter 7].

ℓ is called the fit-to-data function and allows for the incorporation of prior information about the statistics of the noise vector $\boldsymbol{\eta}$. In atmospheric imaging applications

[38], each component of $\mathbf{d}$ in (1.4) is modeled as a realization of a random variable

$$d_i \sim Poisson\left([S\mathbf{f}_{\text{true}}]_i\right) + Normal(0, \sigma^2). \tag{1.7}$$

By this we mean that d_i is the sum of a Poisson random variable with mean and variance $[S\mathbf{f}_{\text{true}}]_i$ and a normally distributed, or Gaussian, random variable with mean 0 and variance σ^2 [2]. We assume that these two random variables are independent. We also assume that d_i is independent of d_j for all $i \neq j$. In this setting, the components of the noise vector $\boldsymbol{\eta}$ in (1.4) are realizations of random variables with components

$$\eta_i = d_i - [S\mathbf{f}_{\text{true}}]_i.$$

We will consider two choices of the function ℓ. The first is appropriate if the Gaussian noise term in (1.7) is dominant, and the variance σ^2 is constant with respect to i. In this case we take

$$\ell(S\mathbf{f}; \mathbf{d}) = ||S\mathbf{f} - \mathbf{d}||_2^2. \tag{1.8}$$

This is proportional to the log likelihood function [42] for the model (1.4) when the components of $\mathbf{d}$ are independent and identically distributed Gaussian random variables with mean 0 and variance σ^2. We will refer to this choice of ℓ as the least squares fit-to-data function. The resulting J_α is quadratic.

If, on the other hand, the Poisson noise term in (1.7) dominates, we can incorporate this information by choosing ℓ to be the log likelihood function for the model (1.4)

when the components of $\mathbf{d}$ are independent Poisson random variables with Poisson parameter $[S\mathbf{f}_{\text{true}}]_i$, which is given by

$$\ell(S\mathbf{f};\mathbf{d}) = \sum_i [S\mathbf{f}]_i - \sum_i d_i \log([S\mathbf{f}]_i). \tag{1.9}$$

Unfortunately, this choice of ℓ is undefined for $\mathbf{f} \geq \mathbf{0}$ such that $[S\mathbf{f}]_i = 0$ for some index i. To overcome this difficulty, we replace (1.9) by

$$\ell(S\mathbf{f};\mathbf{d}) = \sum_i ([S\mathbf{f}]_i + \sigma^2) - \sum_i d_i \log([S\mathbf{f}]_i + \sigma^2), \tag{1.10}$$

where σ^2 is the variance of the Gaussian random variable in equation (1.7). We can statistically justify this choice of ℓ by noting that it is the log likelihood function for the model (1.4) when each component of $\mathbf{d}$ is modeled as a realization of a random variable

$$d_i \sim Poisson\left([S\mathbf{f}_{\text{true}}]_i + \sigma^2\right), \tag{1.11}$$

which has the same variance as the random variable d_i given in equation (1.7). We replace our image vector $\mathbf{d}$ in (1.10) by $\max(\mathbf{d}, \mathbf{0})$. This is consistent with (1.11), since any realization of a Poisson random variable is nonnegative, and has two important consequences. First, it guarantees that ℓ, and hence, the J_α resulting from (1.6), is convex on $\{\mathbf{f} \in \mathbb{R}^N \mid \mathbf{f} \geq \mathbf{0}\}$. Strict convexity for J_α follows if the null spaces of L and S intersect only at $\{\mathbf{0}\}$. Secondly, J_α is Lipschitz continuous on $\{\mathbf{f} \in \mathbb{R}^N \mid \mathbf{f} \geq \mathbf{0}\}$.

The minimization problems we wish to solve are then expressed as

$$\min_{\mathbf{f} \geq \mathbf{0}} J_\alpha(\mathbf{f}), \tag{1.12}$$

where ℓ in (1.6) is given either by (1.8) or (1.10). This problem is in the form of (1.1) with J replaced by J_α, $\mathbf{x}$ replaced by $\mathbf{f}$, and $\Omega = \{\mathbf{f} \in \mathbb{R}^N \mid \mathbf{f} \geq \mathbf{0}\}$.

The Test Problem with Unknown Phase

In certain atmospheric optics applications, the PSF s in (1.2) is unknown, but can be parameterized in terms of a function known as the phase, which we will denote by ϕ. The phase, or wavefront aberration, quantifies the deviation from planarity of the wavefront of light that has propagated through the atmosphere. We assume that the dependence of the PSF s on ϕ is given by

$$s[\phi] = |\mathcal{F}^{-1}\{pe^{i\phi}\}|^2, \tag{1.13}$$

where $\mathcal{F}$ denotes the 2-D Fourier transform, $i = \sqrt{-1}$, and p is the pupil, or aperture, function. See [38] for conditions which justify this assumption. We further assume

$$p(x,y) = \begin{cases} 1, & (x,y) \in A, \\ 0, & \text{otherwise}, \end{cases}$$

where the region A represents the aperture and is an annulus in the case of imaging with large astronomical telescopes. Discretizing the integral in (1.2) and taking into account that corresponding to each image $\mathbf{d}$ there is a (discretized) true phase $\boldsymbol{\phi}_{\text{true}}$, we obtain the matrix-vector equation

$$\mathbf{d} = S[\boldsymbol{\phi}_{\text{true}}]\mathbf{f}_{\text{true}} + \boldsymbol{\eta}. \tag{1.14}$$

This reduces to (1.4) for the case when $\boldsymbol{\phi}_{\text{true}}$, and hence, s, is known.

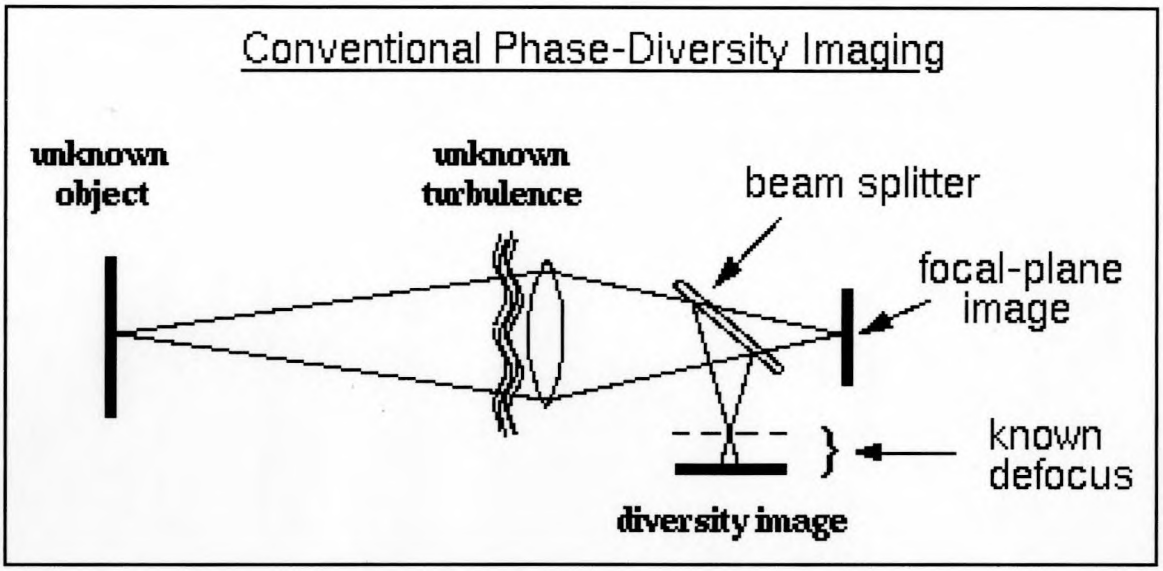


Figure 1. A Conventional Phase Diversity Imaging Setup.

When the vectors $\boldsymbol{\phi}_{\text{true}}$ and $\mathbf{f}_{\text{true}}$ are both unknown then (1.14) is underdetermined. We overcome this difficulty by applying a technique called phase diversity [22]. A second image with a known phase perturbation $\boldsymbol{\theta}$ is collected and is modeled by

$$\mathbf{d}' = S[\boldsymbol{\phi}_{\text{true}} + \boldsymbol{\theta}]\mathbf{f}_{\text{true}} + \boldsymbol{\eta}'. \tag{1.15}$$

See Figure 1 for an illustration. Using least squares fit-to-data functions (see equation (1.8)), and quadratic regularization terms for both the object and the phase we get

$$J_{\alpha,\gamma}(\mathbf{f}, \boldsymbol{\phi}) = \frac{1}{2}||S[\boldsymbol{\phi}]\mathbf{f} - \mathbf{d}||^2 + \frac{1}{2}||S[\boldsymbol{\phi} + \boldsymbol{\theta}]\mathbf{f} - \mathbf{d}'||^2 + \frac{\alpha}{2}\mathbf{f}^T L\mathbf{f} + \frac{\gamma}{2}\boldsymbol{\phi}^T M\boldsymbol{\phi}. \tag{1.16}$$

As in (1.6), L is symmetric positive semidefinite and incorporates prior smoothness information about $\mathbf{f}$. M is also symmetric positive semidefinite and incorporates prior smoothness information about $\boldsymbol{\phi}$. α and γ are positive regularization parameters.

In (1.16), $\boldsymbol{\phi}$ is not a constrained variable. We therefore minimize $J_{\alpha,\gamma}$ over

$$\Omega = \left\{(\mathbf{f}, \boldsymbol{\phi}) \in \mathbb{R}^{2N} \mid \mathbf{f} \geq \mathbf{0}\right\}.$$

The associated minimization problem is then expressed as

$$\min_{(\mathbf{f},\boldsymbol{\phi})\in\Omega} J_{\alpha,\gamma}(\mathbf{f}, \boldsymbol{\phi}). \tag{1.17}$$

This minimization problem is of the form of problem (1.1), with J replaced by $J_{\alpha,\gamma}$, $\mathbf{x}$ replaced by $(\mathbf{f}, \boldsymbol{\phi})$, and $\mathbb{R}^N$ replaced by $\mathbb{R}^{2N}$. If one ignores the nonnegativity constraint on the vector $\mathbf{f}$, (1.17) becomes an unconstrained minimization problem. In [43, 19] this problem is explored.

Unfortunately, the dependence of S on $\boldsymbol{\phi}$ is nonlinear. This results in a nonconvex function $J_{\alpha,\gamma}$. Furthermore, there are twice as many unknowns with the addition of the unknown phase. Consequently, problem (1.17), which we will not attempt to solve in this thesis, is much more difficult to solve than is (1.12), and requires more robust algorithms.

Another important consideration in constrained minimization is degeneracy. This property is discussed in Chapter 2. The solutions of the minimization problems which we will consider are not degenerate, but they are very nearly degenerate. This further increases the difficulty of solving these problems.

Outline of Thesis

Each of the minimization problems presented above is known as a bound constrained minimization problem. Such problems have been studied extensively and present themselves in a large number of applications [32, 23]. Various techniques have been developed for solving these problems.

The gradient projection method for minimizing a continuously differentiable function $J : \mathbb{R}^N \to \mathbb{R}$ over a closed subset Ω of $\mathbb{R}^N$ was introduced in 1964 by Goldstein [20, 21] and independently one year later by Levitin and Polyak [25]. Seminal work was done in the study of the convergence properties of this algorithm by Bertsekas [5]. This work was extended by Calamai and Moré [10]. Unfortunately, the gradient projection algorithm is very slow to converge for ill-conditioned problems.

Active set methods attempt to incorporate Hessian information via the reduced Hessian (see Chapter 2). An active set Newton algorithm for nonlinear bound constrained minimization was developed by Bertsekas [6]. Active set strategies have yielded many successful algorithms for bound constrained quadratic minimization [18]. Unfortunately, these algorithms tend not to perform well for large-scale, ill-conditioned bound constrained problems.

Several of the most effective and robust algorithms for bound constrained minimization can be viewed as two stage algorithms, which combine a constraint identification stage with a subspace minimization stage. Such algorithms include a bound

constrained variant of the limited memory BFGS algorithm [32] known as LBFGS-B [9, 45] and the bound constrained Newton trust region algorithm TRON [26] for general nonlinear functions, and the gradient projection conjugate gradient (GPCG) algorithm [30] and Friedlander and Martinez's GPCG modification [14] for quadratic functions.

Contrasting the previous three approaches, interior point algorithms create a sequence of approximations to a solution of a constrained minimization problem which lie in the interior of the constraint region. Examples of interior point methods include Byrd, Hribar, and Nocedal's KNITRO [8] and MATLAB's large-scale bound constrained minimizer, which is based on the work of Coleman and Li [11, 12].

In Chapter 2 of this thesis we present mathematical preliminaries. These will include a general theory for lower bound constrained minimization, including the necessary and sufficient conditions for a solution to (1.1). We will also present the gradient projection algorithm of Bertsekas [5] and the corresponding convergence theory. We will present a general two stage algorithm of the form of that found in Calamai and Moré [10] and prove that convergence for this algorithm is guaranteed for a strictly convex function. We finish the chapter with a presentation of three unconstrained minimization algorithms: the conjugate gradient (CG) algorithm for quadratic minimization, and the limited memory BFGS and Newton-CG-trust region algorithms for the minimization of general nonlinear functions.

In Chapter 3 we present in detail two algorithms for lower bound constrained quadratic programming. The first is based on the work of Moré and Toreldo [30], while the second is based on the work of Friedlander and Martinez [14]. We conclude the chapter with a numerical implementation and comparison of these two methods on four different imaging test problems.

In Chapter 4 we extend the results of Chapters 2 and 3 to convex lower bound constrained minimization. We present an algorithm for the solution of this problem. Convergence follows from results is Chapter 2. We conclude the chapter with a numerical implementation and comparison of methods on several imaging test problems.

CHAPTER 2

MATHEMATICAL PRELIMINARIES

In this chapter we introduce notation and develop the mathematical tools that will be used in the sequel.

Preliminary Notation and Definitions

We will denote a vector in $\mathbb{R}^N$ by $\mathbf{x} = (x_1, \ldots, x_N)$. The i^{th} standard unit vector $\mathbf{e}_i \in \mathbb{R}^N$ is defined by $[\mathbf{e}_i]_j = \delta_{ij}$, where $\delta_{ij} = 1$ if $i = j$ and is zero otherwise. The Euclidean inner product of $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^N$ is defined by $\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^{N} x_i y_i$, and the associated Euclidean norm of $\mathbf{x}$ is then given by $||\mathbf{x}|| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$.

$\mathbb{R}^{M \times N}$ denotes the set of $M \times N$ matrices with real-valued entries. A matrix $A \in \mathbb{R}^{N \times N}$ is symmetric if $A = A^T$. A is positive definite if for every $\mathbf{p} \neq \mathbf{0}$ in $\mathbb{R}^N$, $\langle A\mathbf{p}, \mathbf{p} \rangle > 0$, while A is positive semidefinite if $\langle A\mathbf{p}, \mathbf{p} \rangle \geq 0$ for all $\mathbf{p}$ in $\mathbb{R}^N$. In the sequel we will use the acronym SPD for symmetric and positive definite. For $A \in \mathbb{R}^{M \times N}$, the spectral norm of A is given by

$$||A|| = \sqrt{\lambda_{max}(A^T A)},$$

where $\lambda_{max}(B)$ denotes the absolute value of the largest eigenvalue of the symmetric matrix B. This is also known as the ℓ^2 operator norm (see [24]).

By $J : \mathbb{R}^N \to \mathbb{R}$ we mean that J is a real valued function of N variables, and, unless otherwise stated, we will assume that J is smooth, i.e. J is sufficiently differentiable.

We define the gradient of J at $\mathbf{x}$ by

$$\text{grad } J(\mathbf{x}) = \left(\frac{\partial J}{\partial x_1}(\mathbf{x}), \ldots, \frac{\partial J}{\partial x_N}(\mathbf{x}) \right)^T.$$

The Hessian of J at $\mathbf{x}$, denoted Hess $J(\mathbf{x})$, is the $N \times N$ matrix with entries

$$[\text{Hess } J(\mathbf{x})]_{ij} = \frac{\partial^2 J}{\partial x_i \partial x_j}(\mathbf{x}).$$

Definition 2.1. If $\mathbf{h} : \mathbb{R}^N \to \mathbb{R}^M$ then $\mathbf{h}$ is Lipschitz continuous if there exists a positive real number γ such that

$$||\mathbf{h}(\mathbf{x}) - \mathbf{h}(\mathbf{y})|| \leq \gamma ||\mathbf{x} - \mathbf{y}||$$

for every $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^N$.

Definition 2.2. Let $J : \mathbb{R}^N \to \mathbb{R}$. J is convex if

$$J(\lambda \mathbf{x} + (1 - \lambda)\mathbf{y}) \leq \lambda J(\mathbf{x}) + (1 - \lambda) J(\mathbf{y}) \tag{2.1}$$

whenever $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$ and $\lambda \in (0, 1)$. If the inequality in (2.1) is strict whenever $\mathbf{x} \neq \mathbf{y}$ then J is said to be strictly convex.

We will now concentrate on the problem of finding a minimizer $\mathbf{x}^*$ of a function $J : \mathbb{R}^N \to \mathbb{R}$ subject to the constraint that $\mathbf{x}^* \in \Omega \subset \mathbb{R}^N$.

Definition 2.3. Let Ω be a proper subset of $\mathbb{R}^N$. Ω will be known as the feasible set, and any $\mathbf{x} \in \Omega$ will be known as a feasible point.

Definition 2.4. $\mathbf{x}^* \in \Omega$ is a local constrained minimizer for J provided that there exists a $\delta > 0$ such that

$$J(\mathbf{x}^*) \leq J(\mathbf{x}) \quad \text{whenever } \mathbf{x} \in \Omega \text{ and } ||\mathbf{x} - \mathbf{x}^*|| < \delta. \tag{2.2}$$

If (2.2) holds for $\delta = \infty$ then $\mathbf{x}^*$ is a global constrained minimizer for J. $\mathbf{x}^*$ is a strict local constrained minimizer for J if the inequality in (2.2) is strict whenever $\mathbf{x} \neq \mathbf{x}^*$.

Remark 2.5. We indicate (2.2) by

$$\mathbf{x}^* = \arg\min_{\mathbf{x} \in \Omega} J(\mathbf{x}).$$

Such an $\mathbf{x}^*$ is then said to be a solution to the problem

$$\min_{\mathbf{x} \in \Omega} J(\mathbf{x}). \tag{2.3}$$

We now introduce the notion of a convex set, and characterize smooth, convex functions defined on convex sets.

Definition 2.6. $\mathcal{C} \subset \mathbb{R}^N$ is convex if $\lambda\mathbf{x} + (1 - \lambda)\mathbf{y} \in \mathcal{C}$ whenever $\mathbf{x}, \mathbf{y} \in \mathcal{C}$ and $0 \leq \lambda \leq 1$.

The following theorem can be found in Luenberger [27].

Theorem 2.7. Let J be a C^1 function defined on a convex set Ω. Then J is convex if and only if for all $\mathbf{x}, \mathbf{y} \in \Omega$

$$J(\mathbf{y}) \geq J(\mathbf{x}) + \langle \text{grad } J(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle . \tag{2.4}$$

PROOF. First suppose J is convex. Then for all $\alpha \in [0,1]$,

$$J(\alpha\mathbf{y} + (1-\alpha)\mathbf{x}) \leq \alpha J(\mathbf{y}) + (1-\alpha)J(\mathbf{x}).$$

This implies that

$$J(\mathbf{y}) \geq J(\mathbf{x}) + \frac{J(\mathbf{x} + \alpha(\mathbf{y}-\mathbf{x})) - J(\mathbf{x})}{\alpha}.$$

Letting $\alpha \to 0$ yields (2.4).

Now suppose (2.4) holds. Fix $\mathbf{x}_1$, $\mathbf{x}_2 \in \Omega$ and $\alpha \in [0,1]$. Setting $\mathbf{x} = \alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2$ and alternatively $\mathbf{y} = \mathbf{x}_1$ and $\mathbf{y} = \mathbf{x}_2$, we get

$$J(\mathbf{x}_1) \geq J(\mathbf{x}) + \langle \text{grad } J(\mathbf{x}), \mathbf{x}_1 - \mathbf{x} \rangle \tag{2.5}$$

$$J(\mathbf{x}_2) \geq J(\mathbf{x}) + \langle \text{grad } J(\mathbf{x}), \mathbf{x}_2 - \mathbf{x} \rangle . \tag{2.6}$$

Multiplying (2.5) by α and (2.6) by $(1-\alpha)$ and adding yields

$$\alpha J(\mathbf{x}_1) + (1-\alpha)J(\mathbf{x}_2) \geq J(\mathbf{x}) + \langle \text{grad } J(\mathbf{x}), \alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2 - \mathbf{x} \rangle .$$

Substituting $\mathbf{x} = \alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2$ then gives

$$\alpha J(\mathbf{x}_1) + (1-\alpha)J(\mathbf{x}_2) \geq J(\alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2),$$

and hence, J is convex. □

The notion of coercivity allows us to prove the existence of a global minimizer for a smooth function J. Strict convexity then guarantees uniqueness.

Definition 2.8. $J : \mathbb{R}^N \to \mathbb{R}$ is coercive if given any $L \in \mathbb{R}$ there exists an $M > 0$ such that if $||\mathbf{x}|| > M$ then $J(\mathbf{x}) > L$.

Theorem 2.9. Let $J : \mathbb{R}^N \to \mathbb{R}$ be smooth and coercive. Then there exists a global constrained minimizer $\mathbf{x}^*$ of J on a closed set Ω.

PROOF. Let $\mathbf{x}_0 \in \Omega$. By coercivity, there exists an $M > 0$ such that whenever $||\mathbf{x}|| > M$, $J(\mathbf{x}) > J(\mathbf{x}_0)$. Let $B = \{\mathbf{x} \in \mathbb{R}^N \mid ||\mathbf{x}|| \leq M\}$. Then $B \cap \Omega$ is nonempty and is closed and bounded, and hence compact. The continuity of J therefore implies that there exists $\mathbf{x}^* \in B \cap \Omega$ at which J attains its minimum on $B \cap \Omega$. Existence follows since $\arg\min_{\mathbf{x} \in \Omega} J(\mathbf{x}) = \arg\min_{\mathbf{x} \in B \cap \Omega} J(\mathbf{x})$. □

Theorem 2.10. If $J : \mathbb{R}^N \to \mathbb{R}$ is strictly convex and a global constrained minimizer exists for J on a convex set Ω, it is unique.

PROOF. Suppose that $\mathbf{x}^*$ is a global constrained minimizer for J on Ω. If $J(\hat{\mathbf{x}}) = J(\mathbf{x}^*)$ with $\hat{\mathbf{x}} \neq \mathbf{x}^*$ then $J(\frac{1}{2}\mathbf{x}^* + \frac{1}{2}\hat{\mathbf{x}}) < \frac{1}{2}J(\mathbf{x}^*) + \frac{1}{2}J(\hat{\mathbf{x}}) = J(\mathbf{x}^*)$, with $\frac{1}{2}\mathbf{x}^* + \frac{1}{2}\hat{\mathbf{x}} \in \Omega$, which is a contradiction. Hence $\mathbf{x}^*$ is unique. □

Definition 2.11. The projection of $\mathbf{x} \in \mathbb{R}^N$ onto a closed and convex set $\mathcal{C}$ is given by

$$\mathcal{P}_{\mathcal{C}}(\mathbf{x}) := \arg\min_{\mathbf{y} \in \mathcal{C}} ||\mathbf{y} - \mathbf{x}||^2.$$

In order to prove that $\mathcal{P}_{\mathcal{C}}$ is continuous we will need the following lemma, which will also prove useful in the sequel. We follow the exposition of Zarantonello [44].

Lemma 2.12.

$$\langle \mathbf{x} - \mathcal{P}_{\mathcal{C}}(\mathbf{x}), \mathcal{P}_{\mathcal{C}}(\mathbf{x}) - \mathbf{y} \rangle \geq 0 \quad \text{for all} \quad \mathbf{y} \in \mathcal{C}. \tag{2.7}$$

PROOF. By definition,

$$||\mathbf{x} - \mathcal{P}_{\mathcal{C}}(\mathbf{x})|| \leq ||\mathbf{x} - \mathbf{y}|| \quad \text{for all} \quad \mathbf{y} \in \mathcal{C}. \tag{2.8}$$

Consequently,

$$\begin{aligned} 0 &\leq ||\mathbf{x} - \mathbf{y}||^2 - ||\mathbf{x} - \mathcal{P}_{\mathcal{C}}(\mathbf{x})||^2 \\ &= ||(\mathbf{x} - \mathcal{P}_{\mathcal{C}}(\mathbf{x})) + (\mathcal{P}_{\mathcal{C}}(\mathbf{x}) - \mathbf{y})||^2 - ||\mathbf{x} - \mathcal{P}_{\mathcal{C}}(\mathbf{x})||^2 \\ &= 2\langle \mathbf{x} - \mathcal{P}_{\mathcal{C}}(\mathbf{x}), \mathcal{P}_{\mathcal{C}}(\mathbf{x}) - \mathbf{y} \rangle + ||\mathcal{P}_{\mathcal{C}}(\mathbf{x}) - \mathbf{y}||^2. \end{aligned}$$

Now replace $\mathbf{y}$ by $\mathbf{y}' = t\mathbf{y} + (1-t)\mathcal{P}_{\mathcal{C}}(\mathbf{x})$, where $t \in [0,1]$. Since $\mathbf{y}' \in \mathcal{C}$

$$0 \leq 2t\langle \mathbf{x} - \mathcal{P}_{\mathcal{C}}(\mathbf{x}), \mathcal{P}_{\mathcal{C}}(\mathbf{x}) - \mathbf{y} \rangle + t^2 ||\mathcal{P}_{\mathcal{C}}(\mathbf{x}) - \mathbf{y}||^2$$

for $t \in [0,1]$. By first dividing by t and then letting $t \to 0$, (2.7) follows. □

Theorem 2.13. If $\mathcal{C}$ is a nonempty, closed and convex set, then $\mathcal{P}_{\mathcal{C}} : \mathbb{R}^N \to \mathbb{R}^N$ is well-defined and continuous.

PROOF. Let $\mathbf{x}_0 \in \mathbb{R}^N$, and define $J : \mathbb{R}^N \to \mathbb{R}$ by $J(\mathbf{y}) = ||\mathbf{y} - \mathbf{x}_0||^2$. In order to prove that $\mathcal{P}_{\mathcal{C}}$ is well-defined, it suffices to show that there exists a unique $\mathbf{y}^* = \arg\min_{\mathbf{y} \in \mathcal{C}} J(\mathbf{y})$.

Fix $\mathbf{y}_0 \in \mathcal{C}$ not equal to $\mathbf{x}_0$. Then $\inf_{\mathbf{y}\in\mathcal{C}} J(\mathbf{y}) = \inf_{\mathbf{y}\in B} J(\mathbf{y})$, where $B = \mathcal{C} \cap \{\mathbf{y} \in \mathbb{R}^N \mid ||\mathbf{y} - \mathbf{x}_0|| < ||\mathbf{y}_0 - \mathbf{x}_0||\}$. B is nonempty, closed, and bounded and is therefore compact. The continuity of J therefore implies that there exists $\mathbf{y}^* = \arg\min_{\mathbf{y}\in\mathcal{C}} J(\mathbf{y})$. To prove uniqueness, suppose $\hat{\mathbf{y}} = \arg\min_{\mathbf{y}\in\mathcal{C}} J(\mathbf{y})$. Let $||\mathbf{x}_0 - \mathbf{y}^*|| = \eta = ||\mathbf{x}_0 - \hat{\mathbf{y}}||$. By the parallelogram inequality $||\mathbf{y}^* - \hat{\mathbf{y}}||^2 = ||(\mathbf{y}^* - \mathbf{x}_0) - (\hat{\mathbf{y}} - \mathbf{x}_0)||^2 = 2||\mathbf{y}^* - \mathbf{x}_0||^2 + 2||\hat{\mathbf{y}} - \mathbf{x}_0||^2 - ||(\mathbf{y}^* - \mathbf{x}_0) + (\hat{\mathbf{y}} - \mathbf{x}_0)||^2 = 4\eta^2 - 4||\frac{1}{2}(\mathbf{y}^* + \hat{\mathbf{y}}) - \mathbf{x}_0||^2$. But $\frac{1}{2}(\mathbf{y}^* + \hat{\mathbf{y}}) \in \mathcal{C}$, so $||\frac{1}{2}(\mathbf{y}^* + \hat{\mathbf{y}}) - \mathbf{x}_0||^2 \geq \eta^2$. Hence $||\mathbf{y}^* - \hat{\mathbf{y}}|| \leq 0$, implying $||\mathbf{y}^* - \hat{\mathbf{y}}|| = 0$, so $\mathbf{y}^* = \hat{\mathbf{y}}$.

We now prove that $\mathcal{P}_\mathcal{C}$ is continuous. Choose $\mathbf{x}$ and $\mathbf{x}'$ in $\mathbb{R}^N$. Replacing $\mathbf{y}$ in (2.7) by $\mathcal{P}_\mathcal{C}(\mathbf{x}')$ we get

$$\langle \mathbf{x} - \mathcal{P}_\mathcal{C}(\mathbf{x}), \mathcal{P}_\mathcal{C}(\mathbf{x}) - \mathcal{P}_\mathcal{C}(\mathbf{x}') \rangle \geq 0. \tag{2.9}$$

Interchanging $\mathbf{x}$ and $\mathbf{x}'$ in inequality (2.9) gives

$$\langle \mathbf{x}' - \mathcal{P}_\mathcal{C}(\mathbf{x}'), \mathcal{P}_\mathcal{C}(\mathbf{x}') - \mathcal{P}_\mathcal{C}(\mathbf{x}) \rangle \geq 0. \tag{2.10}$$

Adding (2.9) and (2.10) gives

$$\langle \mathbf{x} - \mathcal{P}_\mathcal{C}(\mathbf{x}) - (\mathbf{x}' - \mathcal{P}_\mathcal{C}(\mathbf{x}')), \mathcal{P}_\mathcal{C}(\mathbf{x}) - \mathcal{P}_\mathcal{C}(\mathbf{x}') \rangle \geq 0, \tag{2.11}$$

or, equivalently,

$$||\mathcal{P}_\mathcal{C}(\mathbf{x}) - \mathcal{P}_\mathcal{C}(\mathbf{x}')||^2 \leq \langle \mathbf{x} - \mathbf{x}', \mathcal{P}_\mathcal{C}(\mathbf{x}) - \mathcal{P}_\mathcal{C}(\mathbf{x}') \rangle . \tag{2.12}$$

The Cauchy-Schwartz inequality then gives us that $\mathcal{P}_\mathcal{C}$ is continuous. □

We now to problem (2.3).

Definition 2.14. (2.3) is known as a bound constrained minimization problem if

$$\Omega = \{\mathbf{x} \in \mathbb{R}^N \mid L_i \leq x_i \leq U_i \text{ for each } i = 1, \ldots, N\}, \tag{2.13}$$

where $L_i, U_i \in \mathbb{R}$ satisfy $-\infty \leq L_i < U_i \leq \infty$ for each i.

Definition 2.15. (2.3) is known as a lower bound constrained minimization problem if

$$\Omega = \{\mathbf{x} \in \mathbb{R}^N \mid L_i \leq x_i \text{ for each } i = 1, \ldots, N\}, \tag{2.14}$$

where $L_i \in \mathbb{R} \cup \{-\infty\}$.

Remark 2.16. The inequality $L_i \leq x_i$ in (2.14) is called the i^{th} constraint.

Remark 2.17. The feasible set Ω defined by (2.14) or (2.13) is closed and convex, and, therefore, $\mathcal{P}_\Omega$ is well-defined and continuous.

Remark 2.18. In the sequel, unless otherwise stated, Ω will be defined by equation (2.14).

Definition 2.19. Given any feasible point $\mathbf{x}$, the i^{th} constraint is active if $x_i = L_i$ and is inactive if $x_i > L_i$. The active set of indices at $\mathbf{x}$ is given by

$$\mathcal{A}(\mathbf{x}) = \{i \mid x_i = L_i\},$$

and the inactive set by

$$\mathcal{I}(\mathbf{x}) = \{i \mid x_i > L_i\}.$$

Necessary and Sufficient Conditions for a Local Minimizer

Theorem 2.20. (First-Order Necessary Conditions) Let $J \in C^1(\mathbb{R}^N, \mathbb{R})$. If $\mathbf{x}^*$ is a solution of (2.3) then

$$\frac{\partial J}{\partial x_i}(\mathbf{x}^*) \geq 0 \quad \text{if} \quad i \in \mathcal{A}(\mathbf{x}^*), \quad \text{and} \tag{2.15}$$

$$\frac{\partial J}{\partial x_i}(\mathbf{x}^*) = 0 \quad \text{if} \quad i \in \mathcal{I}(\mathbf{x}^*). \tag{2.16}$$

PROOF. First, consider $\phi \in C^1(\mathbb{R}, \mathbb{R})$. From elementary calculus, if $0 = \arg\min_{t \geq 0} \phi(t)$, then $\phi'(0) \geq 0$, while if $0 = \arg\min_{t \geq -\delta} \phi(t)$ for some $\delta > 0$, then $\phi'(0) = 0$. Now define $\phi(t) = J(\mathbf{x}^* + t\mathbf{e}_i)$, where $\mathbf{e}_i$ is the i^{th} standard unit vector. By the chain rule, $\phi'(0) = \langle \mathbf{e}_i, \text{grad } J(\mathbf{x}^*) \rangle = \frac{\partial J}{\partial x_i}(\mathbf{x}^*)$. If $i \in \mathcal{A}(x^*)$ then $0 = \arg\min_{t \geq 0} \phi(t)$ and (2.15) holds. On the other hand, if $i \in \mathcal{I}(\mathbf{x}^*)$ then $0 = \arg\min_{t \geq -\delta} \phi(t)$ for some $\delta > 0$ and (2.16) holds. □

The first-order necessary conditions (2.15) and (2.16) could also have been derived from the Karush-Kuhn-Tucker (KKT) conditions [32, Theorem 12.1]. The KKT conditions provide the first-order necessary conditions for a broad class of constrained optimization problems, of which (2.3) is a special case. The following Corollary gives an alternate formulation of the first-order necessary conditions.

Corollary 2.21. Let $J \in C^1(\mathbb{R}^N, \mathbb{R})$. If $\mathbf{x}^*$ is a solution of (2.3) then

$$\langle \text{grad } J(\mathbf{x}^*), \mathbf{x}^* - \mathbf{y} \rangle \leq 0 \tag{2.17}$$

for all $\mathbf{y} \in \Omega$.

PROOF. If $\mathbf{x}^*$ solves (2.3) then (2.17) follows from (2.15) and (2.16) of Theorem 2.20. □

Theorem 2.22. Let $J \in C^1(\mathbb{R}^N, \mathbb{R})$ be strictly convex. Then $\mathbf{x}$ is a (local) solution of (2.3) if and only if $\mathbf{x}$ is the unique global constrained minimizer for J on Ω.

PROOF. Suppose that $\mathbf{x}$ is a solution of (2.3). Then, by Corollary 2.21, (2.17) holds. Theorem 2.7 then tells us that $J(\mathbf{y}) - J(\mathbf{x}) \geq 0$ for all $\mathbf{y} \in \Omega$. So $\mathbf{x}$ is a global constrained minimizer of J on Ω. Uniqueness follows from Theorem 2.10. The reverse implication is obvious. □

Definition 2.23. If $\mathbf{x} \in \Omega$ satisfies equations (2.15) and (2.16) then $\mathbf{x}$ is a stationary point for problem (2.3). If, in addition, the inequality in (2.15) is strict for each $i \in \mathcal{A}(\mathbf{x})$ then $\mathbf{x}$ is a nondegenerate stationary point for problem (2.3), while if $\frac{\partial J}{\partial x_i}(\mathbf{x}) = 0$ for some $i \in \mathcal{A}(\mathbf{x})$ then $\mathbf{x}$ is a degenerate stationary point for problem (2.3).

Proposition 2.24. Let $\mathbf{x} \in \mathbb{R}^N$. Then $\mathcal{P}_\Omega(\mathbf{x})$ has components

$$[\mathcal{P}_\Omega(\mathbf{x})]_i = \max(x_i, L_i) \text{ for } i = 1, \ldots, N.$$

PROOF. Let $\mathbf{x} \in \mathbb{R}^N$ and $\hat{\mathbf{y}} = \mathcal{P}_\Omega(\mathbf{x})$. Then $\hat{\mathbf{y}} = \arg\min_{\mathbf{y} \in \Omega} J(\mathbf{y})$, where $J(\mathbf{y}) = ||\mathbf{y} - \mathbf{x}||^2$. Furthermore,

$$\frac{\partial J}{\partial y_i}(\hat{\mathbf{y}}) = 2(\hat{y}_i - x_i). \tag{2.18}$$

If $i \in \mathcal{A}(\hat{\mathbf{y}})$, then $\hat{y}_i = L_i$. But by (2.15) and (2.18), $\hat{y}_i \geq x_i$. Similarly, if $i \in \mathcal{I}(\hat{\mathbf{y}})$, then $\hat{y}_i > L_i$. But by (2.16) and (2.18), $\hat{y}_i = x_i$. In either case, $\hat{y}_i = \max(x_i, L_i)$. □

Definition 2.25. The projected gradient of J at $\mathbf{x} \in \Omega$ is given by

$$[\text{grad}_{\mathcal{P}}\ J(\mathbf{x})]_i = \begin{cases} 0, & i \in \mathcal{A}(\mathbf{x}) \text{ and } \frac{\partial J}{\partial x_i}(\mathbf{x}) \geq 0 \\ \frac{\partial J}{\partial x_i}(\mathbf{x}), & \text{otherwise.} \end{cases} \tag{2.19}$$

We can now restate the first-order necessary conditions in terms of the projected gradient.

Theorem 2.26. $\mathbf{x}^*$ is a stationary point for problem (2.3) if and only if $\text{grad}_{\mathcal{P}}\ J(\mathbf{x}^*) = \mathbf{0}$.

PROOF. This follows immediately from (2.15), (2.16), and (2.19). □

The next theorem uses the projected gradient to characterize the global minimizer of a class of convex functions, and follows immediately from Theorems 2.22 and 2.26.

Theorem 2.27. Let $J : \mathbb{R}^N \to \mathbb{R}$ be a smooth, strictly convex function. Then $\text{grad}_{\mathcal{P}}\ J(\mathbf{x}) = \mathbf{0}$ if and only if $\mathbf{x}$ is the unique global constrained minimizer of J on Ω.

In order to determine second-order necessary conditions we will first consider the case in which $\mathcal{A}(\mathbf{x}^*)$ is empty.

Lemma 2.28. Let $J \in C^2(\mathbb{R}^N)$ and suppose $\mathbf{x}^*$ is a solution of (2.3) at which no constraints are active. Then Hess $J(\mathbf{x}^*)$ is positive semidefinite.

PROOF. Suppose Hess $J(\mathbf{x}^*)$ is not positive semidefinite. Then there exists a $\mathbf{p}$ with $||\mathbf{p}|| = 1$ such that $\mathbf{p}^T \text{Hess}\ J(\mathbf{x}^*)\mathbf{p} < 0$. Since at $\mathbf{x}^*$ no constraints are active, there

exists a neighborhood of $\mathbf{x}^*$ which is contained inside of Ω. Therefore, since J is twice continuously differentiable at $\mathbf{x}^*$, by Taylor's Theorem $J(\mathbf{x}^* + \alpha\mathbf{p}) = J(\mathbf{x}^*) + \alpha \langle \mathbf{p}, \text{grad } J(\mathbf{x}^*) \rangle + \frac{\alpha^2}{2} \langle \mathbf{p}, \text{Hess } J(\mathbf{x}^*)\mathbf{p} \rangle + O(\alpha^3)$ as $\alpha \to 0$. By (2.16), grad $J(\mathbf{x}^*) = \mathbf{0}$. Thus $\frac{\alpha^2}{2} \langle \mathbf{p}, \text{Hess } J(\mathbf{x}^*)\mathbf{p} \rangle + O(\alpha^3) = J(\mathbf{x}^* + \alpha\mathbf{p}) - J(\mathbf{x}^*)$. Since $\langle \mathbf{p}, \text{Hess } J(\mathbf{x}^*)\mathbf{p} \rangle < 0$, there exists a $\alpha_0 > 0$ such that $\mathbf{x}^* + \alpha_0\mathbf{p} \in \Omega$ and $\frac{\alpha_0^2}{2} \langle \mathbf{p}, \text{Hess } J(\mathbf{x}^*)\mathbf{p} \rangle + O(\alpha_0^3) < 0$. Thus $J(\mathbf{x}^* + \alpha_0\mathbf{p}) - J(\mathbf{x}^*) < 0$ and $\mathbf{x}^*$ is not a solution of (2.3). □

We now consider the case where $\mathcal{A}(\mathbf{x}^*)$ is not empty. Let $\mathbf{x} = (\boldsymbol{\xi}, \boldsymbol{\eta})$, where, after possibly reordering indices, $\boldsymbol{\xi}$ corresponds to those indices that are inactive at $\mathbf{x}^*$ and $\boldsymbol{\eta}$ to those that are active. Let $\mathbf{x}^* = (\boldsymbol{\xi}^*, \boldsymbol{\eta}^*)$ and define $\hat{J}(\boldsymbol{\xi}) = J(\boldsymbol{\xi}, \boldsymbol{\eta}^*)$. Then $\boldsymbol{\xi}^*$ is an unconstrained local minimizer of $\hat{J}$, and hence, by Lemma 2.28, Hess $\hat{J}(\boldsymbol{\xi}^*)$ is positive semidefinite. We can make no conclusions about second-order derivative information with respect to the active indices. We therefore restrict our attention to the submatrix of Hess $J(\mathbf{x}^*)$ corresponding to Hess $\hat{J}(\boldsymbol{\xi}^*)$. In order to make this precise, we introduce the reduced Hessian.

Definition 2.29. The reduced Hessian of J is the $N \times N$ matrix with entries

$$\text{Hess}_R\, J(\mathbf{x})_{ij} = \begin{cases} \text{Hess } J(\mathbf{x})_{ij}, & \text{if } i \in \mathcal{I}(\mathbf{x}) \text{ or } j \in \mathcal{I}(\mathbf{x}) \\ \delta_{ij}, & \text{otherwise.} \end{cases} \tag{2.20}$$

We can now state the second order necessary conditions for a solution to problem (2.3).

Theorem 2.30. (Second-Order Necessary Conditions) Let $J \in C^2(\mathbb{R}^N)$ and suppose $\mathbf{x}^*$ is a solution of problem (2.3). Then $\text{Hess}_R\, J(\mathbf{x})$ is positive semidefinite.

PROOF. If $\mathbf{x}^*$ has no active indices, $\text{Hess}_R\, J(\mathbf{x}^*) = \text{Hess}\, J(\mathbf{x}^*)$ is positive semidefinite by Lemma 2.28. Otherwise, $\mathbf{x}^*$ has $M < N$ active indices. Define $\hat{J}$ as in the discussion preceding Definition 2.29. Then Hess $\hat{J}(\mathbf{x}^*)$ is a positive semidefinite $(N - M) \times (N - M)$ matrix. Consequently, the $N \times N$ matrix $\begin{bmatrix} \text{Hess}\, \hat{J}(\mathbf{x}^*) & 0 \\ 0 & I_{M\times M} \end{bmatrix}$ is positive semidefinite. Furthermore, it is equivalent to $\text{Hess}_R\, J(\mathbf{x})$ with respect to a reordering of indices, which implies that $\text{Hess}_R\, J(\mathbf{x})$ is also positive semidefinite. □

As in the case of the first-order necessary conditions, the second-order necessary conditions could also have been derived in terms of the KKT conditions [32, Theorem 12.5]. The same can be said for the second-order sufficient conditions [32, Theorem 12.6], which we will present next.

Theorem 2.31. (Second-Order Sufficient Conditions) Let $\mathbf{x}^*$ be a nondegenerate stationary point for problem (2.3), and suppose that $\text{Hess}_R\, J(\mathbf{x}^*)$ is positive definite. Then $\mathbf{x}^*$ is a strict local constrained minimizer for J on Ω.

PROOF. Choose any nonzero $\mathbf{p} \in \mathbb{R}^N$ such that $\mathbf{x}^* + t\mathbf{p} \in \Omega$ for t small and positive. Then $p_i \geq 0$ whenever $i \in \mathcal{A}(\mathbf{x}^*)$. Define $\phi(t) = J(\mathbf{x}^* + t\mathbf{p})$ for all t such that $\mathbf{x}^* + t\mathbf{p} \in \Omega$. It suffices to show that ϕ has a local minimum at $t = 0$. By Taylor's theorem, $\phi(t) = J(\mathbf{x}^*) + t\mathbf{p}^T \text{grad}\, J(\mathbf{x}^*) + \frac{t^2}{2}\mathbf{p}^T \text{Hess}\, J(\mathbf{x}^*)\mathbf{p} + O(t^3)$ for small t. First, suppose there exists $\delta > 0$ such that $\mathbf{x}^* + t\mathbf{p} \in \Omega$ for all $t \in (-\delta, \delta)$. Then $p_i = 0$ whenever $i \in \mathcal{A}(\mathbf{x}^*)$. On the other hand, since $\mathbf{x}^*$ is a stationary point of (2.3), whenever $i \in \mathcal{I}(\mathbf{x}^*)$ $\frac{\partial J}{\partial x_i}(\mathbf{x}^*) = 0$. Hence $\phi'(0) = \mathbf{p}^T \text{grad}\, J(\mathbf{x}^*) = 0$. Furthermore,

$\phi''(0) = \mathbf{p}^T \text{Hess } J(\mathbf{x}^*)\mathbf{p} = \mathbf{p}^T \text{Hess}_R\, J(\mathbf{x}^*)\mathbf{p} > 0$, since $\text{Hess}_R\, J(\mathbf{x}^*)$ is positive definite. Thus ϕ has a local minimum at $t = 0$. In the other case, $p_i > 0$ for some $i \in \mathcal{A}(\mathbf{x}^*)$, and, since $\mathbf{x}^*$ is a nondegenerate stationary point of (2.3), $\frac{\partial J}{\partial x_i}(\mathbf{x}^*) > 0$ whenever $i \in \mathcal{A}(\mathbf{x}^*)$, while $\frac{\partial J}{\partial x_i}(\mathbf{x}^*) = 0$ whenever $i \in \mathcal{I}(\mathbf{x}^*)$. Hence $\phi'(0) = \mathbf{p}^T \text{grad } J(\mathbf{x}^*) = \sum_{i \in \mathcal{A}(\mathbf{x})} p_i \frac{\partial J}{\partial x_i}(\mathbf{x}^*) > 0$. Since $\mathbf{x} + t\mathbf{p} \notin \Omega$ for $t < 0$, this implies that ϕ has a local minimum at $t = 0$. □

The second-order sufficient conditions are not necessary. To see this, consider $J(x) = x^4$ on $\Omega = [-1, \infty)$. Then $x^* = 0$ is a global minimizer of J on Ω. The second-order necessary conditions for J are satisfied at x^*, while, since Hess $J(x) = 12x^2$ is zero at x^*, the second-order sufficient conditions are not satisfied.

The Gradient Projection Algorithm

The gradient projection algorithm can be viewed as a generalization of the method of steepest descent [32] for unconstrained optimization. In this section we follow the exposition of Bertsekas [?]. The gradient projection algorithm generates a sequence $\{\mathbf{x}_k\}$ as follows:

Algorithm 1. (Gradient Projection Algorithm)

(0) Select initial guess $\mathbf{x}_0$, and set $k = 0$.

(1) Compute $\mathbf{p}_k = -\text{grad } J(\mathbf{x}_k)$.

(2) Compute $\lambda_k = \arg\min_{\lambda > 0} J(\mathcal{P}_\Omega(\mathbf{x}_k + \lambda \mathbf{p}_k))$.

(3) Set $\mathbf{x}_{k+1} = \mathcal{P}_\Omega(\mathbf{x}_k + \lambda_k \mathbf{p}_k)$.

(4) If termination criteria are met, STOP.

(5) Otherwise, update $k := k + 1$ and return to (1).

Remark 2.32. Motivated by Theorem 2.26, our objective is to create a sequence $\{\mathbf{x}_k\}$ such that $\text{grad}_{\mathcal{P}}\ J(\mathbf{x}_k) \to \mathbf{0}$. We can take our termination criterion in Step (4) of Algorithm 1 to be

$$||\text{grad}_{\mathcal{P}}\ J(\mathbf{x}_k)|| < \tau,$$

where $\tau > 0$ is a given stopping tolerance.

Remark 2.33. Step (2) above is known as a projected line search. Typically it cannot be solved exactly, but, as we will see, this is not necessary. We begin by defining the function $\phi_k : \mathbb{R} \to \mathbb{R}$ by

$$\phi_k(\lambda) = J(\mathbf{x}_k(\lambda)), \tag{2.21}$$

where

$$\mathbf{x}_k(\lambda) := \mathcal{P}_\Omega(\mathbf{x}_k + \lambda \mathbf{p}_k)$$

and $\mathbf{p}_k = -\text{grad}\ J(\mathbf{x}_k)$. The goal of our algorithm is, given an approximate solution $\mathbf{x}_k$ of problem (2.3) and a search direction $\mathbf{p}_k$, to generate a λ_k which satisfies the sufficient decrease condition

$$\phi_k(\lambda_k) \leq \phi_k(0) - \frac{\mu}{\lambda_k}||\mathbf{x}_k - \mathbf{x}_k(\lambda_k)||^2, \tag{2.22}$$

where $\mu \in (0, \frac{1}{2})$. Typically, μ is set to 10^{-4} [23].

For the remainder of this chapter, we will assume that the Hess $J(\mathbf{x})$ is positive definite for every $\mathbf{x} \in \Omega$, and that there exists λ_{min} and λ_{max} such that

$$0 < \lambda_{min} \leq \lambda(\mathbf{x}) \leq \lambda_{max} \tag{2.23}$$

for every eigenvalue $\lambda(\mathbf{x})$ of Hess $J(\mathbf{x})$ and for every $\mathbf{x} \in \Omega$.

We now present an iterative (approximate) line search algorithm that will be used in the sequel. It is based on the line search algorithm found in Moré and Toreldo [30]. First, we choose our initial approximation λ_k^0 of λ_k to be the minimizer of the quadratic Taylor approximation of ϕ_k,

$$q_k(\lambda) = \phi_k(0) + \lambda\phi_k'(0) + \frac{\lambda^2}{2}\phi_k''(0).$$

In other words, take

$$\lambda_k^0 = \frac{-\langle \text{grad } J(\mathbf{x}_k), \text{grad}_{\mathcal{P}} J(\mathbf{x}_k)\rangle}{\langle \text{Hess } J(\mathbf{x}_k)\text{grad}_{\mathcal{P}} J(\mathbf{x}_k), \text{grad}_{\mathcal{P}} J(\mathbf{x}_k)\rangle}. \tag{2.24}$$

Then, given a λ_k^i that does not satisfy (2.22), we compute λ_k^{i+1} by first defining $\hat{\lambda}_k^{i+1}$ to be the minimizer of the quadratic function q_k that satisfies $q_k(0) = \phi_k(0)$, $q_k'(0) = \phi_k'(0)$, and $q_k(\lambda_k^i) = \phi_k(\lambda_k^i)$. Then

$$\lambda_k^{i+1} \stackrel{\text{def}}{=} \text{median}\left[\frac{1}{100}\lambda_k^i, \hat{\lambda}_k^{i+1}, \frac{1}{2}\lambda_k^i\right]. \tag{2.25}$$

The line search algorithm is then given follows:

Algorithm 2. (Line Search Algorithm)

(0) Compute λ_k^0 as in (2.24), and set $i = 0$.

(1) If λ_k^i satisfies (2.22), set $\lambda_k = \lambda_k^i$ and STOP.

(2) Otherwise, compute λ_k^{i+1} by equation (2.25).

(3) Update $i := i + 1$ and return to (1).

A natural question arises. Will the above algorithm identify a suitable λ_k in a finite number of iterations? The answer is yes, but to prove this fact we will first prove two lemmas.

Lemma 2.34. Define $\mathbf{x}(\lambda) = \mathcal{P}_\Omega(\mathbf{x} - \lambda \text{grad } J(\mathbf{x}))$. Then for every $\mathbf{x}$, $\mathbf{y} \in \Omega$ and $\lambda \geq 0$ we have

$$\lambda \langle \text{grad } J(\mathbf{x}), \mathbf{y} - \mathbf{x}(\lambda) \rangle \geq \langle \mathbf{x} - \mathbf{x}(\lambda), \mathbf{y} - \mathbf{x}(\lambda) \rangle . \tag{2.26}$$

PROOF. By Lemma 2.12, with $\mathbf{x}$ replaced by $\mathbf{x} - \lambda \text{grad } J(\mathbf{x})$ and $\mathcal{C} = \Omega$,

$$\langle \mathbf{x} - \lambda \text{grad } J(\mathbf{x}) - \mathbf{x}(\lambda), \mathbf{y} - \mathbf{x}(\lambda) \rangle \leq 0,$$

for all $\mathbf{x}$ and $\mathbf{y}$ in Ω, which is equivalent to (2.26). □

Lemma 2.35. Assume that grad J is Lipschitz continuous with Lipschitz constant L. Then, for any $\mathbf{x} \in \Omega$, the sufficient decrease condition (2.22) holds for all λ such that

$$0 \leq \lambda \leq \frac{2(1-\mu)}{L}. \tag{2.27}$$

PROOF. Let $\mathbf{x}(\lambda)$ be as in the last lemma, and let $\phi(\lambda)$ be defined by equatin (2.21). Then, for any $\lambda \geq 0$ and $\mathbf{x} \in \Omega$, since $\mathbf{x}(\lambda)$ is continuous and piecewise differentiable,

$$\begin{aligned}
\phi(1) - \phi(0) &= J(\mathbf{x}(\lambda)) - J(\mathbf{x}) \\
&= \langle \text{grad } J(\mathbf{x}), \mathbf{x}(\lambda) - \mathbf{x} \rangle - \\
&\qquad \int_0^1 \langle \text{grad } J[\mathbf{x} - t(\mathbf{x} - \mathbf{x}(\lambda))] - \text{grad } J(\mathbf{x}), \mathbf{x}(\lambda) - \mathbf{x} \rangle \, dt \\
&\leq \frac{-1}{\lambda} ||\mathbf{x} - \mathbf{x}(\lambda)||^2 + \int_0^1 L||t(\mathbf{x} - \mathbf{x}(\lambda))|| \cdot ||\mathbf{x} - \mathbf{x}(\lambda)|| dt \\
&= \left(\frac{-1}{\lambda} + \frac{L}{2} \right) ||\mathbf{x} - \mathbf{x}(\lambda)||^2 \\
&= \left(\frac{\lambda L - 2}{2} \right) \frac{1}{\lambda} ||\mathbf{x} - \mathbf{x}(\lambda)||^2.
\end{aligned}$$

Inequality (2.22) then follows immediately from (2.27). □

We can now prove the theorem that tells us our line search algorithm will eventually generate an acceptable step length λ_k.

Theorem 2.36. Suppose that grad J is Lipschitz continuous. Then, in Algorithm 2, there exists an N such that λ_k^N satisfies the sufficient decrease condition (2.22).

PROOF. Let L be the Lipschitz constant of grad J. There exists an N such that $\lambda_k^0/2^N < \frac{2(1-\mu)}{L}$. Furthermore, (2) of Algorithm 2 tells us that $\lambda_k^N \leq \lambda_k^0/2^N$. The result then follows from Lemma 2.35. □

We can now prove the main result of this section, which gives us a necessary condition for the convergence of Algorithm 1. But first, we will need to prove a lemma.

Lemma 2.37. Let $\lambda_k = \lambda_k^N$ be the output from Algorithm 2 at iteration k of Algorithm 1. Then

$$\min\left\{\frac{1}{\lambda_{max}}, (1-\mu)/(50\cdot L)\right\} \leq \lambda_k \leq \frac{1}{\lambda_{min}} \tag{2.28}$$

for every k.

PROOF. Equation (2.24) can be written

$$\lambda_k^0 = \frac{1}{\langle \text{Hess } J(\mathbf{x}_k)\mathbf{d}_k, \mathbf{d}_k\rangle},$$

where $\mathbf{d}_k = \text{grad}_{\mathcal{P}}\, J(\mathbf{x}_k)/||\text{grad}_{\mathcal{P}}\, J(\mathbf{x}_k)||$. Then

$$\frac{1}{\lambda_{max}} \leq \lambda_k^0 \leq \frac{1}{\lambda_{min}}, \tag{2.29}$$

which gives us the upper bound in (2.28). The lower bound in (2.29) together with Lemma 2.35 and (2) of Algorithm 2 gives us the lower bound in (2.28). □

Theorem 2.38. Assume that J is Lipschitz continuous with Lipschitz constant L. Let $\{\mathbf{x}_k\}$ be a sequence generated by the gradient projection algorithm using an inexact line search satisfying (2.22). Then every limit point of $\{\mathbf{x}_k\}$ is a stationary point for problem (2.3). In particular, if $\mathbf{x}_k \to \hat{\mathbf{x}}$ then $\hat{\mathbf{x}}$ is a stationary point of (2.3).

PROOF. Since our line search algorithm is guaranteed to generate an acceptable λ_k, Algorithm 1 is well defined. If $\{\mathbf{x}_k\}$ generated by Algorithm 1 stops at a stationary point we are done. Otherwise, suppose $\{\mathbf{x}_k\}_{k\in\mathcal{K}}$ is a subsequence which converges to a vector $\bar{\mathbf{x}}$. Since $\{J(\mathbf{x}_k)\}$ is monotone decreasing, $J(\mathbf{x}_k) \to J(\bar{\mathbf{x}})$. By (2.22),