UNIVERSITY OF CALIFORNIA

Los Angeles

Accelerated First-Order Optimization with Orthogonality Constraints

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Mathematics

by

Jonathan Wolfram Siegel

2018

ABSTRACT OF THE DISSERTATION

Accelerated First-Order Optimization with Orthogonality Constraints

by

Jonathan Wolfram Siegel
Doctor of Philosophy in Mathematics
University of California, Los Angeles, 2018
Professor Russel E. Caflish, Chair

Optimization problems with orthogonality constraints have many applications in science and engineering. In these applications, one often deals with large-scale problems which are ill-conditioned near the optimum. Consequently, there is a need for first-order optimization methods which deal with orthogonality constraints, converge rapidly even when the objective is not well-conditioned, and are robust.

In this dissertation we develop a generalization of Nesterov's accelerated gradient descent algorithm for optimization on the manifold of orthonormal matrices. The performance of the algorithm scales with the square root of the condition number. As a result, our method outperforms existing state-of-the-art algorithms on large, ill-conditioned problems. We discuss applications of the method to electronic structure calculations and to the calculation of compressed modes.

The dissertation of Jonathan Wolfram Siegel is approved.

Vidvuds Ozolins

Christopher R. Anderson

Stanley J. Osher

Russel E. Caflish, Committee Chair

University of California, Los Angeles

2018

To My Family

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank my adviser, Russel Caflisch. I am deeply grateful to him for his continued guidance, support, and kindness. This thesis would have been impossible without his encouragement. We have had many interesting conversations which have shaped my understanding of mathematics.

I would also like to thank Stanley Osher and Vidvuds Ozolins for introducing me to the problems studied in this thesis. Their advice greatly aided my research. I am indebted to Chris Anderson as well, for his courses and seminars, as well as his guidance. He played an important role in shaping my understanding of computational math and numerical analysis.

I would like to thank the many great teachers that I had throughout my time at UCLA. In particular, Inwon Kim and Monica Visan helped me develop my understanding of analysis and PDE theory, and I enjoyed their courses very much. I am also indebted to Chris Anderson and Joseph Teran for their courses on numerical analysis and to Frank Jenko for his course on Plasma physics.

I would like to thank all of my collaborators and friends at UCLA, which whom I have had many interesting and inspiring conversations. Joshua Keneda, Dustan Levenstein, Omer Tekin, Edward Chou, Mike Menke, Mike Miller, and Will Oakley, to name just a few. You all made my time at UCLA much more fun and interesting.

I would like to thank the Professors at UC Santa Cruz, who introduced me to mathematics and to whom I am indebted for helping me discover my passion for it. Bruce Cooperstein, Robert Boltje, and Samit Dasgupta, your courses and guidance inspired me to study mathematics. Even though I ultimately chose to study applied math, I am very grateful for everything I learned from you.

I would like to thank my parents and my sister, who have always supported me. It is hard to imagine where I would be without their love and encouragement. Finally, I would like to thank my girlfriend Jessica, who travelled to New York with me and has always helped support and motivate me.

VITA

2011	Honorable Mention in the 72nd Annual William Lowell Putnam Mathematical Competition
2012	Stephen M. Palais Award, Mathematics Department UC Santa Cruz
2013	B.S. in Mathematics, University of California, Santa Cruz
2013-2014	UC Regents Fellow, University of California at Los Angeles
2014-2015	Teaching Assistant, University of California at Los Angeles
2015-2018	Research Assistant, University of California at Los Angeles
2018	Pacific Journal of Mathematics Dissertation Prize

PUBLICATIONS

Siegel, Jonathan, and Tekin, Omer. “Compact support of L^1 penalized variational problems”. In: *Communications in Mathematical Sciences* 15.6 (2017): 1771-1790.

Siegel, Jonathan W. “Shift Invariant Subspaces and Applications To Signal Fragmentation”. *UCLA Cam Reports, To be Submitted*

CHAPTER 1

Introduction

Optimization problems over the set of orthonormal matrices appear naturally in many scientific and engineering problems. Most notably, eigenfunction and electronic structure calculations involve minimizing functions over the set of orthonormal matrices [1, 8, 22]. In these applications, the objective functions are smooth but often ill-conditioned. There are also recent applications which involve non-smooth objectives, most notably the calculation of compressed modes [28], which involve an L^1 penalization of variational problems arising in physics.

Due to the importance of the practical applications, there has been much research directed toward developing efficient optimization procedures which can deal with orthogonality constraints. A general framework for optimization on manifolds and in particular the manifold defined by $X^TX = I$ (which is called the Stiefel manifold) has been developed in [1, 8]. This framework gives a way of generalizing traditional Euclidean optimization procedures to the Stiefel manifold, most notably gradient descent, non-linear conjugate gradient, and Newton's method.

However, for large scale problems Newton's method is far too expensive and existing first-order methods converge slowly when applied to ill-conditioned problems. Since many applications involve large-scale, ill-conditioned problems, there is a need for first-order methods which converge more rapidly when the objective is not well-conditioned.

In this dissertation, we address this problem by adapting momentum-based accelerated gradient methods, such as Nesterov's accelerated gradient descent [24], to the Stiefel manifold. Generalizing these methods involves developing a version of accelerated gradient descent which is specifically designed to deal with the problems that arise on the Stiefel manifold, in addition to introducing efficient algorithms for extrapolating and interpolating on the manifold. We show empirically that the methods are very robust and efficient even when the objective is ill-conditioned.

We begin, in chapter 2, by studying variational problems related to compressed modes. Here we prove that certain L^1 penalized variational problems have compactly supported solutions. In particular, we extend results of Brezis [6] to show that compact support holds even when the L^1 term comes with an inhomogeneous weight.

Chapter 3 is the heart of the dissertation. Here we develop the manifold accelerated gradient descent algorithms in full detail. They are then tested on eigensystem calculations of varying condition number and the relation between the convergence and the condition number is analyzed.

In chapter 4, we consider the application of the algorithms to a simple electronic structure calculation. We demonstrate that our algorithms perform reasonably well and propose them as an alternative to the self-consistent field iteration.

Finally, in chapter 5, we consider the problem of calculating compressed modes. We discuss splitting methods for calculating compressed modes as well as feasible subgradient methods. However, we have discovered that the most robust and efficient way of calculating compressed modes is to smooth the L^1 term and use the algorithms developed in chapter 3 to solve the resulting (ill-conditioned) problem. We provide numerical results and analyze the performance of this approach.

CHAPTER 2

Variational Problems with L^1 Terms

The use of an l^1 penalty has been used with great success in the field of compressed sensing to enforce prior knowledge of sparsity (see [34]). Intuitively, this approach works because the l^1 norm is a convex relaxation of the l^0 norm. In many physical simulations, notably electronic structure calculations of insulators, we have prior knowledge that the solutions are spatially localized. The analogy with compressed sensing has led to the idea that an L^1 penalty term can enforce this prior knowledge, as pioneered in [28] and [5].

In these situations, it is important to known that the solutions to L^1 penalized variational problems in physics have compact support. This chapter is concerned with proving this for eigenvalue and elliptic problems. In particular, our contribution is an extension of existing results to the case were there L^1 norm comes with a non-uniform weight.

To conclude the chapter, we present the results of some numerical experiments in which we solve L^1 penalized elliptic variational problems.

2.1 L^1 Constrained Eigenvalue Problems

In this section, we consider the solution to the following problem on all of $\mathbb{R}^n$

$$\underset{\|u\|_2=1, u\in H^1(\mathbb{R}^n)}{\arg\min} \|\nabla u\|_2^2 + \gamma\|u\|_1 \tag{2.1.1}$$

where $\gamma > 0$. This problem is motivated by the variational formulation of the first eigenvalue/eigenvector of the Laplace operator on a bounded domain. Note that the existence of a solutions to (2.1.1) is non-trivial. In fact, if we remove the L^1 term, this problem has no minimizer as the domain is all of $\mathbb{R}^n$. The existence and compact support of solutions to

(2.1.1) was studied in [4]. We provide a new approach to proving the existence of minimizers which is more flexible and allows us to deal with problems of the form

$$\underset{\|u\|_2=1, u\in H^1}{\arg\min} \|\nabla u\|_2^2 + \|w(x)u\|_1 \tag{2.1.2}$$

as long as the weight, $w(x)$ is a non-decreasing, non-zero, positive radial function.

There are two ingredients to our proof of existence and compact support of solutions to (2.1.2). The first is a rearrangement inequality and the second is a compactness result. We begin with the rearrangement inequality.

Define the symmetric decreasing rearrangement of a function f as follows.

Definition 2.1.1. *Let A be a borel measurable set in $\mathbb{R}^n$. The symmetric rearrangement of A is $A^* = \{x \in \mathbb{R}^d : |x| < r\}$ where r is chosen such that $|A| = |A^*|$. In other words, A^* is the ball centered at the origin with the same measure as A.*

Now let $f : \mathbb{R}^d \to \mathbb{C}$ be a borel measurable function. The symmetric decreasing rearrangement of f is

$$f^*(x) = \int_0^\infty \chi_{\{|f|>\lambda\}^*}(x) d\lambda$$

Note that f^* has the same distribution function as f, i.e. $|\{|f| > \lambda\}| = |\{|f^*| > \lambda\}|$ for all λ. In particular, $\|f\|_p = \|f^*\|_p$ for all p.

We need the following theorem concerning the symmetric decreasing rearrangement, due to Polyá and Szego.

Theorem 2.1.1. *Let $f \in W^{1,p}$ for $1 \le p \le \infty$. Then $f^* \in W^{1,p}$ and*

$$\|\nabla f^*\|_p \le \|\nabla f\|_p$$

Note that the above theorem is related to the isoperimetric inequality. In fact, for $p = 1$ it implies the isoperimetric inequality. We will only need the case $p = 2$ of the above theorem, which can be found in [29].

Next we describe the compactness result. We will prove the following theorem.

Theorem 2.1.2. *Fix $d \geq 2$. Then $H^1_{rad} \cap L^1$ is compactly contained in L^p for $1 < p < \frac{2d}{d-2}$.*

This result is known to be true (due to Gagliardo-Nierenberg) in the case $2 < p < \frac{2d}{d-2}$ even without the L^1 condition. Adding the L^1 condition allows us to use L^p interpolation to extend the result to $1 < p < \frac{2d}{d-2}$.

In order to prove this result we will need the following lemmas from harmonic analysis. (Note that $\lesssim$ means $\leq$ up to a constant factor independent of the function showing up on both sides.)

Lemma 2.1.1. *Let $u \in H^1$, then for $2 \leq p \leq \frac{2d}{d-2}$ (for $d = 1, 2,\ 2 \leq p < \infty$) we have*

$$\|u\|_p \lesssim \|\nabla u\|_2^\theta \|u\|_2^{1-\theta}$$

where $\theta = \frac{2d-p(d-2)}{2p}$.

The previous lemma is the well-known Galgiardo Nierenberg inequality [26]. It follows from the Sobolev embedding theorem in dimension ≥ 3. In dimensions 1 and 2 it is a generalization of Sobolev Embedding.

The next lemma is called the radial Sobolev inequality, which we prove here.

Lemma 2.1.2. *Let $d \geq 2$ and $1 \leq q < \frac{2d}{d-2}$. Let $f \in L^q \cap H^1$ be radial. Then*

$$r^{\frac{2(d-1)}{q+2}} |f(r)| \lesssim \|f\|_q^{\frac{q}{q+2}} \|\nabla f\|_2^{\frac{2}{q+2}}$$

a.e.

Proof. Notice that since $|\nabla|f|| \leq |\nabla f|$ a.e., it suffices to consider the case where $f \geq 0$. We claim that it also suffices to consider the case where f is a Schwartz function. This is so because Schwartz functions are dense in $L^q \cap \dot{H}^1$ and if $f_n \to f$ in $L^q \cap \dot{H}^1$, then a subsequence converges to f a.e.

So assume that f is a non-negative, radial Schwartz function. We have

$$r^{d-1}|f(r)|^{1+\frac{q}{2}} = r^{d-1}\left(1 + \frac{q}{2}\right)\int_r^\infty |f(t)|^{\frac{q}{2}} f'(t) dt$$

Since $t \geq r$ in the above integration we have that the above is bounded by

$$\left(1+\frac{q}{2}\right)\int_r^\infty |f(t)|^{\frac{q}{2}}|f'(t)|t^{d-1}dt$$

We now apply Cauchy-Schwartz to bound the above by

$$\left(1+\frac{q}{2}\right)\left(\int_r^\infty |f(t)|^q t^{d-1}dt\right)^{\frac{1}{2}}\left(\int_r^\infty |f'(t)|^2 t^{d-1}dt\right)^{\frac{1}{2}}$$

$$\leq \left(1+\frac{q}{2}\right)\|f\|_q^{\frac{q}{2}}\|\nabla f\|_2$$

Taking everything to the power $\left(1+\frac{q}{2}\right)^{-1}$, we obtain the lemma. □

We will also need the following characterization of compact subsets of L^p, which is essentially a generalization of the Arzela-Ascoli theorem for $p < \infty$, due to Kolmogorov and Riesz [16, 30].

Lemma 2.1.3. *Let $X \subset L^p$. Then X is precompact in L^p iff the following hold*

1. *X is uniformly bounded, i.e. there exists $M > 0$ s.t. $\|f\|_p < M$ for all $f \in X$.*

2. *X is uniformly equicontinuous, i.e. for all $\epsilon > 0$, there exists a $\delta > 0$ such that*

   $$\|f(x) - f(x-y)\|_{L^p(\mathbb{R}^d)} < \epsilon$$

   whenever $|y| < \delta$, for all $f \in X$.

3. *X is uniformly tight, i.e. for every $\epsilon > 0$, there exists an $R > 0$ such that*

   $$\|f\|_{L^p(B(0,R)^c)} < \epsilon$$

   for all $f \in X$.

We are now in a position to prove Theorem (2.1.2).

Proof. We prove this by verifying each of the conditions given in lemma (2.1.3), when X is the unit ball B_1 in $H^1_{rad} \cap L^1$. First of all, B_1 is uniformly bounded in L^p by the Gagliardo-Nierenberg inequality and interpolation of L^p norms.

Next we must verify equicontinuity. If $p \geq 2$ we see that by Gagliardo-Nierenberg,

$$\|f(x) - f(x-y)\|_{L^p(\mathbb{R}^d)} \lesssim \|\nabla f(x) - \nabla f(x-y)\|^{\theta}_{L^2(dx)} \|f(x) - f(x-y)\|^{1-\theta}_{L^2(dx)}$$

$$\leq 2\|\nabla f\|_2^{\theta} \|\nabla f\|_2^{1-\theta} |y|^{1-\theta} \leq \|f\|_{H^1} |y|^{1-\theta}$$

Now since $1 - \theta > 0$ we get equicontinuity. For $1 < p < 2$ we use interpolation of L^p norms in combination with the result for $p \geq 2$. In particular, we write

$$\|f(x) - f(x-y)\|_{L^p(\mathbb{R}^d)} \lesssim \|f(x) - f(x-y)\|^{\theta}_{L^1(dx)} \|f(x) - f(x-y)\|^{1-\theta}_{L^2(dx)}$$

Now the first term above is bounded by $2\|f\|_1$ and the second term can be bounded as before in terms of a power of $|y|$. Since $p > 1$, $\theta < 1$ and we get the desired equicontinuity.

Finally we verify the tightness. To do this we write

$$\int_{|x|>R} |f(x)|^p dx = \int_{|x|>R} |f(x)|^{\delta} |f(x)|^{p-\delta} dx$$

Now using lemma (2.1.2) with $q = 2$ we see that $|f(x)| \lesssim |x|^{-\frac{d-1}{2}}$. Thus the above integral is

$$\lesssim R^{-\delta \frac{d-1}{2}} \int_{|x|>R} |f(x)|^{p-\delta} dx$$

Setting $\delta = p - 1$ we get

$$\int_{|x|>R} |f(x)|^p dx \lesssim R^{-(p-1)\frac{d-1}{2}} \|f\|_1$$

which completes the proof.

□

In order to show the existence of compactly supported minimizers to (2.1.1), we proceed as follows.

Let x_n be a minimizing sequence, i.e. $\|x_n\|_2 = 1$ and $\|\nabla x_n\|_2^2 + \|x_n\|_1$ converges to the optimal value. By the Polyá-Szego theorem and the trivial properties of the symmetric decreasing rearrangement, we see that taking the symmetric rearrangement of the x_n results in another minimizing sequence. Hence we may assume that the x_n are radial, non-negative, and decreasing. Note that x_n is bounded in $H^1_{rad} \cap L^1$, so by the compactness result we can take a subsequence which converges in L^2. We can also take a further subsequence which converges almost everywhere and weakly in H^1_{rad} (by the BanachAlaoglu theorem). Call u the limit (in L^2) of this sequence. Then we have $\|u\|_2 = 1$ since our sequence converges strongly in L^2. We also have, from the properties of weak convergence, that $\|\nabla u\|_2^2 \leq \lim \|\nabla x_n\|_2^2$. Additionally, since the sequence converges a.e., by Fatou's lemma we have $\|u\|_1 \leq \lim \|x_n\|_1$. But since x_n is a minimizing sequence we can't have strict inequality in the preceeding two inequalities. Hence $\|\nabla u\|_2^2 + \|u\|_1$ is optimal and we have found a minimizer.

We can use the above compactness results to extend this result and show the existence of radial, non-negative, decreasing minimizers to problems of the form

$$\underset{\|u\|_2=1, u\in H^1}{\arg\min} \frac{1}{2}\|\nabla u\|_2^2 + \|w(x)u\|_1 \tag{2.1.3}$$

as long as the weight, $w(x)$ is a non-decreasing, non-zero, positive radial function.

Theorem 2.1.3. *There exist radial, non-negative, decreasing minimizers to*

$$\underset{\|u\|_2=1, u\in H^1}{\arg\min} \frac{1}{2}\|\nabla u\|_2^2 + \|w(x)u\|_1$$

where $w(x)$ is a non-decreasing, non-zero, positive radial function.

Proof. First we will show that

$$\|w(x)u^*\|_1 \leq \|w(x)u\|_1$$

where u^* is the symmetric decreasing rearrangement of u. To show this we note that

$$\|w(x)u\|_1 = \int_{\mathbb{R}^d} w(x)|u(x)|dx$$
$$= \int_{\mathbb{R}^d}\int_0^\infty \chi_{\{w>\lambda\}}(x)d\lambda \int_0^\infty \chi_{\{|u|>\mu\}}(x)d\mu dx$$

Here $\chi_{\{w>\lambda\}}(x)$ is the characteristic function of the set $\{w(x) > \lambda\}$ and $\chi_{\{|u|>\mu\}}(x)$ is the characteristic function of the set $\{w(x) > \mu\}$. This equality follows since

$$|f(x)| = \int_0^\infty \chi_{\{|f|>\lambda\}}(x)d\lambda$$

for all measurable f.

Now we switch the order of integration in the above to obtain

$$\|w(x)u\|_1 = \int_0^\infty \int_0^\infty \int_{\mathbb{R}^d} \chi_{\{w>\lambda\}}(x)\chi_{\{|u|>\mu\}}(x)dxd\lambda d\mu$$
$$= \int_0^\infty \int_0^\infty |\{w > \lambda\} \cap \{|u| > \mu\}|d\lambda d\mu$$

Now we claim that $|\{w > \lambda\} \cap \{|u| > \mu\}| \geq |\{w > \lambda\} \cap \{u^* > \mu\}|$ for all λ and μ. This follows since by assumption, $\{w > \lambda\}$ is the complement of a ball centered at the origin and $\{u^* > \mu\}$ is a ball centered at the origin. Thus if $|\{w > \lambda\} \cap \{u^* > \mu\}| > 0$ then $\{u^* > \mu\}$ covers the entire complement of $\{w > \lambda\}$. Since $|\{u^* > \mu\}| = |\{|u| > \mu\}|$, we have that $|\{w > \lambda\} \cap \{|u| > \mu\}| \geq |\{w > \lambda\} \cap \{u^* > \mu\}|$.

Integrating this with respect to λ and μ we get that

$$\|w(x)u^*\|_1 \leq \|w(x)u\|_1$$

Thus, by taking symmetric decreasing rearrangements we may assume that any minimizing sequence consists of radial functions. Now, as in the previous proof of existence, the compactness result implies the existence of a minimizer if we can uniformly bound the $|\cdot|_1$ norm of the minimizing sequence.

This follows since under the assumptions on w, there is a radius R and a constant $C > 0$ such that $w(x) > C$ if $|x| > R$. Consequently we see that $\|u\|_{L^1(\{|x|>R\})} < C\|w(x)u\|_1$, which implies that $\|u_n\|_{L^1(\{|x|>R\})}$ is uniformly bounded ($\|w(x)u_n\|_1$ is uniformly bounded as it is a minimizing sequence). Now any minimizing sequence satisfies $\|u_n\|_2 = 1$, and thus $\|u_n\|_{L^1(\{|x|<R\})} \leq R^{1/2}\|u_n\|_2$. So, since R only depends on w, we have a uniform bound on $\|u_n\|_1$ for any minimizing sequence.

The above compactness result finishes the proof. □

We now turn to proving the compact support of the solution to (2.1.3).

Theorem 2.1.4. *The radial, non-negative, decreasing solutions to (2.1.1) and (2.1.3) have compact support.*

Proof. Note that the result will follow if we can show that the measure of the support is finite. This is true since a radial, non-negative, decreasing function will have support which is a ball. To this end we note that the solution u satisfies

$$\lambda u \in -\Delta u + w(x)\beta(u)$$

where β is the subdifferential of $|\cdot|$. Multiplying this by the sign of u (u is non-negative so this is just $\chi_{\{u>0\}}$) and integrating we see that

$$\lambda \|u\|_1 = \int_{\mathbb{R}^n} -\Delta u \chi_{\{u>0\}} dx + \int_{\{u>0\}} w(x) dx$$

This is true since $u \cdot \operatorname{sgn}(u) = |u|$ and $\beta(u) \cdot \operatorname{sgn}(u) = \chi_{\{u>0\}}$ (this follows since $\beta(u) = 1$ for $u > 0$ and $\operatorname{sgn}(0) = 0$). Now we consider the term

$$\int_{\mathbb{R}^n} -\Delta u \chi_{\{u>0\}} dx$$

the divergence theorem yields that this is equal to

$$\int_{\partial\{u>0\}} -\nabla u \cdot \nu dS$$

where ν is the outward normal of $\{u > 0\}$. Since $u > 0$ on the interior of $\{u > 0\}$ we have that $-\nabla u \cdot \nu \geq 0$. Thus the above integral is positive. Hence we obtain

$$\int_{\{u>0\}} w(x) dx \leq \lambda \|u\|_1 < \infty$$

Now by the assumptions on w, we have that $w(x) > C$ for $|x| > R$ for some $R > 0$ and $C > 0$. Hence

$$C|\{u > 0\} \cap \{|x| > R\}| \leq \int_{\{u>0\}} w(x) dx < \infty$$

Thus u has finite measure support and thus compact support as desired. □

2.2 L^1 Constrained Elliptic Problems

In this section we prove the compact support of L^1 constrained elliptic variational problems. The problem is as follows. Let $\Omega \subset \mathbb{R}^n$ be an unbounded subset with smooth boundary and

L a second order elliptic operator satisfying the same assumptions as in [6]. Specifically, let

$$L = -\sum_{i,j} a_{ij} \frac{\partial^2}{\partial x_i \partial x_j} + \sum_i a_i \frac{\partial}{\partial x_i} + a \tag{2.2.1}$$

where $a_{ij} \in C^1(\bar{\Omega}) \cap L^\infty(\Omega)$, $a_i, a \in L^\infty(\Omega)$. Additionally, we assume uniform ellipticity on bounded subsets, i.e. for every $r > 0$, there exists an $\alpha(r) > 0$ such that $(a_{ij}(x)) \succeq \alpha(r) I_n$ for $|x| \le r$. Finally, we also assume that a is uniformly bounded away from 0, i.e. $a \ge \delta > 0$.

Let β be a maximal monotone graph in $\mathbb{R}^2$ such that $\beta(0) = [\gamma^-, \gamma^+]$ with $\gamma^- < 0$ and $\gamma^+ > 0$.

We wish to extend the results in [6] by determining when the problem

$$f \in Lu + \mu(x)\beta(u) \tag{2.2.2}$$

on Ω with boundary data $u = \phi$ on $\partial\Omega$ has solutions with compact support.

Note that we may attempt to divide the entire problem by μ to obtain

$$(f/\mu(x)) \in (L/\mu(x))u + \beta(u)$$

Now if μ is bounded away from 0 and positive, then $L/\mu(x)$ will still be an elliptic operator and we are in a position to apply the result from [6]. We wish to extend this to the case where μ can be taken to vanish and be negative. However, we need μ to be large outside of a compact set and we lose uniqueness if μ can be negative.

Precisely, we prove the following

Theorem 2.2.1. *Assume that*

$$\phi \in C_c^2(\partial\Omega) \text{ and } \beta(\phi) \in L^\infty(\partial\Omega)$$

$$f \in L^\infty_{loc} \text{ and } \gamma^- < \liminf_{|x|\to\infty} f \le \limsup_{|x|\to\infty} f < \gamma^+$$

$$\mu \in L^\infty_{loc} \text{ and } \mu(x) \ge 1 \text{ for } x \ge R_0$$

Then all solutions $u \in H^2(\Omega)$ to the above variational problem have compact support. Moreover, if $\mu \ge 0$, then the solution exists and is unique.

A key lemma in the proof will be the following maximal principle

Lemma 2.2.1. *Let $u, v \in H^2(\Omega) \cap C^2(\partial\Omega)$, assume that $\mu \geq 0$, and let $f \in Lu + \mu(x)\beta(u)$ and $g \in Lv + \mu(x)\beta(v)$ with $f, g, \beta(u), \beta(v) \in L^\infty(\Omega)$ and $f \geq g$. Then if $u \geq v$ on $\partial\Omega$, $u \geq v$ a.e. on Ω.*

Proof. Consider the function $w = (v - u)_+ \in H^1(\Omega)$ (note that we can only guarantee that this function will be in $H^1(\Omega)$, not necessarily in $H^2(\Omega)$). We wish to show that $w = 0$. First we define $w^* \in H^1(\mathbb{R}^n)$ such that $w^* = w$ on Ω and $w^* = 0$ elsewhere. This function w^* will be in H^1 since w vanishes at the boundary of Ω and Ω has a smooth boundary. Additionally, extend L to all of $\mathbb{R}^n$ by setting it to be the negative Laplacian outside of Ω. Now we will show that $w^* = 0$. To do so we will show that w^* is a weak subsolution of L, i.e. $Lw^* \leq 0$ in a weak sense. Then the weak Harnack inequality implies that $w^* \leq 0$ (see [10] p.194).

Note first that because $a_{ij} \in C^1(\bar{\Omega})$, be can rewrite L in divergence form, i.e.

$$L = -\partial_i \cdot (a_{ij}\partial_j) + \sum_i \bar{a}_i \frac{\partial}{\partial x_i} + a \tag{2.2.3}$$

where $\bar{a}_j = a_j + \partial_i a_{ij}$.

Now let $q \in C_0^1(\mathbb{R}^n)$, $q \geq 0$ be a test function and integrate by parts to get

$$\langle q, Lw^* \rangle = \int_{\mathbb{R}^n} a_{ij} D_i q D_j w^* + \bar{a}_i D_i w^* q + aqw^* dx$$

where $a_{ij}, \bar{a}_i, a$ are as above within Ω and $a_{ij} = \delta_{ij}, \bar{a}_i = 0, a = 0$ outside of Ω. Notice further that the integral outside of Ω vanishes since w^* and Dw^* are 0 a.e. outside of Ω. So we have

$$\langle q, Lw^* \rangle = \int_{\Omega} a_{ij} D_i q D_j w^* + \bar{a}_i D_i w^* q + aqw^* dx$$

Moreover, since $w^* = (v - u)_+$ within Ω, we have that Dw^* and w^* are 0 whenever $v \leq u$ (at least a.e.). Thus we have that

$$\langle q, Lw^* \rangle = \int_{\{v>u\}} a_{ij} D_i q D_j (v - u) + \bar{a}_i D_i (v - u) q + aq(v - u) dx$$

We now integrate the first term by parts and use the definition of L to see that

$$\langle q, Lw^* \rangle = \int_{\{v>u\}} qL(v - u) dx + \int_{\partial\{v>u\}} q(\nu \cdot a_{ij} D_j (v - u)) dS$$

where ν is the outward normal to $\partial\{v > u\}$. This is valid since the assumptions on L (uniform ellipticity and C^1 coefficients) given in [6] imply that $u, v \in C^{1,\alpha}(\Omega)$ (see [10] Theorem 8.34), which means that the above region is smooth enough for integration by parts.

Note that since $\{v > u\}$ is the set $\{v - u > 0\}$, $D(v - u)$ is a non-negative multiple of the inward pointing normal. Hence, since a_{ij} is positive definite we see that the second integral above is non-positive. Thus we obtain

$$\langle q, Lw^* \rangle \leq \int_{\{v>u\}} qL(v-u)dx = \int_{\{v>u\}} q(g - f - \mu(x)(h - j))dx$$

where $h \in \beta(v)$ and $j \in \beta(u)$ (since $f \in Lu + \mu(x)\beta(u)$ and $g \in Lv + \mu(x)\beta(v)$). But on the set where we are integrating, $v > u$ which implies by the monotonicity of β, that $h \geq j$. Thus since $q \geq 0$ and $\mu \geq 0$, we get that

$$\langle q, Lw^* \rangle \leq \int_{\{v>u\}} q(g - f)dx \leq 0$$

Hence w^* is a weak subsolution of L and thus as remarked above, $w^* \leq 0$. Since we have by definition that $w^* \geq 0$, we see that $w^* = 0$ as desired. □

We now prove Theorem (2.2.1).

Proof. The argument presented in [6] applies to the present situation using the above maximum principle, provided that $\mu \geq 0$. The only difference is that the r_0 which is chosen to satisfy

$$\phi(x) = 0, \ f(x) \leq \gamma^+ - \epsilon \text{ for } |x| \geq r_0 \tag{2.2.4}$$

in [6] must also be chosen larger than the R_0 in our statement of the theorem.

Thus it is only left to consider the case where w is not necessarily positive. First, choose $\epsilon > 0$ let $R > R_0$ so large that $\phi(x) = 0$ and $\gamma^- + \epsilon < f(x) < \gamma^+ - \epsilon$ for $|x| > R$ (this can be done for small enough epsilon by assumption) and consider the domain $\Omega^* = \Omega \cap \{|x| > R\}$. Let $u \in H^2$ be a solution to the given variational problem. We first show that $u \in L^\infty_{loc}(\Omega^*)$.

To this end, we first extend u to u^* on the entire set $\{|x| > R\}$ by setting $u^* = 0$ outside of Ω. Then again we will have $u^* \in H^1$ since u vanishes on $(\partial\Omega) \cap \{|x| > R\}$. It will suffice to show that $u^* \in L^\infty_{loc}(\{|x| > R\})$.

A computation which is essentially the same as the one performed in the above lemma implies that $Lu_+^* \leq 0$ on $\{|x| > R\}$ (this requires that $f - \mu(x)\beta(u) \leq 0$ wherever $u > 0$ as $\mu(x) \geq 1$ and $f < \gamma^+$ for $|x| > R$).

For each point x with $|x| > R$ we choose a ball $B_\rho(x)$ about x which is still contained in $\{|x| > R\}$. We can now use again the weak harnack inequality (see [10] p.194) (as $u \in L^2$ since $u \in H^1$) and the analogous argument applied to u_- to conclude that $u \in L^\infty(B_\rho(x))$.

Now consider the domain $\Omega^* = \Omega \cap \{|x| > R'\}$ where $R' > R$. First we note that u is bounded on $\partial\Omega^*$. This follows since outside of a radius R, $u = 0$ on $\partial\Omega$ and on $\partial\{|x| > R'\}$, u is locally bounded and thus bounded since $\{|x| = R'\}$ is a compact set.

Now we proceed to construct a function $v \in C_c^2(\Omega^*)$ such that $g \in Lv + \mu(x)\beta(v)$ with $g \geq f$. Thus by the above maximum principle applied to Ω^*, $u \leq v$. Analogously we can construct $v \in C_c^2(\Omega^*)$ such that $u \geq v$. This will imply that u has compact support.

In particular, we construct v of the form

$$v(x) = \begin{cases} \frac{\lambda}{2}(|x| - R^*)^2 & \text{for } R' \leq |x| < R^* \\ 0 & \text{for } R^* \leq |x| \end{cases}$$

where λ and R^* are to be determined. Simple computations which are given in [6] imply that

$$Lv \geq -\lambda K' - \lambda K(R^* - |x|) + \frac{1}{2}\delta\lambda(R^* - |x|)^2$$

where $K' = \sup_\Omega \sum_i a_{ii}$, $K^2 = \sum_i \|a_i\|^2_{L^\infty(\Omega)}$ and $\delta > 0$ is such that $a \geq \delta$ (this is one of the assumptions in [6]).

We can now choose λ small enough, so that the above expression is greater than $-\epsilon$ uniformly in R^*. This is because the expression is a quadratic in $(|x| - R^*)$ with positive leading coefficient. Thus there is a minimal value that can be attained which is independent of R^*.

We then simply choose R^* large enough, so that $v \geq u$ on $\{|x| = R'\}$. This can be done since $v(x) = \frac{\lambda}{2}(R^* - R')^2$ on $\{|x| = R'\}$.

Now choose g such that $g \in Lv + \mu(x)\beta(v)$ where $v > 0$ and $g = Lv + \mu(x)\gamma^+$ where $v = 0$

($v \geq 0$ so this covers all cases). Then by definition of γ^+, $g \in Lv + \mu(x)\beta(v)$. Note that since v is a monotone graph, we have $g \geq Lv + \mu(x)\gamma^+$ everywhere. Now $Lv \geq -\epsilon$, $f < \gamma^+ - \epsilon$, and $\mu(x) \geq 1$ on Ω^* imply that $g \geq f$ on Ω^*. Combined with $v \geq u$ on $\{|x| = R'\}$, this implies that $v \geq u$ as desired.

The analogous argument with a subsolution concludes the proof that u must be compactly supported. □

Unfortunately, we don't obtain a bound on the support which is independent of u. In particular, the size of the support depends upon $\|u\|_{L^\infty(\partial\Omega^*)}$ which in turn can be controlled by the L^p norm of u ($p > 1$, by the weak Harnack inequality). Thus we cannot, in general, reduce the existence to a bounded domain. However, if the variational problem arises in the context of a minimization problem which allows one to control the L^p norm of the solution, then existence can be reduced to a bounded domain.

Finally, note that uniqueness fails if μ is allowed to be negative. Indeed, take any $u \in C_c^\infty$, $u \geq 0$ and define $\mu(x) = \Delta u$ if x is in the support of u and $\mu(x) = 1$ otherwise. Additionally, let β be the subdifferential of $|\cdot|$. Then it is easy to see that $0 \in -\Delta u + \mu(x)\beta(u)$. However, we also clearly have $0 \in -\Delta 0 + \mu(x)\beta(0)$. Hence the solution isn't unique in this case.

2.3 Numerical Results for L^1 Constrained Elliptic Problems

In this section we numerically investigate solutions to the L^1 constrained elliptic problem

$$\underset{u \in H^1}{\arg\min} \|\nabla u\|_2^2 - 2\langle f, u\rangle + \|w(x)u\|_1 \tag{2.3.1}$$

Specifically, we solve

$$\underset{u \in H_0^1(\Omega)}{\arg\min} \|\nabla u\|_2^2 - 2\langle f, u\rangle + \|w(x)u\|_1 \tag{2.3.2}$$

where Ω is the unit cube $[0,1]^2$. By making $w(x)$ large enough in relation to f, we can, by the above arguments, force the support of the solution to lie in Ω, and thus obtain a solution to the first problem by solving the second.

To solve the above problem we use a splitting scheme in combination with ADMM.

Specifically, we rewrite the problem as

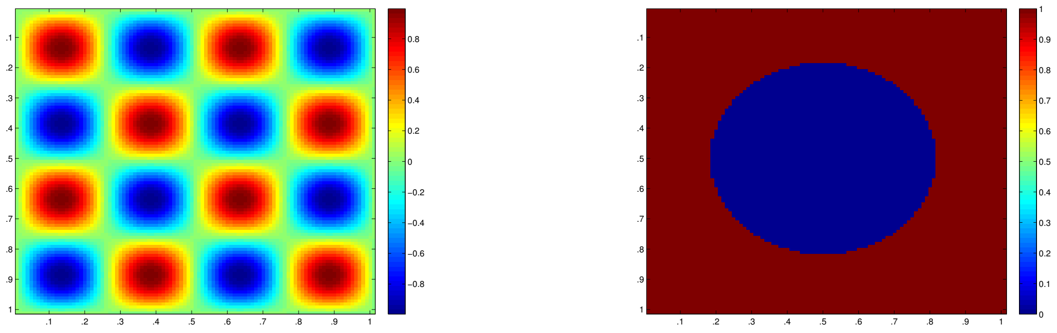
$$\operatorname*{arg\,min}_{u,v\in H_0^1(\Omega)} \|\nabla v\|_2^2 - 2\langle f,u\rangle + \|w(x)u\|_1 \tag{2.3.3}$$

subject to the constraint $u = v$, which we then solve using ADMM. This produces the following iteration

$$\begin{aligned} v_{n+1} &= \operatorname*{arg\,min}_{v\in H_0^1(\Omega)} \|\nabla v\|_2^2 + \frac{\mu}{2}\|v - u_n - \lambda_n\|_2^2 \\ u_{n+1} &= \operatorname*{arg\,min}_{u\in H_0^1(\Omega)} \|w(x)u\|_1 - 2\langle f,u\rangle + \frac{\mu}{2}\|v_{n+1} - u - \lambda_n\|_2^2 \\ \lambda_{n+1} &= \lambda_n + (u_{n+1} - v_{n+1}) \end{aligned}$$

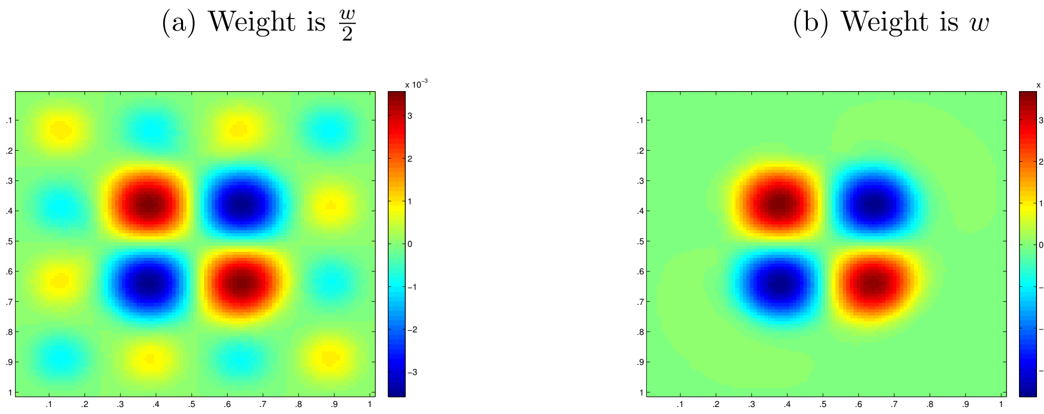
The first of these problems can be solved by solving the Poisson equation. The second minimizer is given in closed form by a pointwise shrink operator.

The results we obtain are as follows. In the first example, we let f and w be as below


(a) f

(b) w

Figure 2.1: Plots of f and the weight w


(a) Weight is $\frac{w}{2}$

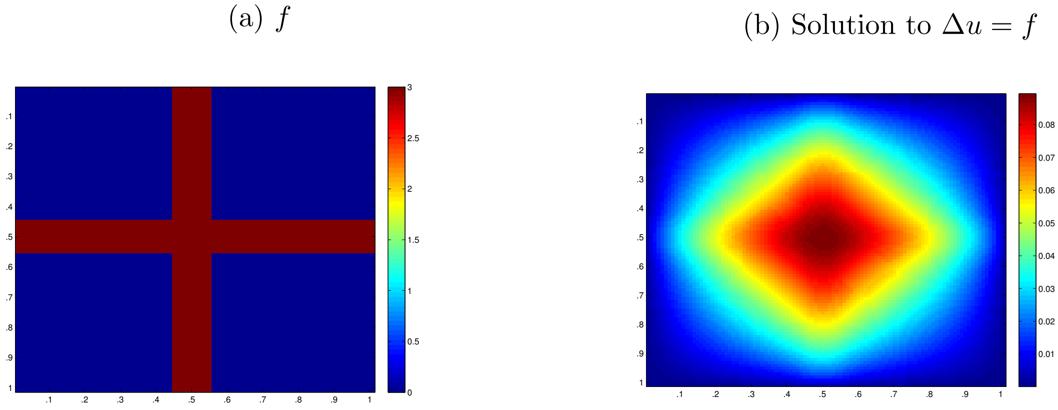
(b) Weight is w

Figure 2.2: Plots of the solution with weights $\frac{w}{2}$ and w

Figure (2.2) shows the results we obtained with two different scalings of the L^1 weight. Notice that $f = \sin(4\pi x)\sin(4\pi y)$ is an eigenfunction for the Dirichlet laplacian, so that the solution to the elliptic problem without the L^1 term is a scaled version of f (namely $\frac{1}{32\pi^2}f$).

We see that although we obtain compact support, the solution is very close to the solution of the laplacian within the circle, where there is no L^1 term.

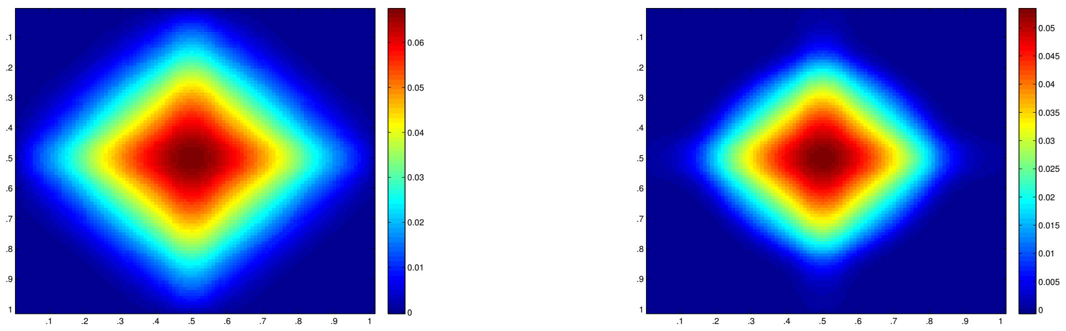
Now we let f be a function which is not an eigenfunction and use the same w as before. The function f and the solution to $\Delta u = f$ are given below.


(a) f

(b) Solution to $\Delta u = f$

Figure 2.3: Plots of new function f and the solution to $\Delta u = f$

In this case, we obtain for two different scalings of the L1 term:



(a) Weight is w

(b) Weight is $3w$

Figure 2.4: Plots of the solution with weights w and $3w$

Again, we see that although we obtain compact support, the solution is close to the solution of the Laplacian within the circle, where there is no L^1 term. We propose that the solutions to such L^1 constrained elliptic problems could be used as C_0^1 local approximations to the unconstrained elliptic problem.

2.4 Conclusion

In this chapter, we used compactness results and inequalities from Harmonic analysis to provide a novel existence proof for L^1 penalized eigenvalue problems. These new techniques apply to the case of a weighted L^1 norm as long as the weight is a radial increasing function. In addition, we prove that the optimizers of L^1 penalized eigenvalue problems and L^1 penalized elliptic problems have compact support. This later result extends work of Brezis to the setting of a weighted L^1 norm. Finally, we provide the results of numerical experiments and propose that these variational problems could be used to construct local approximate solutions of elliptic PDEs.

CHAPTER 3

Accelerated Gradient Descent with Orthogonality Constraints

In this chapter we develop numerical methods for solving smooth optimization problems on the set of $n \times k$ orthonormal matrices. Specifically, we consider problems of the form

$$\arg\min_{X^T X = I_k} f(X) \tag{3.0.1}$$

where X is an $n \times k$ matrix. The set $\{X : X^T X = I_k\}$ is called the Stiefel manifold and we denote it by $S_{n,k}$. Optimization problems of this form have a wide range of applicability including electronic structure and eigensystem calculations. Often the objectives in these cases are ill-conditioned, by which we mean that the Hessian of f at the (local) minimizer is ill-conditioned. This leads to very slow convergence rates for gradient descent type methods.

We compare this to solving μ-strongly convex and L-smooth optimization problems in Euclidean space. In this situation, gradient descent is not an optimal first order method. The number of iterations which gradient descent requires to reach a given level of accuracy is $O(\kappa)$, where $\kappa = L/\mu$ is the condition number of the problem (the eigenvalues of the Hessian are bounded between α and L). However, asymptotically optimal methods such as Nesterov's accelerated gradient descent require only $O(\sqrt{\kappa})$ iterations to achieve the same level of accuracy.

Considering optimization on manifolds again, near a local minimum we expect our function f to be approximately smooth and strongly convex. This suggests that we can achieve much faster convergence by appropriately modifying Nesterov's accelerated gradient descent for optimization on manifolds.

We begin the chapter by describing gradient descent on the Stiefel manifold. Here we give the convergence results that we hope to improve with our method. Then we describe accelerated gradient descent for convex functions on $\mathbb{R}^n$, which we want to generalize. The remainder of the chapter describes the problems which must be solved to do this and gives a detailed description of the method. Finally we conclude with numerical experiments which empirically demonstrate the desired rate of convergence.

3.1 Gradient Descent on Riemannian Manifolds

The goal of this section is to generalize the gradient descent iteration

$$x_{n+1} = x_n - \gamma_n \nabla f(x_n) \tag{3.1.1}$$

to Riemannian manifolds. This problem has been extensively studied in the optimization community, see [1] for a comprehensive treatment. The general idea is that since subtracting the gradient doesn't have meaning on a manifold, we instead follow a curve whose derivative is a descent direction for f.

Before giving the analogue of (3.1.1) on a Riemannian manifold, we must begin by briefly introducing some concepts and notation from differential geometry. For further reference on this topic see [19] or [1].

Let M be a smooth manifold and $x \in M$. We denote the tangent space of M at x by T_xM and the dual tangent space by $(T_xM)^*$. We denote the tangent bundle of M, i.e. the space of pairs (x, v) with $x \in M$ and $v \in T_xM$, by TM, and likewise the dual tangent bundle by $(TM)^*$.

Suppose f is a C^1 function on M. Then the derivative of f at $x \in M$, which we denote by $\nabla f(x)$ is naturally an element of $(T_xM)^*$. In particular, it is the linear functional which maps a tangent vector $v \in T_xM$ to the directional derivative of f in the direction v.

If M is a Riemannian manifold, then each tangent space T_xM is equipped with a positive definite inner product $g : T_xM \times T_xM \to \mathbb{R}$. Because it is positive definite, g induces a norm on the tangent space

$$\|v\|_g^2 = g(v, v) = g_{ij}v^i v^j \tag{3.1.2}$$

For the last expression above we have fixed a coordinate system and g_{ij} are the (covariant) components of g in this coordinate system (we are using the Einstein summation notation).

We also have a dual norm on the dual space

$$\|w\|_{g*} = \sup_{\|v\|_g=1} \langle w, v\rangle = g^{ij} w_i w_j \tag{3.1.3}$$

Here g^{ij} satisfing $g^{ij}g_{jk} = \delta^i_k$ are the (contravariant) components of g.

Additionally, the inner product g provides an isomorphism $\phi_g : (T_xM)^* \to T_xM$ with the property that

$$\|w\|^2_{g*} = \langle w, \phi_g(w)\rangle = \|\phi_g(w)\|^2_g \tag{3.1.4}$$

In terms of the metric, the map ϕ_g is given by raising the indices of w, i.e.

$$\phi_g(w)^i = g^{ij} w_j \tag{3.1.5}$$

and its inverse $\phi_g^{-1} : T_xM \to (T_xM)^*$ is given by lowering the indices of v, i.e.

$$\phi_g^{-1}(v)_i = g_{ij} v^j \tag{3.1.6}$$

Given a smooth curve $c : [0, 1] \to M$, the length of the curve is defined by

$$l(c) = \int_0^1 \|c'(t)\|_g dt \tag{3.1.7}$$

which allows us to define the distance between points $x, y \in M$ as follows

$$d(x, y) = \inf_{\substack{c:[0,1]\to M \\ c(0)=x,\ c(1)=y}} l(c) \tag{3.1.8}$$

Note that the minimizer in the above expression is not unique, even if $d(x, y)$ is very small. This is due to the fact that the length of a curve is invariant under reparametrizations. The unit speed geodesic between x and y is given by

$$\operatorname*{arg\,min}_{\substack{c:[0,d(x,y)]\to M \\ c(0)=x,\ c(d(x,y))=y}} \int_0^{d(x,y)} \|c'(t)\|^2_g dt \tag{3.1.9}$$

which exists and is unique as long as x and y are sufficiently close. The minimizing curve c will satisfy the geodesic equations

$$\frac{d^2c^i}{dt^2} + \Gamma^i_{kl}\frac{dc^k}{dt}\frac{dc^l}{dt} = 0 \tag{3.1.10}$$

where Γ^i_{kl} are the Christoffel symbols.

Assuming that the geodesic equations can be solved globally in time (which is true for the Stiefel manifold that we are interested in) we can define the exponential map $\exp_x : T_xM \to M$ as follows

$$\exp_x(v) = c_v(1) \tag{3.1.11}$$

where c_v is the (unique) unit speed geodesic satisfying $c_v'(0) = v$.

Let $x \in M$ and $f \in C^2(M)$. We define the following quadratic form on T_xM, called the Hessian of f (which generalizes the Hessian in Euclidean space), as follows

$$Hf(x)(v) = \left.\frac{d^2 f(\exp_x(tv))}{dt^2}\right|_{t=0} \tag{3.1.12}$$

The Hessian can be given in matrix form as follows

$$Hf(x)_{ij} = \frac{\partial^2 f}{\partial_i \partial_j} - \Gamma^k_{ij}\frac{\partial f}{\partial_k} \tag{3.1.13}$$

The condition number of $Hf(x)$ is the condition number of the matrix representation with respect to an orthonormal basis of T_xM. Alternatively, it is the ratio

$$\kappa(Hf(x)) = \frac{\sup_{\|v\|_g=1} Hf(x)(v)}{\inf_{\|v\|_g=1} Hf(x)(v)} \tag{3.1.14}$$

We now show how to generalize gradient descent using the concept of a retraction, which is central to the problem of optimizing functions on manifolds.

3.1.1 Retractions and Gradient Descent

The issue with the gradient descent iteration (3.1.1) when x is constrained to a manifold is that the linear operation of subtracting $\gamma_n \nabla f(x_n)$ doesn't make sense. Instead we will follow curves which have the "correct" derivative at the current iterate x_n. To do this, we must specify a family of curves to follow and also clarify what the "correct" direction is.

The first of these issues is dealt with via the notion of a retraction.

Definition 3.1.1. *Let M be a (smooth) manifold. A retraction on M is a (smooth) map $R : TM \to M$ (here TM denotes the tangent bundle of M) satisfying for all $x \in M$ and*

$v \in T_xM$.

$$R(x,0) = x \tag{3.1.15}$$

$$\left.\frac{d}{dt}\right|_{t=0} R(x,tv) = v \tag{3.1.16}$$

(Here I write $R(x,v)$ for the image of the point $(x,v) \in TM$ under R.)

Intuitively, given any $x \in M$ and any $v \in T_xM$, a retraction gives a curve on the manifold starting at x and moving initially in the direction v.

The notion of a retraction is a central concept in optimization on manifolds. The efficiency of many optimization methods depend crucially on the choice of a retraction which can be efficiently calculated. Later, we will give detailed descriptions of some retractions on the Stiefel manifold and how to compute them.

However, first we will discuss the second issue of finding the "correct" direction and introduce gradient descent on Riemannian manifolds as an algorithm template which requires the specification of a retraction.

We begin by calculating the derivative of the function f along the curve defined by our retraction. I.e. we calculate

$$\left.\frac{d}{dt}\right|_{t=0} f(R(x,tv)) \tag{3.1.17}$$

Using the chain rule (note that $\langle\cdot,\cdot\rangle$ represents the pairing between T_xM and its dual $(T_xM)^*$), we get

$$\left.\frac{d}{dt}\right|_{t=0} f(R(x,tv)) = \langle v, \nabla f(x)\rangle \tag{3.1.18}$$

Now the Riemannian structure on M comes into play. It gives us measure of the length of the tangent vector v and the dual vector $\nabla f(x)$. In order to generalize gradient descent we wish to maximize the objective decrease subject to $\|v\|_g = \|\nabla f(x)\|_{g^*}$. We note that this is achieved by setting $v = -\phi_g(\nabla f(x))$ (ϕ_g is the isomorphism $(T_xM)^* \to T_xM$ induced by g) since

$$|\langle v, \nabla f(x)\rangle| \leq \|v\|_g \|\nabla f(x)\|_{g^*} = \|\nabla f(x)\|_{g^*}^2 \tag{3.1.19}$$

by the definition of the dual norm, and

$$\langle -\phi_g(\nabla f(x)), \nabla f(x)\rangle = -\|\nabla f(x)\|_{g^*}^2 \tag{3.1.20}$$

by the definition of $\phi(g)$.

This leads to the following generalization of gradient descent (3.1.1) on Riemannian manifolds.

$$x_{n+1} = R(x_n, -\gamma_n \phi_g(\nabla f(x_n))) \tag{3.1.21}$$

3.1.2 Step Size Selection and Convergence Properties

Just as in $\mathbb{R}^n$, the selection of the step size γ_n is very important in ensuring that the gradient descent iteration (3.1.21) has desirable convergence properties. The approach that we will take is to choose the step size to satisfy a sufficient decrease condition called the Armijo rule (with parameter 1/2, see [2])

$$f(x_{n+1}) \leq f(x_n) - \frac{1}{2}\gamma_n \|\nabla f(x_n)\|_{g^*}^2 \tag{3.1.22}$$

We simply note here that if f is smooth, then (3.1.22) can always be satisfied by making γ_n sufficiently small. For instance, if $f : \mathbb{R}^n \to \mathbb{R}$ has L-lipschitz gradient, then $\gamma_n = 1/L$ will work. In practice, we use a line search to find an appropriate γ_n (the precise details can be found in [1]).

The gradient descent method (3.1.21) enjoys similar convergence properties as its counterpart in $\mathbb{R}^n$. Namely, in [1], the following theorem is proved concerning its convergence.

Theorem 3.1.1 (Theorem 4.3.1 in [1])**.** *Let a sequence of points x_k in M be generated by iteration (3.1.21) with step size γ_n satisfying (3.1.22). Then every accumulation point of x_k is a critical point of f.*

A simple corollary of this result is that if M is compact, then $\|\nabla f(x_k)\|_g \to 0$ as $k \to \infty$. This is promising since it implies that the iteration (3.1.21) is guaranteed to converge in the sense that the gradient norm can be made arbitrarily small.

There is also a convergence rate result which holds once the iterates are close enough to a local minimizer x^*. This result depends on the Hessian of the function at x^*, which is analogous to the situation in $\mathbb{R}^n$.

Theorem 3.1.2 (Theorem 4.5.6 in [1])**.** *Let a sequence of points x_k in M be generated by iteration (3.1.21) with step size given by the Armijo rule (see [2]). Assume additionally that x_k converges to a local minimizer x^*. Then for some $c > 0$, there exists an N such that for $n > N$ we have*

$$f(x_{n+1}) - f(x^*) \leq (1 - c\kappa^{-1})(f(x_n) - f(x^*)) \tag{3.1.23}$$

where κ is the condition number of the Hessian of f at x^*.

This theorem implies that the number of iterations required to reach a certain accuracy is $O(\kappa)$ once we are close enough to a local minimum. This result is good if the functions we are optimizing are well-conditioned near their local minima. However, many applications of interest, including electronic structure calculations and eigensystem problems do not have this property. In these situations, the number of iterations required for gradient descent makes the approach of manifold optimization unfeasible.

Methods to overcome this problem have been proposed, most notably conjugate gradient type methods (see [8], [37]). However, these methods are much more complicated and expensive on manifolds and have not, to our knowledge, achieved a satisfactory improvement for ill-conditioned problems.

We take a different approach to reducing the iteration count to $O(\sqrt{\kappa})$. Instead of attempting to generalize conjugate gradient methods, we generalize accelerated first order methods for convex optimization, in particular Nesterov's gradient descent, as we will discuss in a later section.

3.2 Retractions on the Stiefel Manifold

In order to convert (3.1.21), or any algorithm template which uses retractions, into an algorithm, we must specify the retraction R and provide a method for calculating it. In this section, we define a collection of retractions on the Stiefel manifold and give algorithms for computing them. These retractions will form the backbone of our optimization method.

3.2.1 Geometry of the Stiefel Manifold

The Stiefel manifold $S_{n,k}$ is the set of $n \times k$ orthonormal matrices, i.e.

$$S_{n,k} = \{X \in \mathbb{R}^{n\times k} : X^T X = I_k\}$$

We begin by describing the Riemannian metric which we put on the Stiefel manifold and giving formulas for calculating inner products, raising and lowering indices and geodesics. Of course, the metric which we consider is not unique as any diffeomorphism of M onto itself provides a new metric (which is the same only if the diffeomorphism is an isometry).

In fact, there are two metrics commonly put on the Stiefel manifold in the literature. One is obtained by viewing $S_{n,k} \subset \mathbb{R}^{nk}$ and considering the metric induced by the ambient space $\mathbb{R}^{nk}$. The other, called the canonical metric and which we will be considering for the remainder of this chapter, is obtained by viewing $S_{n,k} = O(n)/O(n-k)$ as the quotient of the orthogonal group $O(n)$ by the right action of $O(n-k)$. Specifically, the action is given by right multiplication by

$$\begin{bmatrix} I_{k\times k} & 0_{k\times n} \\ 0_{n\times k} & O_{(n-k)\times(n-k)} \end{bmatrix} \tag{3.2.1}$$

where $O_{(n-d)\times(n-d)} \in O(n-k)$. This induces a quotient metric on $S_{n,k}$. For more details on the former metric and the differences between these two viewpoints, see [8].

Before we describe the metric in more detail, we must fix a representation of the elements of $S_{n,k}$ and its tangent and dual tangent space. Throughout, the elements of $S_{n,k}$ will be represented by $n \times k$ orthonormal matrices (even though our metric is induced by viewing $S_{n,k}$ as a quotient $O(n)/O(n-k)$). The tangent space at a point $X \in S_{n,k}$ is then naturally identified with the set $T_X = \{V \in \mathbb{R}^{n\times k} : V^T X + X^T V = 0\}$. We choose to represent the dual space by elements of the same set, i.e. $(T_X)^* = \{W \in \mathbb{R}^{n\times k} : W^T X + X^T W = 0\}$, with the pairing between T_X and $(T_X)^*$ given by $\langle V, W\rangle = \mathrm{Tr}(V^T W)$ (i.e. the usual inner product on $\mathbb{R}^{nk}$).

Using these representations, the metric on $S_{n,k}$ is given by (see [8])

$$g(Y, Z) = \mathrm{Tr}\left(Y^T \left(I - \frac{1}{2} X X^T\right) Z\right) \tag{3.2.2}$$

where $Y, Z \in T_X S_{n,k}$. The formula for the inner product on the dual space is

$$g^*(Y,Z) = \mathrm{Tr}\left(Y^T\left(I - \frac{1}{2}XX^T\right)^{-1} Z\right) \tag{3.2.3}$$

for $Y, Z \in (T_X S_{n,k})^*$. Since X is orthonormal, it follows that XX^T is a projection, and we thus have

$$\left(I - \frac{1}{2}XX^T\right)^{-1} = \left(I + XX^T\right) \tag{3.2.4}$$

So we can rewrite the dual space inner product as

$$g^*(Y,Z) = Tr\left(Y^T\left(I + XX^T\right)Z\right) \tag{3.2.5}$$

Finally, the maps corresponding to raising and lowering the indices are

$$\phi_g(W) = \left(I - \frac{1}{2}XX^T\right)^{-1} W = \left(I + XX^T\right)W \tag{3.2.6}$$

and

$$\phi^g(V) = \left(I - \frac{1}{2}XX^T\right)V \tag{3.2.7}$$

respectively.

3.2.2 The Geodesic Retraction

The advantage of using canonical metric, i.e. the metric induced by the quotient structure of $S_{n,k}$, is that geodesics can be computed using the matrix exponential. In fact, the constant-speed geodesic starting at $X \in S_{n,k}$ and moving initially in the direction $V \in T_X S_{n,k}$ is given by (see [8] for details)

$$X(t) = \exp\left(t(VX^t - XV^t + XV^tXX^t)\right)X \tag{3.2.8}$$

This leads naturally to the geodesic retraction on $S_{n,k}$, defined by

$$R_G(X,V) = \exp\left(VX^t - XV^t + XV^tXX^t\right)X \tag{3.2.9}$$

Plugging this retraction into the gradient descent iteration (3.1.21) gives

$$X_{n+1} = R_G(X_n, -\gamma_n\phi_g(\nabla f(X_n))) \tag{3.2.10}$$

and it remains to explain how to compute $\phi_g(\nabla f(X))$.

We wish to calculate this in terms of the component-wise derivative of f, i.e. in terms of

$$G_{ij} = \frac{\partial}{\partial X_{ij}} f$$

Luckily this is relatively straightforward. We must first project G onto the dual tangent space $T_X S_{n,k}$ and then apply ϕ_g to raise the indices.

Since the pairing between the dual and tangent spaces under our representation is just the inner product in $\mathbb{R}^{nk}$, the projection of G onto the dual tangent space is just a projection in Euclidean space onto the set $(T_X)^* = \{W \in \mathbb{R}^{n\times k} : W^T X + X^T W = 0\}$. This is easily seen to be

$$P_{(T_X)^*}(G) = G - \frac{1}{2}X(X^T G + G^T X) \tag{3.2.11}$$

Applying the map ϕ_g and calculating, we see that

$$\phi_g(\nabla F(X)) = \left(I + XX^T\right)\left(G - \frac{1}{2}X(X^T G + G^T X)\right) = G - XG^T X \tag{3.2.12}$$

Finally, plugging this back into (3.2.9), we obtain the following gradient descent iteration

$$X_{n+1} = \exp\left(-\gamma_n(G_n X_n^T - X_n^T G_n)\right)X_n \tag{3.2.13}$$

Since the matrix $GX^T - XG^T$ has rank $2k$, this exponential can be calculated by diagonalizing a $2k \times 2k$ antisymmetric matrix, as shown in [8].

3.2.3 Approximate Geodesic Retractions

By approximating the exponential in (3.2.9) we can obtain retractions which can be computed more efficiently. It is these approximations to the geodesic retraction, first considered in [36], which will be used in our method.

The idea behind these approximate geodesic retractions is to replace the matrix exponential in (3.2.9) by its symmetric Padé approximant

$$e^x \approx P_{r,r}(x) = \left(\sum_{n=0}^{r} \frac{(2r-n)!r!}{(2r)!(r-n)!}\frac{x^n}{n!}\right) \Big/ \left(\sum_{n=0}^{r} \frac{(2r-n)!r!}{(2r)!(r-n)!}\frac{(-x)^n}{n!}\right) \tag{3.2.14}$$

The approximant $P_{r,r}$ is an order $2r$ approximation to the exponential at $x = 0$. Moreover, it has the remarkable property that $P_{r,r}(x)P_{r,r}(-x) = 1$. This property is vital if we want to use our approximation to define a retraction. This is because any retraction must certainly satisfy $R(X, V) \in S_{n,k}$, which is not necessarily true if we replace the exponential in (3.2.9) by some approximation. However, the mentioned property implies that if M is an antisymmetric matrix and $P_{r,r}(M)$ exists (i.e. the matrix in the denominator of the above expression is invertible), then in fact

$$P_{r,r}(M)P_{r,r}(M)^T = P_{r,r}(M)P_{r,r}(M^T) = P_{r,r}(M)P_{r,r}(-M) = I \tag{3.2.15}$$

so that $P_{r,r}(M)$ is an orthogonal transformation. Now we can verify that if $V \in T_X M$, then the matrix $VX^t - XV^t + XV^tXX^t$ in the argument of the exponential in (3.2.9) is antisymmetric. This means that approximating this exponential by $P_{r,r}$ does in fact result in a valid retraction. So we define the approximate geodesic retraction of order $2r$

$$R_r(X, V) = P_{r,r}(VX^t - XV^t + XV^tXX^t)X \tag{3.2.16}$$

Note that if we wish to apply the approximate retraction to a dual vector or function gradient, we simply obtain (analogous to the computation of the previous section)

$$R_r(X, \phi_g(W)) = P_{r,r}(WX^T - X^TW)X \tag{3.2.17}$$

and

$$R_r(X, \phi_g(\nabla f(X))) = P_{r,r}(GX^T - X^TG)X \tag{3.2.18}$$

where G is the component wise derivative of F. The corresponding gradient descent iteration is

$$X_{n+1} = P_{r,r}\left(-\gamma_n(G_nX_n^T - X_n^TG_n)\right)X_n \tag{3.2.19}$$

We now discuss how these approximate geodesic retractions can be computed efficiently. For the remainder of this section, we will focus on calculating $R_r(X, \phi_g(W))$. For numerical stability reasons we must first project G onto the dual tangent space when calculating $R_r(X, \phi_g(\nabla f(X)))$ (instead of just using formula (3.2.18)), as first noted in [14]. The calculation of $R_r(X, V)$ is similar but a bit more computationally expensive and will not appear in our method.

The key is to utilize the low rank of $WX^T - XW^T$. Indeed, note that

$$WX^T - XW^T = UZ^T \tag{3.2.20}$$

where $U = [W, X]$ and $Z = [X, -W]$.

This allows us to rewrite $P_{r,r}(WX^T - XW^T) = P_{r,r}(UZ^T)$ as

$$\left(I + U\left(\sum_{n=0}^{r-1} C_{n+1,r} Z^T (UZ^T)^n\right)\right)\left(I - U\left(\sum_{n=0}^{r-1}(-1)^n C_{n+1,r} Z^T (UZ^T)^n\right)\right)^{-1} \tag{3.2.21}$$

where $C_{n,r} = \frac{(2r-n)!r!}{(2r)!(r-n)!n!}$.

We now use the following variant of the Sherman-Morrison-Woodbury formula (initially introduced in [32]).

Proposition 3.2.1. *If $I - T^TU$ is invertible, then $I - UT^T$ is invertible and*

$$(I - UT^T)^{-1}(I + US^T)X = X + U(I - T^TU)^{-1}(T + S)^T X \tag{3.2.22}$$

Proof. Recall that by the Sherman-Morrison-Woodbury formula (see [13]), we have that

$$(I - UT^T)^{-1} = I + U(I - T^TU)^{-1}T^T \tag{3.2.23}$$

Multiplying this by $(I + US^T)X$ and expanding, we see get

$$(I - UT^T)^{-1}(1 + US^T)X = X + US^TX + U(I - T^TU)^{-1}T^T(I + US^T)X \tag{3.2.24}$$

which we rewrite as

$$X + U(S^T + (I - T^TU)^{-1}T^T(I + US^T))X \tag{3.2.25}$$

Thus we will be done if we can show that

$$(I - T^TU)^{-1}(T + S)^T = S^T + (I - T^TU)^{-1}T^T(I + US^T) \tag{3.2.26}$$

So we multiply the right side of this equation by $(I - T^TU)$ to obtain

$$S^T - T^TUS^T + T^T(I + US^T) = S^T + T^T$$

This proves (3.2.26) and we are done. □

Applying (3.2.1) with

$$T = \sum_{n=0}^{r-1} (-1)^n C_{n+1,r} Z(U^T Z)^n \tag{3.2.27}$$

and

$$S = \sum_{n=0}^{r-1} C_{n+1,r} Z(U^T Z)^n \tag{3.2.28}$$

we can calculate $f_{r,r}(-UW^T)X$ via an $(n \times 2k)(2k \times k)$ matrix product, a $(2k \times n)(n \times k)$ matrix product, a $2k \times 2k$ inversion, and $2r$ $2k \times 2k$ matrix products using a total of $4nk^2 + O(rk^3)$ floating point operations.

The special case of $r = 1$ is particularly efficient and was first considered in [36]. In this situation we obtain the following formula.

$$R_1(X, \phi_g(W)) = \left(I - \frac{1}{2}(WX^T - X^T W)\right)^{-1} \left(I + \frac{1}{2}(WX^T - X^T W)\right) X \tag{3.2.29}$$

which, by the Sherman-Morrison-Woodbury formula is just (with $U = [W, X]$ and $Z = [X, -W]$)

$$R_1(X, \phi_g(W)) = X + 2U(I - Z^T U)^{-1} Z^T X \tag{3.2.30}$$

since $T = S = Z$ when $r = 1$.

In [37], the approximate retraction for general r is considered, but it is not noted that this higher order approximation can also be efficiently calculated. To the best of our knowledge, we are the first to note that retractions based on higher order Padé approximants can also be efficiently calculated.

3.3 Accelerated First Order Methods in Euclidean Space

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable convex function. We say that f is μ-strongly convex if

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|x - y\|_2^2 \tag{3.3.1}$$

We also say that f is L-smooth if

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|x - y\|_2^2 \tag{3.3.2}$$

One way of thinking about these definitions is that μ-strong convexity implies that the eigenvalues of the Hessian of f at every point are $\geq \mu$ and L-smoothness implies that the eigenvalues are $\leq L$.

In his seminal paper [24], Nesterov introduced first-order methods which achieve the asymptotically optimal objective error for the class of L-smooth convex functions and for the class of L-smooth and μ-strongly convex functions. These methods rely on a 'momentum step' and take the following form

$$x_0 = y_0,\ x_{n+1} = y_n - \gamma_n \nabla f(y_n),\ y_{n+1} = x_{n+1} + \alpha_n(x_{n+1} - x_n) \tag{3.3.3}$$

The choice of γ_n and α_n depend on whether the function f is strongly convex (as opposed to only convex and L-smooth), and also on the precise parameters μ and L.

If f is μ-strongly convex and L-smooth, then setting $\alpha_n = \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$ and $\gamma_n = 1/L$ produces the asymptotically optimal objective error of $O((1-\sqrt{\frac{\mu}{L}})^{-n})$ (compared with $O((1-\frac{\mu}{L})^{-n})$ for gradient descent), as the following theorem shows.

Theorem 3.3.1. *Assume that f is μ-strongly convex and L smooth. Let x^* be the minimizer of f. If we let $\alpha_n = \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$ and $\gamma_n = 1/L$ in (3.3.3), then we have that*

$$f(x_n) - f(x^*) \leq 2\left(1 - \sqrt{\frac{\mu}{L}}\right)^n (f(x_0) - f(x^*)) \tag{3.3.4}$$

Although proofs of this theorem are well-known, we present our own proof (adapting the argument in [33] to the strongly convex case) for completeness.

Proof. We consider the Lyapunov function

$$J_n = \left(1 - \sqrt{\frac{\mu}{L}}\right)^{-n} \left[(f(x_n) - f(x^*)) + \frac{1}{2}\|\sqrt{\mu}(y_n - x^*) + \sqrt{L}(y_n - x_n)\|_2^2\right] \tag{3.3.5}$$

We proceed to show that $\Delta J_n = J_{n+1} - J_n \leq 0$. To make this calculation simpler, we set $J_n^1 = (f(x_n) - f(x^*))$ and $J_n^2 = \frac{1}{2}\|\sqrt{\mu}(y_n - x^*) + \sqrt{L}(y_n - x_n)\|_2^2$. Then, by setting $C = \left(1 - \sqrt{\frac{\mu}{L}}\right)^{-1}$, we see that

$$\Delta J_n = C^n(CJ_{n+1}^1 - J_n^1 + CJ_{n+1}^2 - J_n^2) = C^n((C-1)J_{n+1}^1 + \Delta J_n^1 + (C-1)J_n^2 + C\Delta J_n^2) \tag{3.3.6}$$

As $C > 0$, it suffices to show that

$$(C-1)J^1_{n+1} + \Delta J^1_n + (C-1)J^2_n + C\Delta J^2_n \leq 0 \tag{3.3.7}$$

We now consider each of these terms separately. First we see that

$$J^1_{n+1} = (f(x_{n+1}) - f(x^*)) = (f(x_{n+1}) - f(y_n)) + (f(y_n) - f(x^*)) \tag{3.3.8}$$

We now use the L-smoothness of f and the fact that $x_{n+1} - y_n = -(1/L)\nabla f(y_n)$ to conclude that

$$f(x_{n+1}) - f(y_n) \leq -\frac{1}{2L}\|\nabla f(y_n)\|_2^2 \tag{3.3.9}$$

The strong convexity of f implies that

$$f(y_n) - f(x^*) \leq \nabla f(y_n) \cdot (y_n - x^*) - \frac{\mu}{2}\|y_n - x^*\|_2^2 \tag{3.3.10}$$

so that

$$J^1_{n+1} \leq -\frac{1}{2L}\|\nabla f(y_n)\|_2^2 + \nabla f(y_n) \cdot (y_n - x^*) - \frac{\mu}{2}\|y_n - x^*\|_2^2 \tag{3.3.11}$$

Likewise $\Delta J^1_n = f(x_{n+1}) - f(x_n) = (f(x_{n+1}) - f(y_n)) + (f(y_n) - f(x_n))$, so that

$$\Delta J^1_n \leq -\frac{1}{2L}\|\nabla f(y_n)\|_2^2 + \nabla f(y_n) \cdot (y_n - x_n) - \frac{\mu}{2}\|y_n - x_n\|_2^2 \tag{3.3.12}$$

Now we expand

$$J^2_n = \frac{1}{2}\|\sqrt{\mu}(y_n - x^*) + \sqrt{L}(y_n - x_n)\|_2^2 \tag{3.3.13}$$

as

$$J^2_n = \frac{\mu}{2}\|y_n - x^*\|_2^2 + \frac{L}{2}\|y_n - x_n\|_2^2 + \sqrt{\mu L}(y_n - x^*) \cdot (y_n - x_n) \tag{3.3.14}$$

Finally, we consider ΔJ^2_n. We set $t_n = \sqrt{\mu}(y_n - x^*) + \sqrt{L}(y_n - x_n)$ and note that since $J^2_n = \frac{1}{2}\|t_n\|_2^2$ we have

$$\Delta J^2_n = \Delta t_n \cdot t_n + \frac{1}{2}\|\Delta t_n\|_2^2 \tag{3.3.15}$$

We calculate $\Delta t_n = t_{n+1} - t_n$ as follows

$$t_{n+1} - t_n = \sqrt{\mu}(y_{n+1} - y_n) + \sqrt{L}(y_{n+1} - y_n) - \sqrt{L}(x_{n+1} - x_n) \tag{3.3.16}$$

Using

$$y_{n+1} - y_n = (y_{n+1} - x_{n+1}) + (x_{n+1} - y_n) = \alpha_n(x_{n+1} - x_n) - \frac{1}{L}\nabla f(y_n)$$

and

$$x_{n+1} - x_n = (x_{n+1} - y_n) + (y_n - x_n) = -\frac{1}{L}\nabla f(y_n) + (y_n - x_n)$$

we see that (since $\alpha_n = \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$)

$$t_{n+1} - t_n = -(\sqrt{\mu} + \sqrt{L})\frac{1}{L}\nabla f(y_n) + (\sqrt{L} - \sqrt{\mu})(x_{n+1} - x_n) - \sqrt{L}(x_{n+1} - x_n) \tag{3.3.17}$$

and so

$$\begin{aligned} t_{n+1} - t_n &= -(\sqrt{\mu} + \sqrt{L})\frac{1}{L}\nabla f(y_n) + \sqrt{\mu}\frac{1}{L}\nabla f(y_n) - \sqrt{\mu}(y_n - x_n) \\ &= -\frac{1}{\sqrt{L}}\nabla f(y_n) - \sqrt{\mu}(y_n - x_n) \end{aligned} \tag{3.3.18}$$

Thus we see that (by expanding out equation (3.3.15))

$$\begin{aligned} \Delta J_n^2 = &-\frac{\sqrt{\mu}}{\sqrt{L}}\nabla f(y_n)\cdot(y_n - x^*) - \mu(y_n - x_n)\cdot(y_n - x^*) \\ &- \nabla f(y_n)\cdot(y_n - x_n) - \sqrt{\mu L}\|y_n - x_n\|_2^2 \\ &+ \frac{1}{2L}\|\nabla f(y_n)\|_2^2 + \frac{\mu}{2}\|y_n - x_n\|_2^2 + \frac{\sqrt{\mu}}{\sqrt{L}}\nabla f(y_n)\cdot(y_n - x_n) \end{aligned} \tag{3.3.19}$$

Combining equations (3.3.11), (3.3.12), (3.3.14), and (3.3.19) with equation (3.3.7), collecting all of the terms, and noting that (recall that $C = \left(1 - \sqrt{\frac{\mu}{L}}\right)^{-1}$)

$$(C-1) = C\frac{\sqrt{\mu}}{\sqrt{L}} \tag{3.3.20}$$

$$C\left(\frac{\sqrt{\mu}}{\sqrt{L}} - 1\right) + 1 = 0 \tag{3.3.21}$$

$$(C-1)\sqrt{\mu L} - C\mu = C\frac{\sqrt{\mu}}{\sqrt{L}}\sqrt{\mu L} - C\mu = 0 \tag{3.3.22}$$

and

$$\frac{(C-1)L}{2} + C\left(\frac{\mu}{2} - \sqrt{\mu L}\right) = \frac{C\sqrt{\mu L}}{2} + C\left(\frac{\mu}{2} - \sqrt{\mu L}\right) = \frac{C}{2}\left(\mu - \sqrt{\mu L}\right) \leq 0 \tag{3.3.23}$$

we finally see that

$$\Delta J_n = J_{n+1} - J_n \leq 0 \tag{3.3.24}$$

To complete the proof, we note that this implies that $J_n \leq J_0$. So we have (as $x_0 = y_0$)

$$\left(1 - \sqrt{\frac{\mu}{L}}\right)^{-n}(f(x_n) - f(x^*)) \leq J_n \leq J_0 = (f(x_0) - f(x^*)) + \frac{\mu}{2}\|x_0 - x^*\|_2^2 \tag{3.3.25}$$

Using the fact that strong convexity implies $\frac{\mu}{2}\|x_0 - x^*\|_2^2 \leq f(x_0) - f(x^*)$ we finally get

$$f(x_n) - f(x^*) \leq 2\left(1 - \sqrt{\frac{\mu}{L}}\right)^n (f(x_0) - f(x^*)) \tag{3.3.26}$$

as desired. □

One disadvantage of the method analyzed in Theorem (3.3.1) is that setting the proper step size and momentum parameter requires knowing the smoothness parameter L and the strong convexity parameter α.

The optimal method for L-smooth functions is more flexible. In particular, no knowledge about the smoothness parameter is needed. One can use a line search to determine the correct step size and still obtain the optimal objective error of $O(n^{-2})$ (compared with $O(n^{-1})$ for gradient descent). In particular, we have the following result (which generalizes the results in [33] to obtain a larger family of accelerated schemes).

Theorem 3.3.2. *Assume that f is convex and differentiable with minimizer x^*.*

Let q_n be any sequence of non-negative real numbers satisfying $q_0 = 0$ and $(q_{n+1} + 1)^2 \leq (q_n + 2)^2 + 1$ (in particular $q_{n+1} \leq q_n + 1$ works).

Then, if in iteration (3.3.3), γ_n is chosen so that $\gamma_n \leq \gamma_{n-1}$ and $f(x_{n+1}) \leq f(y_n) - (\gamma_n/2)\|\nabla f(y_n)\|_2^2$, and $\alpha_n = \frac{q_n}{2+q_{n+1}}$, we have

$$f(x_n) - f(x^*) \leq 2(\gamma_n q_n(q_n + 2))^{-1}\|x_0 - x^*\|_2^2 \tag{3.3.27}$$

Note that in the above theorem we made no assumption that f was L-smooth. This emphasizes that our scheme is independent of the particular value of L. We choose the step size γ_n to provide a sufficient decrease in the objective. Such a γ_n can be found using a line search and will be about $1/L$ in the worst case (within a constant depending on the precise line search scheme).

Also, setting $q_n = \alpha n$ and $\gamma_n = 1/L$ for $\alpha \leq 1$ recovers the result from [33] (with $r = 1 + 2/\alpha$). In particular, the special case $\alpha = 1$ gives $f(x_n) - f(x^*) \leq 2Ln^{-2}\|x_0 - x^*\|_2^2$.

Proof. Consider the Lyapunov function

$$J_n = \gamma_n q_n(q_n + 2)(f(x_n) - f(x^*)) + \frac{1}{2}\|2(y_n - x^*) + q_n(y_n - x_n)\|_2^2 \tag{3.3.28}$$

We will show that $J_{n+1} \leq J_n$ which proves the theorem since $\gamma_n q_n(q_n+2)(f(x_n)-f(x^*)) \leq J_n$ and $J_0 = 2\|y_0 - x^*\|_2^2 = 2\|x_0 - x^*\|_2^2$. To this end, we denote

$$J_n^1 = \gamma_n q_n(q_n + 2)(f(x_n) - f(x^*)) \tag{3.3.29}$$

and

$$J_n^2 = \frac{1}{2}\|2(y_n - x^*) + q_n(y_n - x_n)\|_2^2 \tag{3.3.30}$$

Then we see that

$$\begin{gathered} J_{n+1}^1 - J_n^1 = \gamma_n q_n(q_n + 2)(f(x_{n+1}) - f(x_n)) + \\ (\gamma_{n+1} q_{n+1}(q_{n+1} + 2) - \gamma_n q_n(q_n + 2))(f(x_{n+1}) - f(x^*)) \end{gathered} \tag{3.3.31}$$

Since by assumption $\gamma_{n+1} \leq \gamma_n$ and $q_{n+1}(q_{n+1} + 2) = (q_{n+1} + 1)^2 - 1 \leq (q_n + 2)^2$ we see that $\gamma_{n+1} q_{n+1}(q_{n+1} + 2) \leq \gamma_n(q_n + 2)^2$ and the bottom line in the above equation is bounded by

$$(\gamma_n(q_n + 2)^2 - \gamma_n q_n(q_n + 2))(f(x_{n+1}) - f(x^*)) = 2\gamma_n(q_n + 2)(f(x_{n+1}) - f(x^*)) \tag{3.3.32}$$

Thus we see that

$$J_{n+1}^1 - J_n^1 \leq \gamma_n(q_n + 2)[2(f(x_{n+1}) - f(x^*)) + q_n(f(x_{n+1}) - f(x_n))] \tag{3.3.33}$$

The step sizes γ_n are chosen so that $f(x_{n+1}) - f(y_n) \leq -(1/2)\gamma_n\|\nabla f(y_n)\|_2^2$ and so we can rewrite the above to obtain

$$\begin{aligned} J_{n+1}^1 - J_n^1 \leq & \gamma_n(q_n + 2)[2(f(y_n) - f(x^*)) + q_n(f(y_n) - f(x_n))] \\ & - \frac{(\gamma_n(q_n + 2))^2}{2}\|\nabla f(y_n)\|_2^2 \end{aligned} \tag{3.3.34}$$

The convexity of f implies that $f(y_n) - f(x^*) \leq \nabla f(y_n) \cdot (y_n - x^*)$ and $f(y_n) - f(x_n) \leq \nabla f(y_n) \cdot (y_n - x_n)$ so we get

$$\begin{aligned} J_{n+1}^1 - J_n^1 \leq & \gamma_n(q_n + 2)\nabla f(y_n) \cdot [2(y_n - x^*) + q_n(y_n - x_n)] \\ & - \frac{(\gamma_n(q_n + 2))^2}{2}\|\nabla f(y_n)\|_2^2 \end{aligned} \tag{3.3.35}$$

Now we consider $J_{n+1}^2 - J_n^2$. Note that $J_n^2 = (1/2)\|t_n\|_2^2$ with

$$t_n = 2(y_n - x^*) + q_n(y_n - x_n)$$

Thus

$$J_{n+1}^2 - J_n^2 = (t_{n+1} - t_n) \cdot t_n + \frac{1}{2}\|(t_{n+1} - t_n)\|_2^2 \tag{3.3.36}$$

Considering that

$$J_{n+1}^1 - J_n^1 \leq \gamma_n(q_n + 2)\nabla f(y_n) \cdot t_n - \frac{(\gamma_n(q_n + 2))^2}{2}\|\nabla f(y_n)\|_2^2 \tag{3.3.37}$$

we will be done if we can show that $t_{n+1} - t_n = -\gamma_n(q_n + 2)\nabla f(y_n)$. To this end we compute

$$t_{n+1} - t_n = 2(y_{n+1} - y_n) + q_n(y_{n+1} - y_n) - q_n(x_{n+1} - x_n) + (q_{n+1} - q_n)(y_{n+1} - x_{n+1}) \tag{3.3.38}$$

using our update formulas we see that

$$y_{n+1} - x_{n+1} = \alpha_n(x_{n+1} - x_n)$$

and

$$y_{n+1} - y_n = -\gamma_n\nabla f(y_n) + \alpha_n(x_{n+1} - x_n)$$

so that this simplifies to

$$t_{n+1} - t_n = -(q_n + 2)\gamma_n\nabla f(y_n) + (2\alpha_n + q_{n+1}\alpha_n - q_n)(x_{n+1} - x_n) \tag{3.3.39}$$

which is equal to $-(q_n + 2)\gamma_n\nabla f(y_n)$ by our choice of α_n. □

This concludes our discussion of accelerated first order methods in Euclidean space. Our goal in the remainder of the chapter will be to extend these methods to the setting of the Stiefel manifold. There are three fundamental problems we need to solve in the process of doing this.

First, the (local, i.e. near the minimizer) strong convexity parameter μ and smoothness parameter L are not known. This problem occurs when applying accelerated methods to convex functions in $\mathbb{R}^n$ as well. The previous theorem shows that estimating the smoothness parameter L is not an issue as we can use a line search to find a point satisfying a sufficient decrease condition. Getting around knowledge of the strong convexity parameter is a much more difficult problem.

Second, the functions which we will be minimizing are non-convex. This is due to the fact that all globally convex functions on the Stiefel manifold are constant (since the manifold is compact). Because of this, we cannot hope to obtain a global convergence rate. However, we want a method which is guaranteed to converge and which will achieve an accelerated rate once it is close enough to the (local) minimizer.

Finally, we must find an efficient way of generalizing the momentum step

$$y_{n+1} = x_{n+1} + \alpha_n(x_{n+1} - x_n)$$

of (3.3.3) to the Stiefel manifold. We will develop a very efficient method for averaging and extrapolating on the Stiefel manifold, which can be used to design a variety of other optimization methods as well.

3.4 Adaptive Restart

We first address the lack of knowledge of the smoothness and strong convexity parameters. This issue arises even when considering convex optimization in $\mathbb{R}^n$. Recall that setting the proper the momentum and step size parameters for smooth strongly convex functions in iteration (3.3.3) requires knowing μ and L. In general, μ and L are not known and many researchers have considered the problem of estimating them adaptively (see [25], [20] and [27], for instance).

In developing our method, we build upon the work presented in [27]. The methods introduced there are based on the following observation.

Suppose we are given a μ-strongly convex, L-smooth function f. Then since f is convex and L-smooth, we can run iteration (3.3.3) with the parameters given in Theorem (3.3.2) (setting $q_n = n$) and obtain the following objective error

$$f(x_n) - f(x^*) \leq 2Ln^{-2}\|x_0 - x^*\|_2^2 \tag{3.4.1}$$

The strong convexity of f now allows us to bound the iterate error by the objective error, since strong convexity implies that $(\mu/2)\|x_n - x^*\|_2^2 \leq f(x_n) - f(x^*)$. Combining this with equation (3.4.1) we see that

$$\|x_n - x^*\|_2^2 \leq 4(L/\mu)n^{-2}\|x_0 - x^*\|_2^2 = 4\kappa n^{-2}\|x_0 - x^*\|_2^2 \tag{3.4.2}$$

where $\kappa = (L/\mu)$ is the condition number of f. This implies that after $n = \sqrt{8\kappa}$ iterations, we will have

$$\|x_n - x^*\|_2^2 \leq \frac{\|x_0 - x^*\|_2^2}{2} \tag{3.4.3}$$

So by restarting the method (i.e. setting $x_0 = x_n$ and resetting the momentum parameter) every $\sqrt{8\kappa}$ iterations, we halve the iterate error every time we restart. This means that it takes $O(\sqrt{\kappa}\log(\epsilon))$ iterations to attain an ϵ-accurate solution and thus restarting the method at this frequency recovers the asymptotically optimal convergence rate (for μ-strongly convex L-smooth functions).

Of course, in order to apply this scheme, we must know the condition number κ in order to determine the correct restart frequency. To get around this, the method proposed in [27] adaptively chooses when to restart based on an observable condition on the iterates. Specifically, they consider two restart conditions

- Function Restart Scheme: Restart when $f(x_k) > f(x_{k-1})$
- Gradient Restart Scheme: Restart when $\nabla f(y_{k-1}) \cdot (x_k - x_{k-1}) > 0$

Both of these restart conditions are based upon the analysis of a quadratic objective and it is an open problem to fully analyze their behavior when applied to an arbitrary strongly convex, smooth function. However, experimental results in [27] show empirically that the adaptively restarted methods perform well in practice.

We will show how to generalize these adaptively restarted methods to solve optimization problems on the Stiefel manifold. In the next subsection, we will modify the function restart scheme to additionally address the problem of non-convexity of functions on the Stiefel manifold. Later on, we will also show how to adapt the gradient restart scheme to the manifold setting.

3.4.1 Restart for Non-convex Functions

When adapting accelerated gradient methods to the Stiefel manifold, we are faced with the issue that the manifold is compact and so the only convex functions are constant. Consequently, the functions which we are optimizing are necessarily non-convex. In this case the convergence results of Theorems (3.3.1) and (3.3.2) don't apply and in fact we cannot hope for a 'global' convergence rate.

Instead, what we note is that in a small neighborhood of a local optimum X^* the function will be strongly convex and smooth, provided that the Hessian at X^* is positive definite. Moreover, the ratio of the strong convexity and smoothness parameters in this neighborhood will be the close to the condition number of $\nabla^2 f(X^*)$, which we denote by $\kappa(X^*)$.

Thus the accelerated gradient method analyzed in Theorem (3.3.1) suggests that we should be able to find a method which achieves a convergence rate of $O((1-\kappa(X^*)^{-1/2})^n)$ once it is close enough to the local minimum X^*. But since we have to deal with functions which are not globally convex, we hope to design a method which is guaranteed to converge to a local minimum even for non-convex functions, but which achieves the optimal convergence rate once it is close enough to the local minimum.

Our approach is to modify the function restart scheme considered in [27] and described in the previous section. We introduce the following restart condition, which forces a sufficient decrease in the objective.

- Modified Function Restart Scheme: Restart when

$$f(x_{n+1}) > f(x_n) - c_R\gamma_n\|\nabla f(y_n)\|_2^2 \tag{3.4.4}$$

  where c_R is a parameter we take to be a small constant (recall that γ_n is the step size at step n).

We now prove that with this restart condition, the algorithm converges in an appropriate sense.

Theorem 3.4.1. *Let f be a differentiable, L-smooth function, i.e. ∇f is Lipschitz with constant L. Assume also that f is bounded below.*

Consider the iteration (3.3.3) with step size γ_n chosen to satisfy $c/L \leq \gamma_n \leq \gamma_{n-1}$ for some $c \leq 1$ and $f(x_{n+1}) \leq f(y_n) - (\gamma_n/2)\|\nabla f(y_n)\|_2^2$.

If this iteration is restarted whenever (3.4.4) holds (with the new γ_0 chosen to be $\leq \gamma_n$), then we have

$$\lim_{n\to\infty} \|\nabla f(x_n)\|_2 \to 0 \tag{3.4.5}$$

Proof. Note first that our condition on the step size γ_n can always be satisfied, since by the L-smoothness of f we have that $\gamma_n = c/L$ will always work.

Also note that since $x_0 = y_0$, the condition on the step size always guarantees that $f(x_1) \leq f(x_0) - c_R\gamma_0\|\nabla F(y_0)\|_2^2$.

So we can always run the algorithm (3.3.3) in a way which satisfies the conditions of the theorem.

To complete the proof, we note that the restart condition combined with the observation that we always take at least one step implies that

$$f(x_{n+1}) \leq f(x_n) - c_R\gamma_n\|\nabla f(y_n)\|_2^2 \tag{3.4.6}$$

Summing this, we obtain

$$c_R\sum_{n=0}^{n} \gamma_n\|\nabla f(y_n)\|_2^2 \leq f(x_0) - f(x_n) \tag{3.4.7}$$

Since f is bounded below, say by M and $\gamma_n \geq c/L$ we see that

$$\sum_{n=0}^{\infty} \|\nabla f(y_n)\|_2^2 \leq \frac{L(f(x_0) - M)}{c_R c} < \infty \tag{3.4.8}$$

This implies that $\|\nabla f(y_n)\| \to 0$. Now we simply note that since f is L-smooth and $x_n = y_{n-1} - \gamma_n\nabla f(y_{n-1})$, we have that

$$\|\nabla f(x_n)\|_2 \leq (1 + L\gamma_n)\|\nabla f(y_{n-1})\|_2 \leq (1 + L\gamma_0)\|\nabla f(y_{n-1})\|_2 \tag{3.4.9}$$

where the last inequality is because $\gamma_n \leq \gamma_0$ by assumption. Thus, $\|\nabla f(x_n)\| \to 0$ as desired. □

3.5 Extrapolation and Interpolation on the Stiefel Manifold

In the previous sections, we have seen how to get around knowing the strong convexity and smoothness parameters and how to deal with non-convex functions in the process. In this section, we address the third difficulty mentioned at the end of section (3.3). Namely, we consider the problem of generalizing the momentum step of (3.3.3)

$$Y_{n+1} = X_{n+1} + \alpha_n(X_{n+1} - X_n) \tag{3.5.1}$$

to the manifold setting.

More generally, we will consider the problem of efficiently extrapolating and interpolating on the Stiefel manifold, i.e. given two points $X, Y \in S_{n,k}$ and $\alpha \in \mathbb{R}$, we want to calculate points $(1-\alpha)X + \alpha Y$ on a curve through X and Y. By setting $\alpha \in (0, 1)$ this gives a way of averaging points on the manifold and by setting $\alpha < 1$ we can extrapolate as in (3.5.1).

One very simple approach would be to perform the extrapolation or interpolation in Euclidean space and then project back onto the Stiefel manifold. However, this projection step is quite expensive. One could also replace the projection by a reorthogonalization procedure such as Gram-Schmidt (or a QR factorization). However, this is quite inaccurate if k (the number of vectors) is large and is also relatively expensive and difficult to parallelize.

The approach we take is both simpler and easier to parallelize. What we propose for generalizing

$$(1-\alpha)X + \alpha Y \tag{3.5.2}$$

is to solve for a $V \in (T_X S_{n,k})^*$ which satisfies (here R is a retraction which we have fixed in the course of designing our method)

$$Y = R(X, \phi_g(V)) \tag{3.5.3}$$

and to then extrapolate or average by setting

$$(1-\alpha)X + \alpha Y = R(X, \phi_g((1+\alpha)V)) \tag{3.5.4}$$

Note that the use of ϕ_g simply allows us to work in the dual tangent space.

The obvious difficulty with this is solving equation (3.5.3) for V, i.e. finding a V such that $R(X, \phi_g(V)) = Y$ for some given X and Y. However, if we take our retraction to be R_1 from the previous section (this is the Cayley retraction introduced in [36]), then this boils down to solving

$$\left(I+\frac{1}{2}(VX^T-XV^T)\right)X=\left(I-\frac{1}{2}(VX^T-XV^T)\right)Y \tag{3.5.5}$$

for V. Since $X^TX=Y^TY=I$, one can now easily check that $V=2Y(I+X^TY)^{-1}$ solves this equation (of course V is not unique, one can add XS to it where S is an arbitrary symmetric matrix). Thus, for this particular choice of retraction, this problem is computationally very easy to solve (it only requires solving a $k\times k$ linear system).

This gives us a computationally efficient procedure for averaging and extrapolating on the Stiefel manifold. We have already essentially described how this can be used to generalize accelerated gradient methods to the Stiefel manifold and in the next sections we will describe these methods in full detail. We also propose that this averaging and extrapolation procedure could potentially be a building block in other novel optimization algorithms on the manifold. As an example, we show how this idea can be used to generalize the gradient restart scheme in section (3.4) to the manifold.

3.5.1 Gradient Restart Scheme

We propose the following method for generalizing the gradient restart scheme to the Stiefel manifold. Recall that the gradient restart scheme restarts iteration (3.3.3) whenever

$$\nabla f(y_{k-1})\cdot(x_k-x_{k-1})>0$$

We begin by noting that $x_k=y_{k-1}-\gamma_{k-1}\nabla f(y_k)$ and so we can rewrite this condition as

$$-\gamma_{k-1}\|\nabla f(y_{k-1})\|_2^2+\nabla f(y_{k-1})\cdot(y_{k-1}-x_{k-1})>0 \tag{3.5.6}$$

Now it is clear that on the manifold $\|\nabla f(y_{k-1})\|_2^2$ should become $\|\nabla f(y_{k-1})\|_{g*}^2$. The tricky part is generalizing $\nabla f(y_{k-1})\cdot(y_{k-1}-x_{k-1})$. What we propose is to solve for a $V\in(S_{y_{k-1}})^*$ such that

$$x_{k-1}=R(y_{k-1},\phi_g(V)) \tag{3.5.7}$$

This element V then serves as $x_{k-1}-y_{k-1}$ and the analogue of the gradient restart condition becomes

$$-\gamma_{k-1}\|\nabla f(y_{k-1})\|_{g*}^2-\langle\nabla f(y_{k-1}),V\rangle_{g*}>0 \tag{3.5.8}$$

As in the previous section, we see that equation (3.5.7) can be efficiently solved for V if the retraction we are using is R_1 (the Cayley retraction introduced in [36]).

3.6 Numerical Results

We will analyze the numerical properties of the algorithms in tables (3.1) and (3.2) below, which were motivated and described in detail in the previous sections.

3.6.1 Single Eigenvector Calculations

We begin by testing our algorithms on the sphere (which is a special case of the Stiefel manifold $S_{n,k}$ with $k = 1$). The problem we solve is the eigenvector calculation

$$\arg\min_{X \in S^n} \frac{1}{2} X^T A X \tag{3.6.1}$$

where A is a symmetric matrix. The solution to this problem is the eigenvector corresponding to the smallest eigenvalue of A.

In order to evaluate the performance of our algorithm, we must investigate how the number of iterations scales with the condition number of (3.6.1) (not to be confused with the condition number of A). We now show how to calculate this condition number in terms of the eigenvalues of A.

Let $\lambda_1, ..., \lambda_n$ be the eigenvalues of A and let $v_1, ..., v_n$ be the associated eigenvectors. We know that v_1 is the minimizer and we are interested in calculating the condition number of $f(v) = \frac{1}{2} v^T A v$ at this minimum. Given a vector $v \in T_{v_1} S^n$ (which is just the space of vectors orthogonal to v_1) the unit speed geodesic in the direction v is

$$c_v(t) = \cos(t) v_1 + \sin(t) v \tag{3.6.2}$$

This allows us to calculate

$$\left. \frac{d^2}{dt^2} f(c_v(t)) \right|_{t=0} = v^T A v - v_1^T A v_1 \tag{3.6.3}$$

```
Algorithm 1: Accelerated Gradient Descent with Function Restart Scheme
Data: $f$ a smooth function, $\epsilon$ a tolerance, $c_R$ a small restart parameter
Result: A point $X_n$ such that $\|\nabla f(X_n)\|_{g^*} < \epsilon$
$X_0 \leftarrow$ initial point;
$Y_0 \leftarrow X_0$;
$n \leftarrow 0$;
$k \leftarrow 0$;
while $\|\nabla f(X_n)\|_{g^*} \geq \epsilon$ do
    $X_{n+1} \leftarrow R_1(Y_n, \phi_g(-\gamma_n \nabla f(Y_n)))$ evaluated using equation (3.2.30) with $\gamma_n$ chosen
     so that $f(X_{n+1}) \leq f(Y_n) - \frac{1}{2}\gamma_n \|\nabla f(Y_n)\|^2_{g^*}$ (Armijo condition) and $\gamma_n \leq \gamma_{n-1}$;
    if $f(X_{n+1}) > f(X_n) - c_R \gamma_n \|\nabla f(Y_n)\|^2_{g^*}$ (Restart Condition) then
        $X_{n+1} \leftarrow X_n$;
        $Y_n \leftarrow X_{n+1}$;
        $k \leftarrow 0$;
    else
        $V_n \leftarrow 2X_{n+1}(I + X^T_{n+1} X_n)^{-1}$;
        $Y_{n+1} \leftarrow R_1(X_n, (1 + \frac{k}{k+3})\phi_g(V_n))$ (apply momentum);
        $k \leftarrow k + 1$;
    end
    $n \leftarrow n + 1$;
end
```

Table 3.1: Accelerated Gradient Descent with Function Restart Scheme

Algorithm 2: Accelerated Gradient Descent with Gradient Restart Scheme

```
Data: $f$ a smooth function, $\epsilon$ a tolerance
Result: A point $X_n$ such that $\|\nabla f(X_n)\|_{g^*} < \epsilon$
$X_0 \leftarrow$ initial point;
$Y_0 \leftarrow X_0$;
$n \leftarrow 0$;
$k \leftarrow 0$;
while $\|\nabla f(X_n)\|_{g^*} \geq \epsilon$ do
    $X_{n+1} \leftarrow R_1(Y_n, \phi_g(-\gamma_n \nabla f(Y_n)))$ evaluated using equation (3.2.30) with $\gamma_n$ chosen
     so that $f(X_{n+1}) \leq f(Y_n) - \frac{1}{2}\gamma_n \|\nabla f(Y_n)\|^2_{g^*}$ (Armijo condition) and $\gamma_n \leq \gamma_{n-1}$;
    $W_n \leftarrow 2X_n(I + X_n^T Y_n)^{-1}$;
    if $\langle \nabla f(Y_n), W_n \rangle_{g*} < -\gamma_n \|\nabla f(Y_n)\|^2_{g^*}$ (Restart Condition) then
        $X_{n+1} \leftarrow X_n$;
        $Y_n \leftarrow X_{n+1}$;
        $k \leftarrow 0$;
    else
        $V_n \leftarrow 2X_{n+1}(I + X_{n+1}^T X_n)^{-1}$;
        $Y_{n+1} \leftarrow R_1(X_n, (1 + \frac{k}{k+3})\phi_g(V_n))$ (apply momentum);
        $k \leftarrow k + 1$;
    end
    $n \leftarrow n + 1$;
end
```

Table 3.2: Accelerated Gradient Descent with Gradient Restart Scheme

Applying formula (3.1.14) and noting that $v \in T_{v_1}S^n$ we see that

$$\kappa(Hf(v_1)) = \frac{\sup_{\substack{\|v\|_2=1 \\ v\cdot v_1=0}} v^T A v - v_1^T A v_1}{\inf_{\substack{\|v\|_2=1 \\ v\cdot v_1=0}} v^T A v - v_1^T A v_1} = \frac{\lambda_n - \lambda_1}{\lambda_2 - \lambda_1} \tag{3.6.4}$$

In figure (3.1) we present the results of applying our algorithms to the above problem with $A_n = \mathrm{diag}(0, 1, ..., n)$ on S^{n+1}. By the above calculation, the condition number of this problem is n. We initialize X_0 at a uniformly random point on the sphere and plot the number of iterations (with the tolerance $\epsilon = 1\mathrm{e}{-3}$) vs the condition number n, for $n = 100$ through $n = 2000$.

From these results, we see that our method appears to achieve the desired convergence behavior. In particular, we plot a log linear fit to the data and note that the coefficients are slightly larger than .5 in both cases, which suggests that the iteration count scales approximately as the square root of the condition number. We note that the 'staircase' behavior is due to the step size selection rule. During the course of the algorithm, the line search will decrease the step size by a constant factor until it is at most $1/L$ (L being the smoothness of the objective). In our example, the smoothness L is proportional to n and the steps occur whenever L becomes large enough to force the step size to decrease further than before.

We test this empirical observation over a larger range of condition numbers by solving the same problem with $n = 100, (1.5) \cdot 100, (1.5^2) \cdot 100, ..., (1.5^{20}) \cdot 100$. To reduce the random fluctuations, we solve each problem 10 times (with different random starting points) and plot the average number of iterations against the condition number in figure (3.2).

3.6.2 Multiple Eigenvector Calculations

We now test our algorithms on the Stiefel manifold $S_{n,k}$ with $k > 1$. The problem we consider is that of calculating the smallest k eigenvectors of a symmetric linear operator A. We begin by reformulating this as the optimization problem

$$\arg\min_{X \in S_{n,k}} \frac{1}{2} \sum_{i=0}^{k} \alpha_k \langle X_k, AX_k \rangle \tag{3.6.5}$$

where X_k denotes the k-th column of X and $0 < \alpha_1 < \alpha_2 < ... < \alpha_k$ are coefficients which force the minimizer to consist of eigenvectors of A rather than eigenvectors up to an orthogonal transformation.

As before, we want to investigate how the number of iterations depends upon the condition number of (3.6.5) and so we begin by calculating this condition number. For this calculation, we will use the following equivalent form of equation (3.1.14)

$$\kappa(Hf(x)) = \frac{\sup_{\|w\|_{g*}=1} Hf(x)(\phi_g(w))}{\inf_{\|w\|_{g*}=1} Hf(x)(\phi_g(w))} \tag{3.6.6}$$

because geodesics following dual directions have a simpler formula.

Using equation (3.2.8) we see that the geodesic starting a point X with initial derivative $\phi_g(W)$ is given by

$$c_{W,X}(t) = \exp(t(WX^T - XW^T))X \tag{3.6.7}$$

Rewriting out objective function as

$$f(X) = \frac{1}{2}\mathrm{Tr}(X^T AXD_\alpha) \tag{3.6.8}$$

where D_α is a diagonal matrix with entries $\alpha_1, ..., \alpha_k$, we calculate

$$\frac{d}{dt}f(c_{W,X}(t)) = \mathrm{Tr}(D_\alpha c_{W,X}(t)^T A(WX^T - XW^T)c_{W,X}(t)) \tag{3.6.9}$$

whence the second derivative is

$$\begin{aligned}\left.\frac{d^2}{dt^2}f(c_{W,X}(t))\right|_{t=0} =&\mathrm{Tr}(D_\alpha X^T A(WX^T - XW^T)^2 X) - \\ &\mathrm{Tr}(D_\alpha X^T(WX^T - XW^T)A(WX^T - XW^T)X)\end{aligned} \tag{3.6.10}$$

Now let $v_1, ..., v_n$ and $\lambda_1, ..., \lambda_n$ be the eigenvectors and eigenvalues of A and X^* be the minimizer of f, i.e. $X^*_k = v_1, ..., X^*_1 = v_k$. This means that $AX^* = X^* D_\lambda$, where D_λ is the diagonal matrix with diagonal entries $\lambda_k, ..., \lambda_1$. This, along with the fact that $X^{*T}X^* = I$ and $W^T X^* + X^{*T} W = 0$ (since $W \in T_{X^*}S_{n,k}$) allows us the rewrite the second derivative as

$$\begin{aligned}\left.\frac{d^2}{dt^2}f(c_{W,X^*}(t))\right|_{t=0} =&\mathrm{Tr}(D_\alpha W^T AW) - \mathrm{Tr}(D_\alpha D_\lambda W^T W) + \\ &3\mathrm{Tr}(D_\alpha D_\lambda X^{*T} WX^{*T} W) - 3\mathrm{Tr}(D_\alpha X^{*T} WD_\lambda X^{*T} W)\end{aligned} \tag{3.6.11}$$

We now decompose W as $W = X^*N + W^p$ where N is antisymmetric and $X^{*T}W^p = 0$ (N is just $X^{*T}W$). Using formula (3.2.5), we see that $\|W\|_{g*}^2 = 2\|N\|_F^2 + \|W^p\|_F^2$. Plugging this decomposition into (3.6.11) we obtain

$$\begin{aligned}\frac{d^2}{dt^2} f(c_{W,X^*}(t))\bigg|_{t=0} =&\mathrm{Tr}(D_\alpha W^{pT} A W^p) - \mathrm{Tr}(D_\alpha D_\lambda W^{pT} W^p) + \\ &4\mathrm{Tr}(D_\alpha D_\lambda N^2) - 4\mathrm{Tr}(D_\alpha N D_\lambda N)\end{aligned} \tag{3.6.12}$$

We now compute the numerator in formula (3.6.6),

$$\sup_{\|W\|_{g*}=1} Hf(X^*)(\phi_g(W)) = \sup_{\|W\|_{g*}=1} \frac{d^2}{dt^2} f(c_{W,X^*}(t))\bigg|_{t=0} \tag{3.6.13}$$

As $\|W\|_{g*} = 2\|N\|_F^2 + \|W^p\|_F^2$ and the objective in (3.6.12) is the sum of a term which is quadratic in W^p and a term which is quadratic in N, we see that the above maximum will be achieved either when $N = 0$ or when $W^p = 0$. So we independently consider the problems

$$\sup_{\substack{\|W^p\|_F^2=1 \\ X^{*T}W^p=0}} \mathrm{Tr}(D_\alpha W^{pT} A W^p) - \mathrm{Tr}(D_\alpha D_\lambda W^{pT} W^p) \tag{3.6.14}$$

and

$$\sup_{\substack{\|N\|_F^2=.5 \\ N+N^T=0}} 4(\mathrm{Tr}(D_\alpha D_\lambda N^2) - \mathrm{Tr}(D_\alpha N D_\lambda N)) \tag{3.6.15}$$

and then take the maximum. We easily see that the optimizer in (3.6.14) is attained when $W_k^p = v_n$ and $W_i^p = 0$ for $i < k$ (i.e. the last column of W consists of the largest eigenvector of A and the other columns are 0). This gives a value of $\alpha_k(\lambda_n - \lambda_1)$.

To handle the problem (3.6.15), we note first that the antisymmetry of N implies that

$$4(\mathrm{Tr}(D_\alpha D_\lambda N^2) - \mathrm{Tr}(D_\alpha N D_\lambda N)) = 2\langle D_\lambda N - N D_\lambda, N D_\alpha - D_\alpha N\rangle_F \tag{3.6.16}$$

so that we obtain

$$\sup_{\substack{\|N\|_F^2=.5 \\ N+N^T=0}} 4\sum_{i<j} N_{ij}^2(\lambda_j - \lambda_i)(\alpha_j - \alpha_i) \tag{3.6.17}$$

The maximum here is clearly obtained when $N_{1k} = -N_{k1} = .5$, which produces a value of $(\lambda_k - \lambda_1)(\alpha_k - \alpha_1)$. Notice that this value is always less than $\alpha_k(\lambda_n - \lambda_1)$ and so the numerator in formula (3.6.6) is just $\alpha_k(\lambda_n - \lambda_1)$.

Switching the suprema to infema in the above analysis easy shows that the denominator in (3.6.6) is either $\alpha_1(\lambda_{k+1}-\lambda_k)$ (arising from W^p) or $\min_{i<k}(\lambda_{i+1}-\lambda_i)(\alpha_{i+1}-\alpha_i)$, whichever is smaller. So we finally obtain the formula

$$\kappa(Hf(X^*)) = \frac{\alpha_k(\lambda_n - \lambda_1)}{\min\{\alpha_1(\lambda_{k+1}-\lambda_k), \min_{i<k}(\lambda_{i+1}-\lambda_i)(\alpha_{i+1}-\alpha_i)\}} \tag{3.6.18}$$

This calculation leads to the interesting question of how to choose the weights α_i given the eigenvalues of A to minimize the condition number. Of course, for practical applications one probably will not have access to the eigenvalues of A. However, if one could obtain appropriate estimates on the eigenvalues, then this information could be used to guide the choice of weights.

In the following, we will restrict our attention to the situation where the eigenvalues of A are distinct. If this is not the case, then it is possible for the condition number to be infinite (regardless of the choice of weights). What this means is that the function that we are optimizing is not strongly convex in a neighborhood of the optimum (although it is convex). This leads to a slow down of the method to an objective error of $O(n^{-2})$.

We break down the problem of choosing optimal weights into two pieces. First, we fix α_1 and α_k and note that minimizing the condition number over all other α_j is equivalent to maximizing

$$\min_{i<k}(\lambda_{i+1}-\lambda_i)(\alpha_{i+1}-\alpha_i) \tag{3.6.19}$$

An elementary argument implies that this maximum will be achieved when $(\lambda_{i+1}-\lambda_i)(\alpha_{i+1}-\alpha_i)$ is constant for all i (i.e. when the minimum is achieved for each value of i). This implies that the optimal value is $(\alpha_k-\alpha_1)C_{\lambda,k}$ with

$$C_{\lambda,k} = \left(\sum_{i=1}^{k-1}\frac{1}{\lambda_{i+1}-\lambda_i}\right)^{-1} \tag{3.6.20}$$

Now, we minimize

$$\frac{\alpha_k(\lambda_n-\lambda_1)}{\min\{\alpha_1(\lambda_{k+1}-\lambda_k), (\alpha_k-\alpha_1)C_{\lambda,k}\}} \tag{3.6.21}$$

over α_1 and α_k. Again, an elementary argument implies that this minimum will be achieved when

$$\alpha_1(\lambda_{k+1}-\lambda_k) = (\alpha_k-\alpha_1)C_{\lambda,k} \tag{3.6.22}$$

This allows us to solve for the optimal values of α_1 and α_k (up to scaling, of course). Plugging all of this into equation (3.6.18), we see that the smallest we can make the condition number (by choosing optimal weights α_i) is

$$\kappa(Hf(X^*))_{opt} = (\lambda_n - \lambda_1)\left(\sum_{i=1}^{k} \frac{1}{\lambda_{i+1} - \lambda_i}\right) \tag{3.6.23}$$

Finally, we come to numerical results. As in the single eigenvector calculations, we let $A_n = \text{diag}(0, ..., n)$ and calculate the first k eigenvectors by optimizing over $S_{n+1,k}$. Setting the weights $\alpha_i = i$ produces the optimal condition number of kn. We set $k = 10$, initialize X_0 at a uniformly random point on $S_{n+1,10}$ and plot the number of iterations (with the tolerance $\epsilon = 1\text{e}{-3}$) vs the condition number, for $n = 100$ through $n = 2000$. The results can be found in figure (3.3).

We again see that our method empirically achieves the desired convergence behavior. Indeed, we plot a log linear fit whose coefficient is slightly larger than .5 in both cases, similar to what was observed on the sphere (we note the same 'staircase' behavior, as well, which has the same explanation). This indicates that the method also works well when optimizing over a larger set of orthonormal vectors.

As in the case of a sphere, we test this observation rigorously over a large range of condition numbers. We solve the same problem with $n = 100, (1.5)\cdot 100, (1.5^2)\cdot 100, ..., (1.5^{20})\cdot 100$. To reduce the random fluctuations, we solve each problem 10 times (with different random starting points) and plot the average number of iterations against the condition number in figure (3.4).

3.7 Conclusion

In this chapter, we developed novel accelerated first-order optimization methods designed to handle orthogonality constraints. The algorithms developed are a generalization of Nesterov's gradient descent to the Stiefel manifold. In the process, we constructed an efficient way of averaging and extrapolating points on the manifold, which we believe can be useful in developing other novel optimization algorithms. Numerical results indicate that our methods achieve the desired scaling with the condition number of the problem.

We conclude by noting that if the objective has some group of symmetries, then our algorithm behaves as if it were running on the quotient of $S_{n,k}$ by this group of symmetries. Thus we recover linear convergence (even though the objective is not strongly convex if the symmetry group is continuous) and the important quantity is the condition number of the objective as a function on this quotient manifold. We have observed this behavior experimentally and will include the relevant experiments in a future paper.

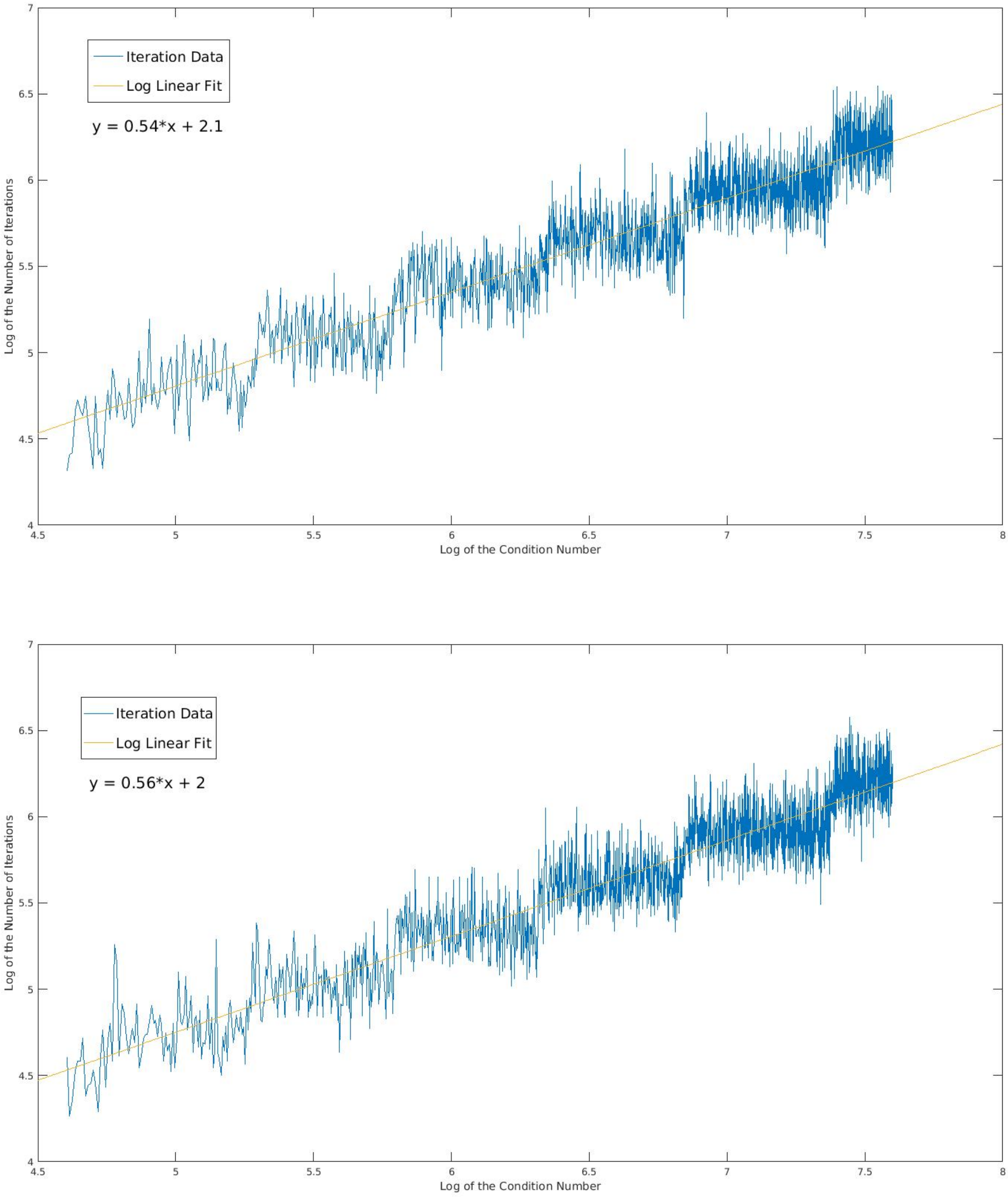
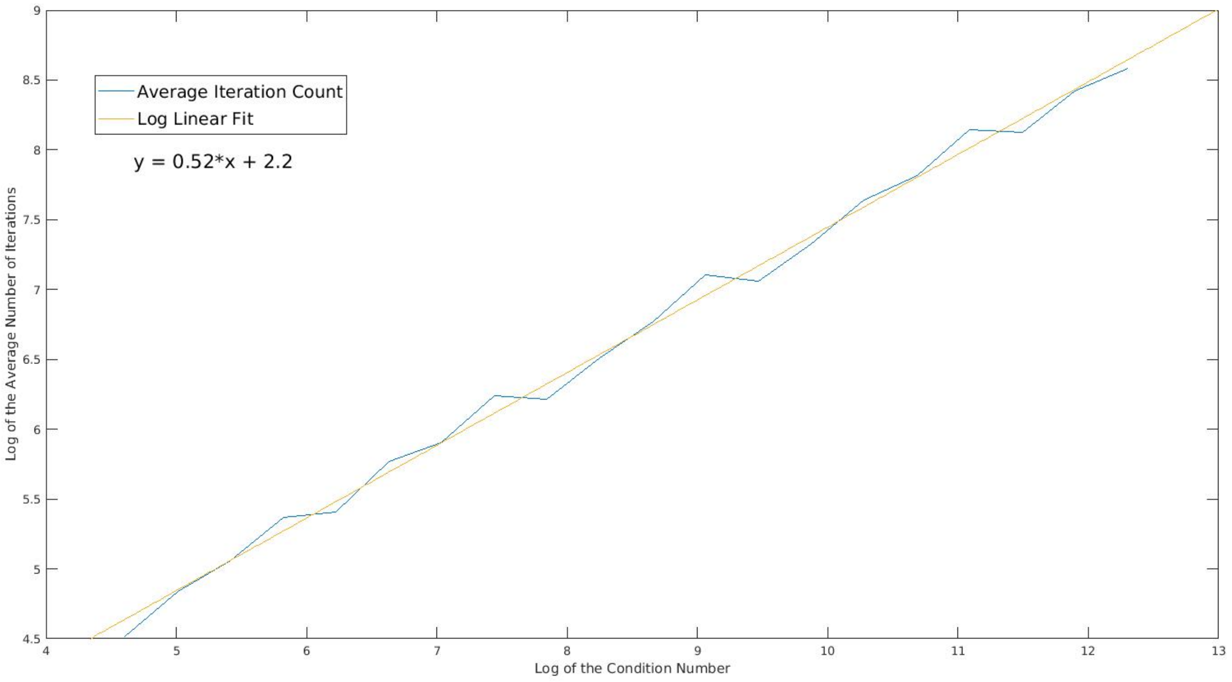


Figure 3.1: Iteration Count vs Condition Number (Sphere), Function Restart (top) Gradient Restart (bottom)



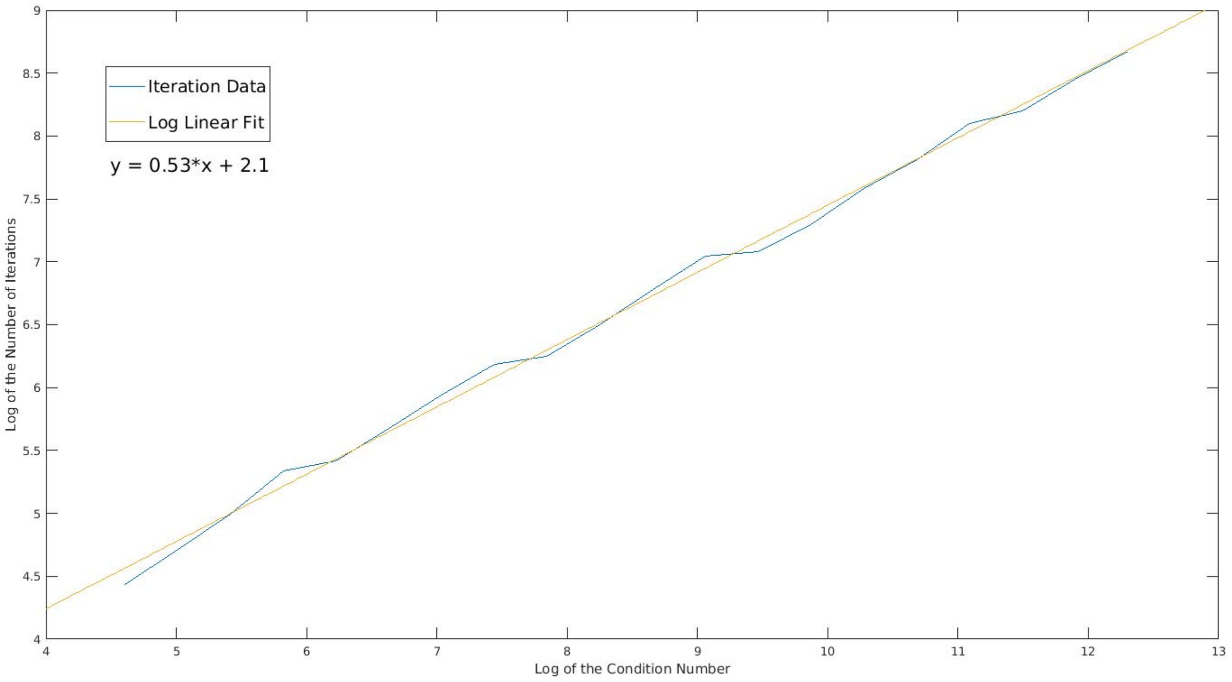
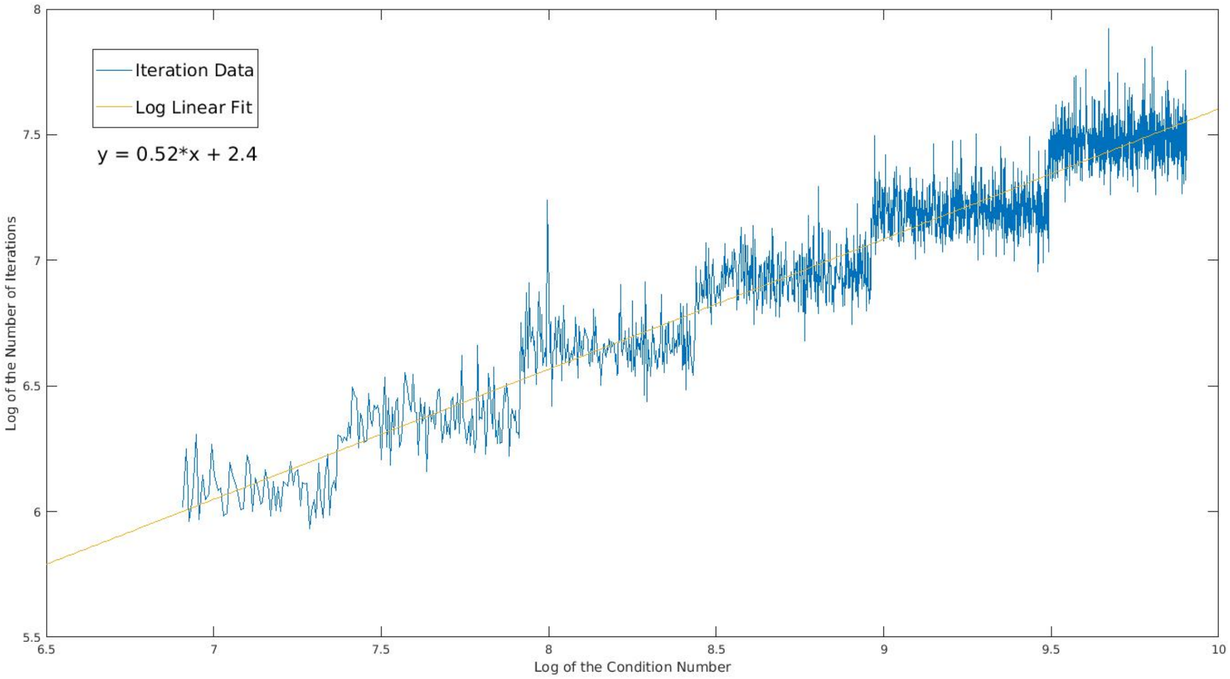


Figure 3.2: Iteration Count vs Condition Number (Sphere, larger range of condition numbers), Function Restart (top), Gradient Restart (bottom)



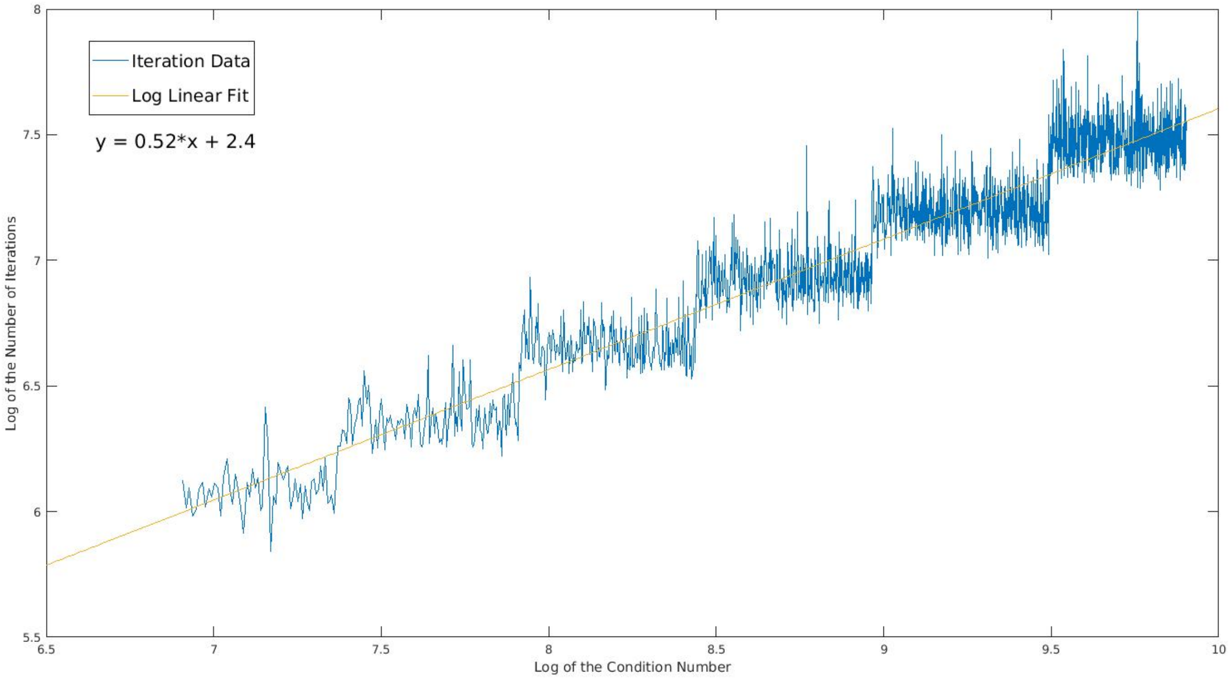


Figure 3.3: Iteration Count vs Condition Number with $k = 10$, Function Restart (top), Gradient Restart (bottom)

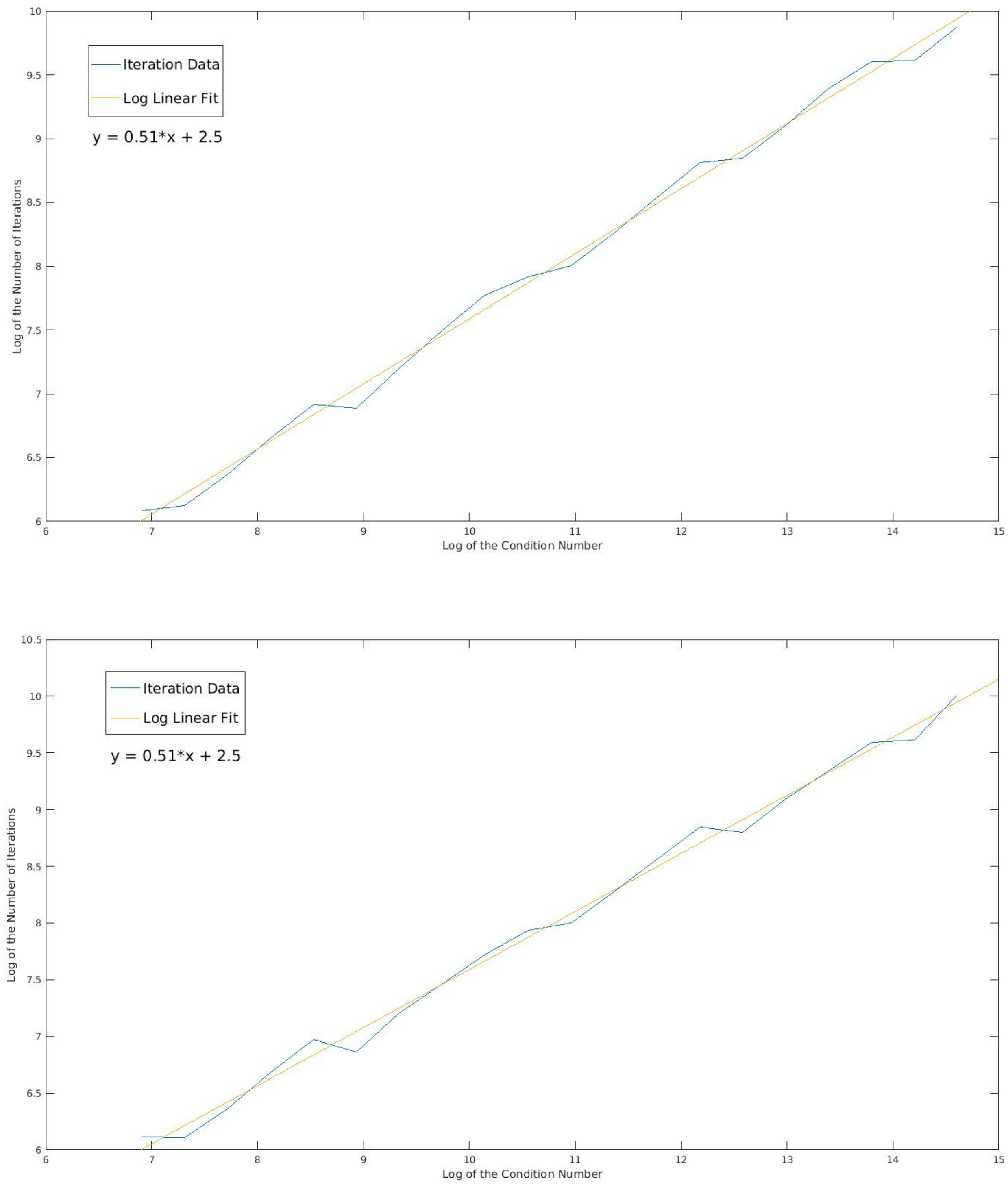


Figure 3.4: Iteration Count vs Condition Number ($k = 10$, larger range of condition numbers), Function Restart (top), Gradient Restart (bottom)

CHAPTER 4

Applications to Electronic Structure Calculations

Electronic structure calculations are concerned with first principles modelling of quantum mechanical systems, particularly electrons in a fixed external potential. The governing physical equation for such a system is the non-relativistic Schrödinger equation (see, for instance, [31])

$$i\hbar\frac{\partial}{\partial t}\Psi(x_1, ..., x_n, t) = \mathcal{H}\Psi(x_1, ..., x_n, t) \tag{4.0.1}$$

Here the x_i are variables corresponding to the position and spin of the i-th electron and $\mathcal{H}$ is the Hamiltonian of the system. In addition, the joint wave function Ψ is constrained to be anti-symmetric with respect to interchange of the variables $x_1, ..., x_n$ (since we are modelling electrons, which are fermions).

In computational chemistry and materials science, one is often interested in the electron density of the ground state or of the equilibrium configuration at a particular temperature [22], i.e. one is interested in solving the eigensystem (the time-independent Schrödinger equation)

$$\mathcal{H}\Psi(x_1, ..., x_n) = \lambda\Psi(x_1, ..., x_n) \tag{4.0.2}$$

or in calculating the diagonal of the Gibbs canonical ensemble [9]

$$\frac{\exp(-\beta\mathcal{H})}{\operatorname{Tr}\exp(-\beta\mathcal{H})} \tag{4.0.3}$$

In both of these cases, the quantity of interest is typically the electron density

$$\rho(x) = n\int|\Psi(x, x_2, ..., x_n)|^2 dx_2...dx_n \tag{4.0.4}$$

In the present work we will only be concerned with the calculation of the ground state electron density, i.e. with the approximate solution of (4.0.2). Solving the eigensystem (4.0.2) directly is infeasible, because Ψ is a function of $3n$ variables and discretizing the operator H on such a high dimensional space is computationally impossible even for small values of n. To get around this, we use the well known Kohn-Sham approximation of Density Functional Theory [22].

The starting point for this approximation is to notice that the eigenfunction problem can be written as follows

$$\arg\min_{\Psi}\langle\Psi,\mathcal{H}\Psi\rangle = \arg\min_{\Psi}\left[\langle\Psi,\mathcal{H}_e\Psi\rangle + \int_{\mathbb{R}^3} V(x)\rho(x)dx\right] \tag{4.0.5}$$

where H_e is the energy of the system not due to the interaction of the electrons and the potential (H_e consists of the kinetic energy of the electrons and the Couloumb interactions of the electrons with themselves). The important point is that the dependence of the total energy on the potential is through the electron density. This allows us to solve for the ground state electron density as follows

$$\rho^* = \arg\min_{\rho}\left[\arg\min_{\Psi\to\rho}\langle\Psi,\mathcal{H}_e\Psi\rangle\right] + \int_{\mathbb{R}^3} V(x)\rho(x)dx \tag{4.0.6}$$

Here the inner minimization is over all anti-symmetric wavefunctions which give rise to the density ρ. If one could determine the functional

$$F(\rho) = \arg\min_{\Psi\to\rho}\langle\Psi,\mathcal{H}_e\Psi\rangle \tag{4.0.7}$$

then one could replace the eigenfunction problem (4.0.2), which is in $3n$ dimensions, by the above optimization over functions on $\mathbb{R}^3$.

In practice, the functional F above is impossible to determine exactly and very difficult to approximate. The Kohn-Sham approximation makes the assumption that the ground state electron density is the same as the ground state density of some system of non-interacting electrons. We then decompose the functional F as (we assume for simplicity that the number of electrons is even)

$$F(\rho) = \arg\min_{\langle\phi_i,\phi_j\rangle=\delta_{ij}}\sum_{i=1}^{n/2}\|\nabla\phi_i\|_2^2 + E(\rho) \tag{4.0.8}$$

where the minimization includes the constraint that $\rho(x) = 2\sum_{i=1}^{n/2}|\phi_i(x)|^2$. Here the first term is the kinetic energy of a system of non-interacting electrons (which still obey the Pauli exclusion principle). In the literature, the remaining term $E(\rho)$ is also decomposed into the electrostatic interaction of the electrons

$$E_{Hartree}(\rho) = \frac{1}{2} \int \int \frac{\rho(x)\rho(y)}{|x-y|} dx dy \tag{4.0.9}$$

and all remaining terms $E_{xc}(\rho)$, which is called the exchange correlation functional. Plugging this into (4.0.6) we obtain the following optimization problem for the ground state electron density

$$\underset{\langle \phi_i, \phi_j \rangle = \delta_{ij}}{\arg\min}\ 2 \sum_{i=1}^{n/2} \|\nabla \phi_i\|_2^2 + \int_{\mathbb{R}^3} V(x)\rho(x) dx + E_{Hartree}(\rho) + E_{xc}(\rho) \tag{4.0.10}$$

where $\rho(x) = 2 \sum_{i=1}^{n/2} |\phi_i(x)|^2$.

Determining approximations to the E_{xc} functional which accurately model the behavior of molecules and solid state materials is an active area of research in computational chemistry and materials science. Another great challenge is numerically solving the approximation (4.0.10). The method typically used is the self-consistent field iteration, which is essentially a fixed point iteration in which each step involves solving an eigensystem [22]. However, it is known that the self-consistent field iteration often converges very slowly or fails to converge entirely [17, 21] and complicated heuristics must be developed in these situations.

We note that when discretizing (4.0.10) with an orthonormal basis the resulting problem is the optimization of a smooth function with orthogonality constrains. The algorithms developed in the previous chapter provide an efficient way of solving this problem. Previous approaches to directly optimizing (4.0.10) on the Stiefel manifold have run into the issue that the objective is ill-conditioned and convergence is slow [8]. Our algorithm specifically addresses this issue and we propose it as an efficient and robust alternative to the self-consistent field iteration.

Our hope is that this will allow researchers to study new exchange-correlation functionals and new potentials without having to worry about fine-tuning a self-consistent field iteration, they can simply apply our algorithm 'out of the box'. In this chapter, we demonstrate the use of our algorithms to perform a simple one dimensional DFT calculation.

4.1 1D-Jellium Calculation

Jellium is a physical model consisting of a cloud of electrons interacting with a fixed positive background density [11]. The goal is to understand the quantum mechanical effects of the interacting electrons without needing to precisely model the positions of the nuclei in a material. In this section, we present the simulation of an infinite slab of jellium. Although the calculation is primitive, it demonstrates that our method can be used successfully to solve electronic structure problems.

Since the slab of jellium is assumed to be infinite in two directions, the calculation reduces to a one-dimensional calculation. In particular, the functions ϕ_i in (4.0.10) are functions of a single variable, which we take to have zero Dirichlet boundary conditions (which corresponds to putting the whole system into an infinite potential well). Taking into account the form of the Couloumb interaction in one dimension, we obtain the following energy functional

$$\underset{\langle\phi_i,\phi_j\rangle=\delta_{ij}}{\arg\min}\ 2\sum_{i=1}^{k/2}\left\|\frac{\partial^2}{\partial x^2}\phi_i\right\|_2^2 + \int_{-L}^{L} V(x)\rho(x)dx + \frac{1}{2}\int_{-L}^{L}\int_{-L}^{L}\rho(x)\rho(y)|x-y|dxdy \qquad (4.1.1)$$

where L is the length of the system and we have ignored the exchange-correlation functional in the present simulation. Also, we won't worry about constants which arise from the choice of units as this is just a proof of concept.

We discretize the interval using $n = 1000$ points and simulate $k = 40$ electrons. A balancing amount of positive charge is uniformly distributed on the middle third of the interval. We solve the energy optimization problem using the function restart algorithm in (3.1). The results with two very different interval widths L are shown in figure (4.1), where we also indicate the number of iterations. Notice that the narrower interval is a more difficult problem because the dominance of the kinetic energy term results in a more poorly conditioned objective.

4.2 Conclusion

In this chapter, we proposed our accelerated first-order optimization methods for orthogonality constrained problems as an alternative to the self-consistent field iteration in Density Functional Theory. We provided numerical experiments demonstrating their use on a one-dimensional jellium calculation.

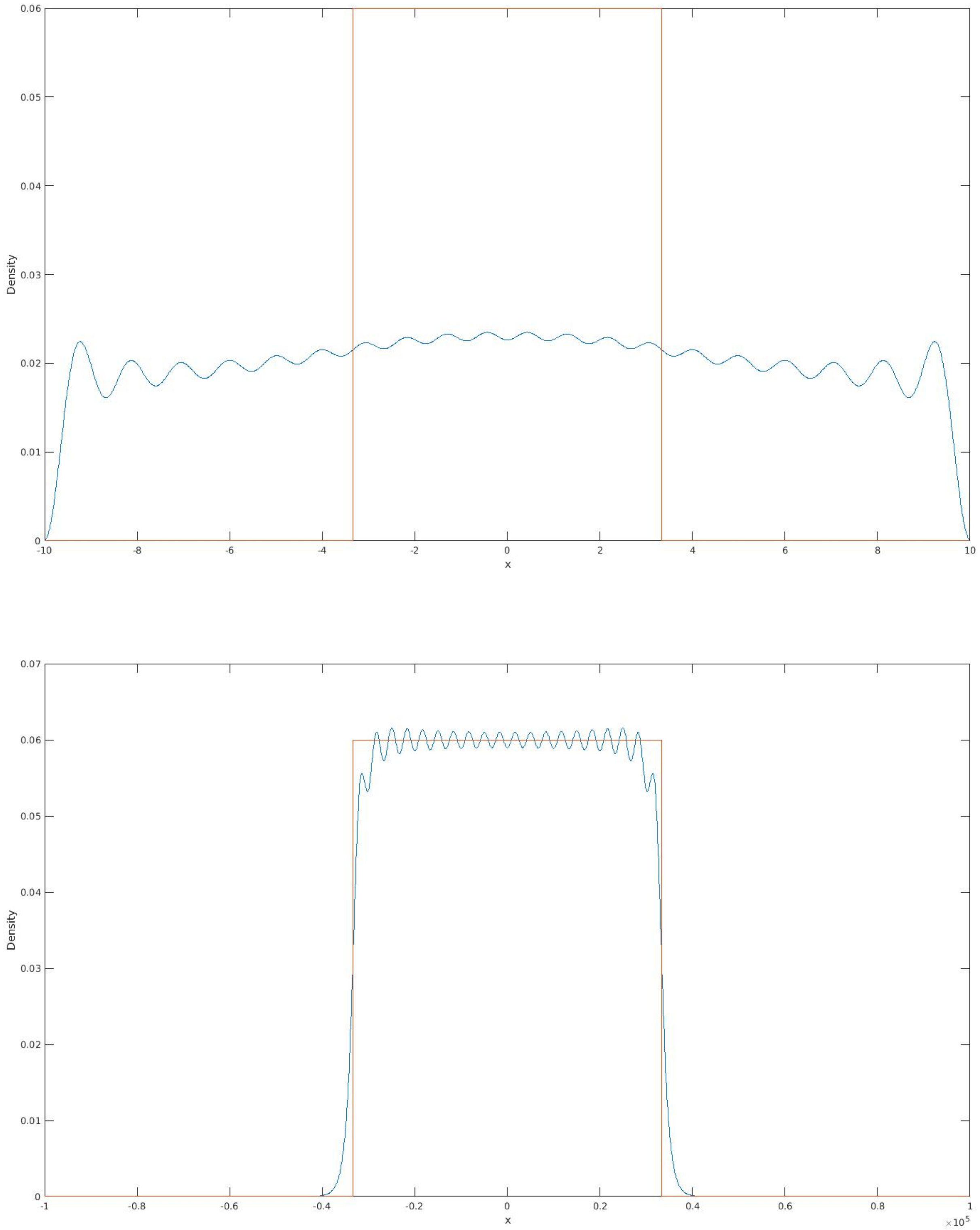


Figure 4.1: Jellium results with L = 10 (top, 3096 iterations) and L = 100000 (bottom, 966 iterations)

CHAPTER 5

Compressed Modes

In this chapter we study methods for calculating compressed modes. Compressed modes, which were first introduced in [28], are defined to be the solutions of

$$\underset{\substack{f_1,\ldots,f_n:\Omega\to\mathbb{R} \\ \langle f_i f_j\rangle=\delta_{ij}}}{\arg\min} \sum_{i=1}^{n} \langle f_i, Hf_i\rangle + \gamma\|f_i\|_1 \tag{5.0.1}$$

Here $\Omega \subset \mathbb{R}^n$ or $\Omega = \mathbb{R}^n/\Gamma$ with Γ a group of isometries of $\mathbb{R}^n$, and H is an elliptic operator on $H^1(\Omega)$. Generally, H will be the Hamiltonian operator for a quantum mechanical system; in particular, H of the form $\Delta + V$ for some potential $V : \Omega \to \mathbb{R}$ is typical.

To explain the motivation behind this problem, we first consider removing the L^1 term and analyze the problem

$$\underset{\substack{f_1,\ldots,f_n:\Omega\to\mathbb{R} \\ \langle f_i f_j\rangle=\delta_{ij}}}{\arg\min} \sum_{i=1}^{n} \langle f_i, Hf_i\rangle \tag{5.0.2}$$

If H has a discrete spectrum, which will hold if Ω is compact or if V grows rapidly enough at infinity, then the solution to this problem consists of any orthonormal basis for the space spanned by the smallest n eigenfunctions of H. In particular, the problem is highly degenerate since the orthogonal group O_n leaves the solution space invariant.

It is often the case that we desire a particular solution of (5.0.2). For instance, we may wish to know the eigenfunctions themselves as opposed to simply a basis for their span. Alternatively, in many physical applications, it is desirable to find a basis which is particularly well-localized in space. This is the idea behind the Wannier functions analyzed in [15].

For some simple problems, it is possible to analytically determine a well-localized basis. However, for more complicated systems this is infeasible. The idea behind the compressed modes (5.0.1) is that adding an L^1 penalty to this problem will result in a sparse, and thus localized, approximate basis. This can then be used to systematically find localized bases similar to the Wannier functions [22].

Many promising analytic results toward this end are known. In particular, in [4] an explicit bound on the measure of the support of each of the compressed modes f_i is derived. Also, in [3] it is shown that as $\mu \to 0$ the optimizers f_i converge to a solution of (5.0.2), i.e. to a basis for the space spanned by the smallest n eigenfunctions of H. This suggests that solving (5.0.1) with a small value of μ is a promising approach to systematically generating Wannier-type functions for arbitrary operators H.

One of the biggest difficulties with this approach is that numerically solving (5.0.1) is extremely challenging. The most popular approach to this problem is a complicated splitting method based on ADMM which was introduced in [28].

In this chapter we study numerical methods for solving (5.0.1). We begin by discussing ADMM-based splitting methods for solving (5.0.1) and then show how the ideas of the previous chapters can be used to derive generalizations of averaged subgradient descent. Unfortunately, these methods are not very robust or efficient. Finally, in the last section we show how smoothing the L^1 term and using the accelerated gradient descent methods developed in the earlier chapters provide a robust and efficient way of solving the problem.

5.1 Splitting Methods

For numerical calculations, we consider the case where Ω is a rectangle or a torus (i.e. rectangle with periodic boundary condition). As in [28], we use a regular grid to discretize (5.0.1), which results in the problem

$$\arg\min_{X\in S_{n,k}} \mathrm{Tr}(X^T H X) + \gamma\|X\|_1 \tag{5.1.1}$$

where $S_{n,k}$ is the Stiefel manifold of order (n,k).

The splitting method introduced in [28] and [18] solves (5.1.1) by rewriting the problem as

$$\underset{X=Z,\ X=Y}{\arg\min}\ \operatorname{Tr}(X^THX)+\gamma\|Y\|_1+\chi_{S_{n,k}}(Z) \tag{5.1.2}$$

Here $\chi_{S_{n,k}}$ is the characteristic function of the Stiefel manifold, i.e. it is 0 if $Z\in S_{n,k}$ and ∞ otherwise. We then proceed by introducing Lagrange multipliers B^1 and B^2 for the constraints $X=Z$ and $X=Y$ and iteratively solve

$$\begin{aligned}X_{n+1}&=\underset{X}{\arg\min}\,L_{\mu,\nu}(X,Y_n,Z_n,B_n^1,B_n^2)\\Y_{n+1}&=\underset{Y}{\arg\min}\,L_{\mu,\nu}(X_{n+1},Y,Z_n,B_n^1,B_n^2)\\Z_{n+1}&=\underset{Z}{\arg\min}\,L_{\mu,\nu}(X_{n+1},Y_{n+1},Z,B_n^1,B_n^2)\end{aligned}$$

and then updating the Lagrange multipliers $B_{n+1}^1=B_n^1+\mu(X-Z)$ and $B_2+\nu(X-Y)$. Here $L_{\mu,\nu}$ is the augmented Lagrangian

$$\begin{aligned}L_{\mu,\nu}(X,Y,Z,B^1,B^2)=&\operatorname{Tr}(X^THX)+\gamma\|Y\|_1+\chi_{S_{n,k}}(Z)+\langle B^1,(X-Z)\rangle+\\&\langle B^2,(X-Y)\rangle+\frac{\mu}{2}\|X-Z\|_F^2+\frac{\nu}{2}\|X-Y\|_F^2\end{aligned} \tag{5.1.3}$$

The motivation behind this algorithm is the split-Bregman method introduced in [12], for which there is a detailed convergence theory in the convex case, i.e. when applied to

$$\underset{Ax+By=z}{\arg\min}\,f(x)+g(y) \tag{5.1.4}$$

where f and g are convex functions. In the compressed modes calculation, we are splitting twice (since we have two constraints $X=Z$ and $X=Y$) and our characteristic function is not convex (since the Stiefel manifold is not a convex set). Consequently, the split-Bregman convergence theory doesn't apply.

A rigorous convergence theory for ADMM applied to non-convex problems is developed in [35]. Here it is shown that the very general algorithm given by iteratively solving

$$x_{n+1}^i=\underset{x}{\arg\min}\,L_\mu(x_{n+1}^1,...,x_{n+1}^{i-1},x,x_n^{i+1},...,x_n^k,\lambda_n) \tag{5.1.5}$$

for $i=1,...,k$ and updating the Lagrange multiplier $\lambda_{n+1}=\lambda_n+\mu(A_1x_{n+1}^1+...+A_kx_{n+1}^x+b)$ will converge to a stationary point for the problem

$$\underset{A_1x^1+...+A_kx^k+b=0}{\arg\min}\quad f_1(x^1)+...+f_k(x^k) \tag{5.1.6}$$

provided some minor technical assumptions on $f_1, ..., f_k$ and $A_1, ..., A_k$. Here the augmented Lagrangian is given by

$$L_\mu(x^1, ..., x^k, \lambda) = f_1(x^1) + ... + f_k(x^k) + \langle \lambda, A_1x^1 + ... + A_kx^k + b\rangle + \frac{\mu}{2}\|A_1x^1 + ... + A_kx^k + b\|_2^2 \tag{5.1.7}$$

The most restrictive assumption made in [35] is the assumption that

$$\mathrm{Im}(A_1) + ... + \mathrm{Im}(A_{k-1}) \subset \mathrm{Im}(A_k) \tag{5.1.8}$$

This means that whichever values for $x_1, ..., x_{k-1}$ show up in the iteration, there exists an assignment for x_k which satisfies the constraint. This assumption holds for a two-variable splitting scheme of the form

$$\underset{x=y}{\arg\min}\, f(x) + g(y) \tag{5.1.9}$$

but clearly doesn't hold for a three-variable splitting scheme of the form

$$\underset{x=y=z}{\arg\min}\, f(x) + g(y) + h(z) \tag{5.1.10}$$

since if $x_n \neq y_n$, then no assignment to z_n will cause the constraint to be satisfied.

This means that this convergence theory doesn't apply to the method introduced in [28], since this algorithm involves a three-way variable splitting. What we propose is to modify the scheme to only involve splitting two variables. We rewrite the problem as

$$\underset{X=Y}{\arg\min}\, \gamma\|Y\|_1 + \chi_{S_{n,k}}(Y) + \mathrm{Tr}(X^THX) \tag{5.1.11}$$

and then use the general ADMM method analyzed in [35]. This algorithm is guaranteed to converge, however it involves solving the following sub-problem for Y_n

$$\underset{Y^TY=I}{\arg\min}\, \frac{1}{2}\|Y - V\|_2^2 + \gamma\|Y\|_1 \tag{5.1.12}$$

which is an L^1 regularized projection onto the Stiefel manifold. As far as we are aware, there is no efficient method for solving this problem for general k (the number of columns of Y). However, in the particular case when $k = 1$, i.e. we are only calculating a single compressed mode, this sub-problem can be solved efficiently, as the following result shows.

Theorem 5.1.1. *The solution to the problem*

$$\arg\min_{\|y\|_2=1} \frac{1}{2}\|y - v\|_2^2 + \gamma\|y\|_1 \tag{5.1.13}$$

is given by $proj_S(shrink_\gamma(v))$ if $shrink_\gamma(v) \neq 0$. (Here $proj_S(x) = x/\|x\|_2$ is projection onto the unit sphere, and $shrink_\gamma$ is the soft-thresholding operator.)

Proof. Let y be the minimizer and note that y must satisfy

$$\lambda y \in y - v + \gamma \text{sgn}(y) \tag{5.1.14}$$

for some λ. I will rewrite this as

$$v \in (1 - \lambda)y + \gamma \text{sgn}(y) \tag{5.1.15}$$

First consider the case when $\lambda \geq 1$ and let i be such that $y_i \neq 0$. We have $v_i \in (1 - \lambda)y_i + \gamma \text{sgn}(y_i)$. Suppose that $y_i \geq 0$ (the argument is entirely symmetrical if $y_i \leq 0$). Then this implies that $v_i \leq \gamma$. This means that either $\text{shrink}_\gamma(v_i) = 0$ or v_i has the opposite sign as y_i. The latter is absurd since replacing y_i by $-y_i$ would clearly decrease the objective. Hence $\text{shrink}_\gamma(v_i) = 0$.

This must be true for all i such that $y_i \neq 0$. However, if $y_i = 0$, then $v_i \in [-\gamma, \gamma]$ so that $\text{shrink}_\gamma(v_i) = 0$ as well. Thus $\lambda \geq 1$ implies that $\text{shrink}_\gamma(v) = 0$.

So suppose that $\lambda < 1$ and so $(1 - \lambda) > 0$. We can multiply by its inverse to obtain

$$(1 - \lambda)^{-1}v \in y + (1 - \lambda)^{-1}\gamma \text{sgn}(y) \tag{5.1.16}$$

Note that because $(1 - \lambda)$ is positive, the unique y which satisfies this is (by the definition of soft-thresholding)

$$\text{shrink}_{(1-\lambda)^{-1}\gamma}((1 - \lambda)^{-1}v) = (1 - \lambda)^{-1}\text{shrink}_\gamma(v) \tag{5.1.17}$$

Now we simply $(1 - \lambda)^{-1}$ so that y has unit norm. This yields, as desired the following formula for the minimizer.

$$y = \text{proj}_S(\text{shrink}_\gamma(v)) \tag{5.1.18}$$

□

This results in a provably convergent method for calculating a single compressed mode, i.e. for solving

$$\arg\min_{\|x\|_2=1}\langle x, Hx\rangle + \gamma\|x\|_1 \tag{5.1.19}$$

For solving the general compressed modes problem using a splitting method, we believe that an efficient procedure for solving (5.1.12) must be developed and we propose this an interesting research problem.

5.2 Feasible Methods

We now discuss non-smooth feasible methods for calculating compressed modes. Unlike splitting methods, feasible methods only consider points on the manifold and use retractions to generate new iterates. Instead of only considering the compressed modes problem, we consider the more general problem

$$\arg\min_{X^TX=I} G(X) \tag{5.2.1}$$

where $G : \mathbb{R}^{nk} \rightarrow \mathbb{R}$ is a (not necessarily smooth) convex function on the whole domain $\mathbb{R}^{nk}$. Note that the constrain $X^TX = I$ is what makes this a difficult non-convex problem.

In this section we propose a generalization of averaged subgradient descent to solve (5.2.1). We begin by giving the convergence properties of subgradient descent for convex functions. Recall the definition of the subdifferential of a convex function.

Definition 5.2.1. *Let $x \in \mathbb{R}^n$ and let $G : \mathbb{R}^n \rightarrow \mathbb{R}$ be a convex function. The subgradient of G at x is*

$$\partial G(x) = \{v \in \mathbb{R}^n : \ \forall y \ G(y) \geq G(x) + v \cdot (y - x)\} \tag{5.2.2}$$

The averaged subgradient descent iteration takes the following form

$$x_{n+1} = x_n - \gamma_n g_n, \ y_{n+1} = (1 - \alpha_n)y_n + \alpha_n x_{n+1} \tag{5.2.3}$$

with $g_n \in \partial G(x_n)$ and y_N is the output of the algorithm. Essentially, in each step we take a step in the direction given by some subgradient and output a weighted average of the iterates (with the weights determined by the parameters α_n).

The convergence properties of subgradient descent are well-known. In particular, we have the following theorems regarding convex and α-strongly convex functions (see [7]).

Theorem 5.2.1 (Theorem 3.2 in [7])**.** *Assume that G is convex and L-Lipschitz, i.e. $|G(x) - G(y)| \leq L\|x-y\|_2$. Let $T \in \mathbb{N}$ and consider iteration (5.2.3) with $\gamma_n = 1/\sqrt{T}$ (note that the step size depends on T) and $\alpha_n = 1/(n+1)$ (this implies that y_n is the average of $x_1, ..., x_n$). Then we have*

$$G(y_T) - G^* \leq \frac{L}{\sqrt{T}}\|x_1 - x^*\|_2 \tag{5.2.4}$$

Theorem 5.2.2 (Theorem 3.9 in [7])**.** *Assume that G is α-strongly convex and L-Lipschitz on a (necessarily bounded) domain $U \subset \mathbb{R}^n$. Consider iteration (5.2.3) with $\gamma_n = \frac{2}{\alpha(n+1)}$ and $\alpha_n = \frac{2}{n+2}$. Then if all the iterates are contained in U, we have*

$$G(y_T) - G^* \leq \frac{2L^2}{\alpha(T+1)} \tag{5.2.5}$$

Note that although both of these objective error bounds are quite weak, strong convexity provides a significant speed up of the convergence if the iterates are carefully averaged. Under the assumption that the objective G is strongly convex in a neighborhood of its local minimizer, this indicates that carefully averaging the subgradient descent iterates can improve performance.

In chapter 3, in the process of developing accelerated method for smooth optimization, we already showed how to take steps in a given dual tangent direction and also how to average on the Stiefel manifold. Applying these results to iteration (5.2.3) we obtain the algorithm in table (5.1) below, which generalizes averaged subgradient descent to the Stiefel manifold. However, we have not observed a significant improvement of this method over non-averaged subgradient descent (i.e. iteration (5.2.3) without the y_n iterates) when applied to the compressed modes problem. In particular, it converges very slowly and is sensitive to the initial iterate. Nonetheless, we believe it may be useful for other non-smooth optimization problems with orthogonality constraints.

```
Algorithm 3: Averaged Subgradient Descent
  Data: $G$ a function, $T$ the total number of iterations
  Result: An approximate minimizer $Y_T$
  $X_0 \leftarrow$ initial point;
  $Y_0 \leftarrow X_0$;
  $n \leftarrow 1$;
  while $n \leq T$ do
      $X_n \leftarrow R_1(X_{n-1}, \phi_g(-\gamma_n g_n))$ with $g_n \in \partial G(X_{n-1})$;
      $V_n \leftarrow 2X_n(I + X_n^T Y_{n-1})$;
      $Y_n \leftarrow R_1(Y_{n-1}, \alpha_n \phi_g(V_n))$;
      $n \leftarrow n + 1$;
  end
```

Table 5.1: Averaged Subgradient Descent

5.3 Smoothing the L1 Term

In this section, we propose smoothing the L^1 term in (5.1.1) and using our algorithms for smooth optimization with orthogonality constraints to calculate compressed modes. This will turn out to be extremely fruitful and we will obtain a robust and efficient numerical method for solving (5.0.1).

Specifically, we consider replacing the L^1 norm by its Moreau-Yosida envelope [23]

$$f_\epsilon(x) = \min_y |y| + \frac{1}{2\epsilon}(x-y)^2 = \begin{cases} \frac{1}{2\epsilon}x^2 & |x| \leq \epsilon \\ |x| - \frac{\epsilon}{2} & |x| > \epsilon \end{cases} \tag{5.3.1}$$

where ϵ is a parameter quantifying the trade-off between smoothness and closeness to the L^1 norm. This results in the following smooth optimization problem on the Stiefel manifold

$$\operatorname*{arg\,min}_{X \in S_{n,k}} \operatorname{Tr}(X^T H X) + \gamma \sum_{i,j} f_\epsilon(X_{ij}) \tag{5.3.2}$$

which we solve using the accelerated gradient descent algorithms developed in the previous sections.

5.3.1 Numerical Results

We now demonstrate this approach to calculating compressed modes. The first problem we consider is the problem of calculating compressed modes on an interval with zero Dirichlet boundary conditions. Specifically, we consider the optimization problem

$$\operatorname*{arg\,min}_{\substack{\phi_i \in H_0^1([0,1]) \\ \langle \phi_i, \phi_j \rangle_{L^2} = \delta_{ij}}} \sum_{i=1}^{k} \|\nabla \phi_i\|_2^2 + \mu \sum_{i=1}^{k} \|\phi_i\|_1 \tag{5.3.3}$$

We discretize the interval $[0, 1]$ using $n = 1000$ points and calculate $k = 20$ compressed modes. The results of this computation for various values of μ and smoothing level ϵ are shown in figures (5.1) and (5.2).

We see that the condition number of the problem becomes worse as ϵ is decreased or μ is increased, which is expected since this makes the objective much less smooth, but also that $\epsilon = .001$ seems to be small enough to accurately calculate the compressed modes. We also note that in these calculations our initial point was chosen uniformly at random on the manifold and we nonetheless found the correct optimizer. This demonstrates that our method is very robust.

In practice, we observed that choosing a better initial point reduced the number of iterations required substantially. In fact, we recommend solving the problem for larger values of ϵ to find good initial points for smaller values of ϵ when applying this algorithm in practice.

We proceed to test our method on the two dimensional compressed modes calculation shown in [28]. For this problem, the domain is the torus, i.e. $[0, 1]^2$ with periodic boundary conditions and the objective is

$$\operatorname*{arg\,min}_{\substack{\phi_i \in H^1([0,1]^2) \\ \langle \phi_i, \phi_j \rangle_{L^2} = \delta_{ij}}} \sum_{i=1}^{k} \|\nabla \phi_i\|_2^2 + \mu \sum_{i=1}^{k} \|\phi_i\|_1 \tag{5.3.4}$$

The results of our calculation with $k = 25$ and $\mu = 5$ are shown in figure (5.3) (we discretized with 200 points in each direction). As before, we start with a uniformly random initial point on the manifold. We have also run experiments with potentials and inhomogeneous Laplace operators and the method is very robust and efficient in all of these cases.

5.4 Conclusion

In this chapter, we considered the problem of calculating compressed modes. We analyzed splitting methods, subgradient descent methods, and introduced an approach which smoothes the L^1 penalty and uses the accelerated methods developed in chapter 3. While the splitting methods and subgradient methods don't come with convergence guarantees and aren't particularly robust, we provide numerical experiments which show that this new method is robust and efficient.

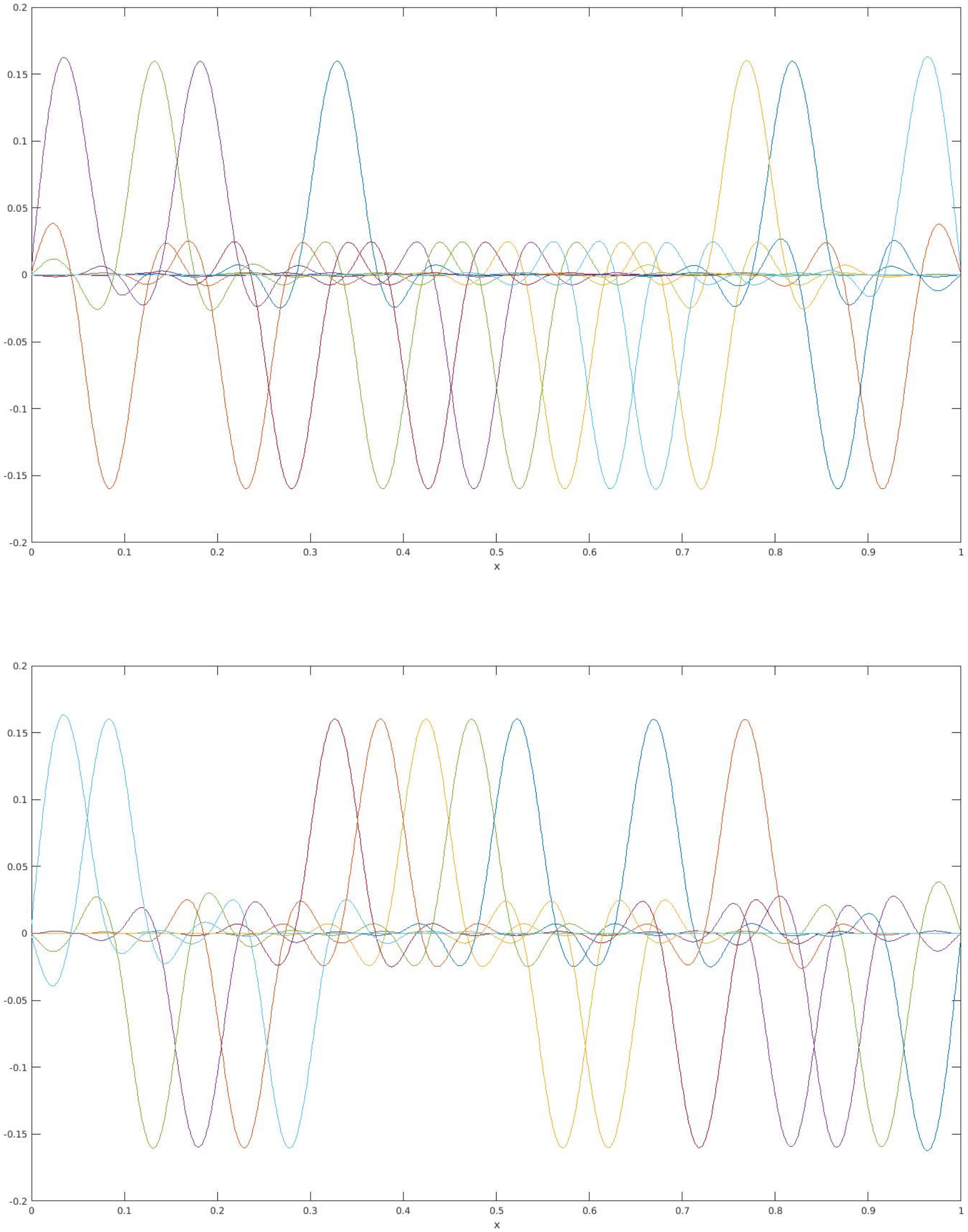
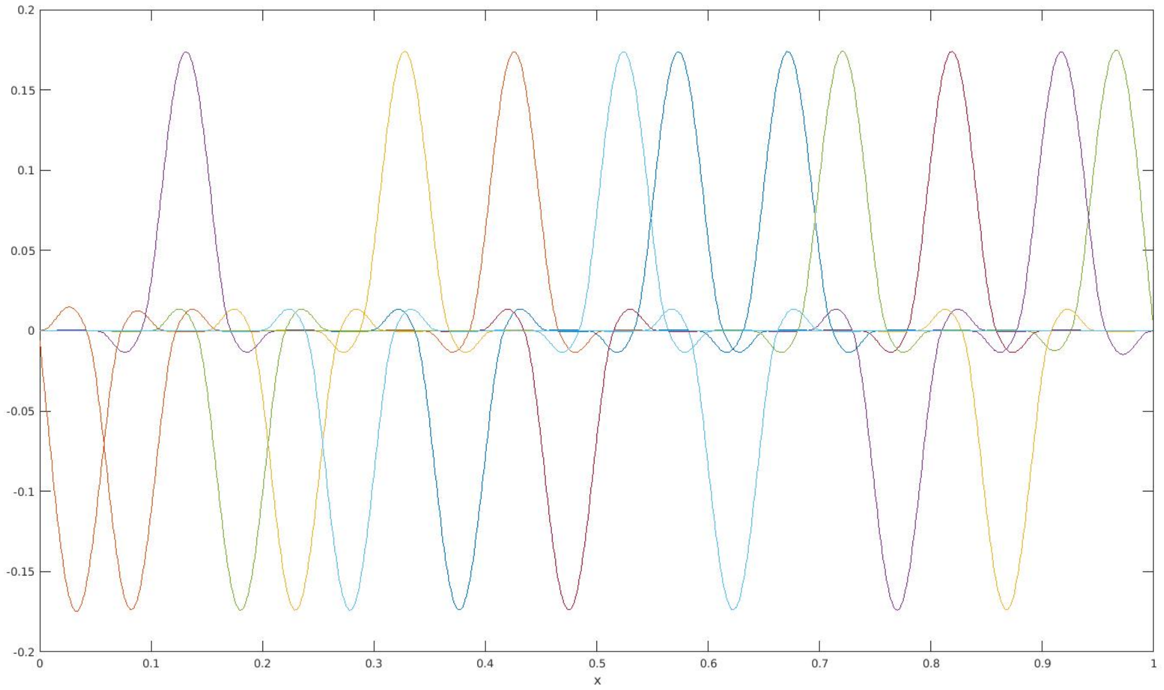


Figure 5.1: Compressed modes with $\mu = 50$, $\epsilon = .001$ (top, 2880 iterations) and $\epsilon = .0001$ (bot, 5755 iterations)



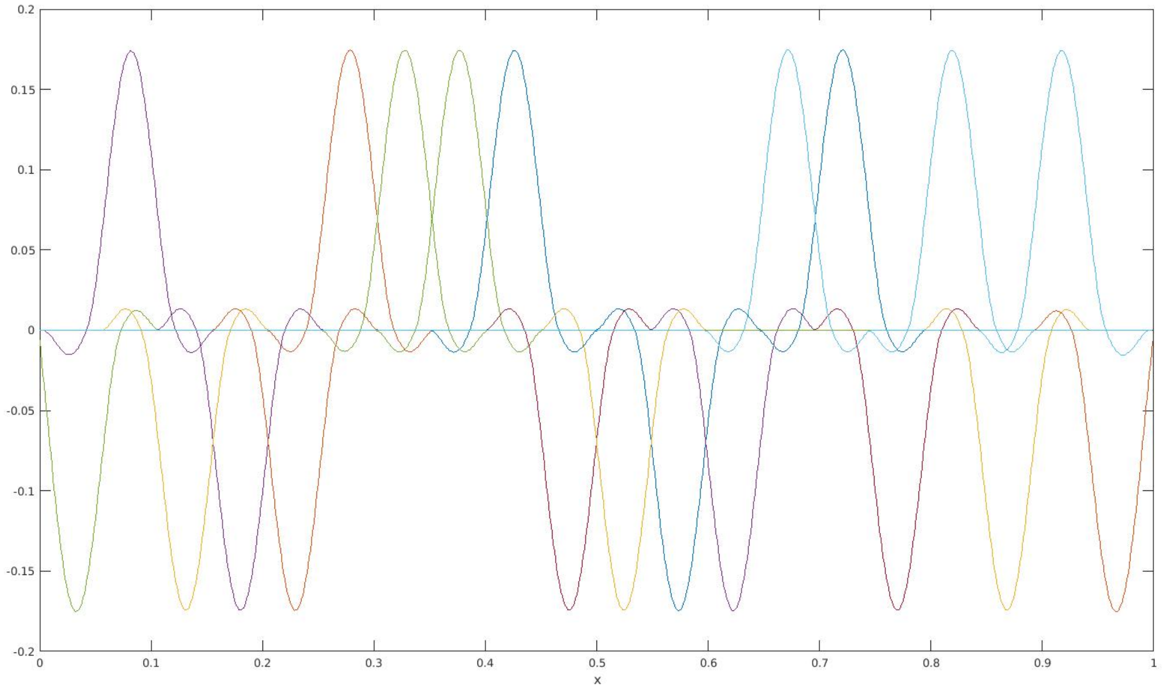


Figure 5.2: Compressed modes with $\mu = 200$, $\epsilon = .001$ (top, 5376 iterations) and $\epsilon = .0001$ (bot, 12883 iterations)

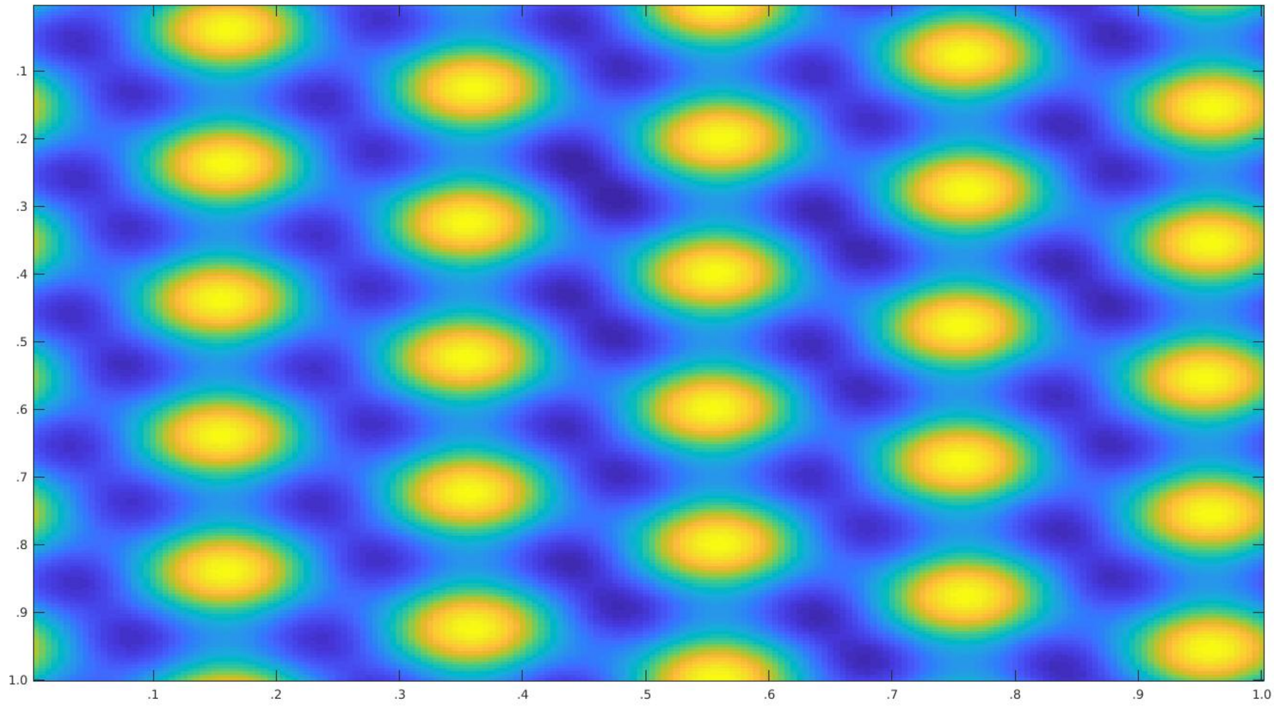


Figure 5.3: Compressed modes on the torus ($k = 25, \mu = 5$, we have plotted the sum of all of the modes)

Bibliography

[1] P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.

[2] Larry Armijo. "Minimization of functions having Lipschitz continuous first partial derivatives". In: *Pacific Journal of mathematics* 16.1 (1966), pp. 1–3.

[3] Farzin Barekat. "On the consistency of compressed modes for variational problems associated with the Schrödinger operator". In: *SIAM Journal on Mathematical Analysis* 46.5 (2014), pp. 3568–3577.

[4] Farzin Barekat, Russel Caflisch, and Stanley Osher. "On the support of compressed modes". In: *SIAM Journal on Mathematical Analysis* 49.4 (2017), pp. 2573–2590.

[5] Farzin Barekat et al. "Compressed Wannier modes found from an L_1 regularized energy functional". In: *arXiv preprint arXiv:1403.6883* (2014).

[6] Haim Brezis. "Solutions with compact support of variational inequalities". In: *Russian Mathematical Surveys* 29.2 (1974), pp. 103–108.

[7] Sébastien Bubeck et al. "Convex optimization: Algorithms and complexity". In: *Foundations and Trends® in Machine Learning* 8.3-4 (2015), pp. 231–357.

[8] Alan Edelman, Tomás A Arias, and Steven T Smith. "The geometry of algorithms with orthogonality constraints". In: *SIAM journal on Matrix Analysis and Applications* 20.2 (1998), pp. 303–353.

[9] J Willard Gibbs. *Elementary principles in statistical mechanics*. Courier Corporation, 2014.

[10] David Gilbarg and Neil S Trudinger. *Elliptic partial differential equations of second order*. springer, 2015.

[11] Gabriele Giuliani and Giovanni Vignale. *Quantum theory of the electron liquid*. Cambridge university press, 2005.

[12] Tom Goldstein and Stanley Osher. “The split Bregman method for L1-regularized problems”. In: *SIAM journal on imaging sciences* 2.2 (2009), pp. 323–343.

[13] William W Hager. “Updating the inverse of a matrix”. In: *SIAM review* 31.2 (1989), pp. 221–239.

[14] Bo Jiang and Yu-Hong Dai. “A framework of constraint preserving update schemes for optimization on Stiefel manifold”. In: *Mathematical Programming* 153.2 (2015), pp. 535–575.

[15] Walter Kohn. “Analytic properties of Bloch waves and Wannier functions”. In: *Physical Review* 115.4 (1959), p. 809.

[16] Andrej N Kolmogorov. *Über kompaktheit der funktionenmengen bei der konvergenz im mittel.* Weidmann, 1931.

[17] Jaroslav Kouteckỳ and Vlasta Bonačić. “On Convergence Difficulties in the Iterative Hartree-Fock Procedure”. In: *The Journal of Chemical Physics* 55.5 (1971), pp. 2408–2413.

[18] Rongjie Lai and Stanley Osher. “A splitting method for orthogonality constrained problems”. In: *Journal of Scientific Computing* 58.2 (2014), pp. 431–449.

[19] John M Lee. “Smooth manifolds”. In: *Introduction to Smooth Manifolds.* Springer, 2003, pp. 1–29.

[20] Qihang Lin and Lin Xiao. “An adaptive accelerated proximal gradient method and its homotopy continuation for sparse optimization”. In: *International Conference on Machine Learning.* 2014, pp. 73–81.

[21] Xin Liu et al. “On the Convergence of the Self-Consistent Field Iteration in Kohn–Sham Density Functional Theory”. In: *SIAM Journal on Matrix Analysis and Applications* 35.2 (2014), pp. 546–558.

[22] Richard M Martin. *Electronic structure: basic theory and practical methods.* Cambridge university press, 2004.

[23] Jean-Jacques Moreau. “Proximité et dualité dans un espace hilbertien”. In: *Bull. Soc. Math. France* 93.2 (1965), pp. 273–299.

[24] Yurii Nesterov. “A method of solving a convex programming problem with convergence rate O $(1/k^2)$”. In: *Soviet Mathematics Doklady*. Vol. 27. 2. 1983, pp. 372–376.

[25] Yurii Nesterov et al. *Gradient methods for minimizing composite objective function*. 2007.

[26] Louis Nirenberg. “On elliptic partial differential equations”. In: *Il principio di minimo e sue applicazioni alle equazioni funzionali*. Springer, 2011, pp. 1–48.

[27] Brendan O’donoghue and Emmanuel Candes. “Adaptive restart for accelerated gradient schemes”. In: *Foundations of computational mathematics* 15.3 (2015), pp. 715–732.

[28] Vidvuds Ozoliņš et al. “Compressed modes for variational problems in mathematics and physics”. In: *Proceedings of the National Academy of Sciences* 110.46 (2013), pp. 18368–18373.

[29] George Pólya and G Szegö. “Inequalities for the capacity of a condenser”. In: *American Journal of Mathematics* 67.1 (1945), pp. 1–32.

[30] Marcel Riesz. “Sur les ensembles compacts de fonctions sommables”. In: *Acta Szeged Sect. Math* 6 (1933), pp. 136–142.

[31] Ramamurti Shankar. *Principles of quantum mechanics*. Springer Science & Business Media, 2012.

[32] Jack Sherman and Winifred J Morrison. “Adjustment of an inverse matrix corresponding to a change in one element of a given matrix”. In: *The Annals of Mathematical Statistics* 21.1 (1950), pp. 124–127.

[33] Weijie Su, Stephen Boyd, and Emmanuel J Candes. “A differential equation for modeling Nesterov’s accelerated gradient method: theory and insights”. In: *Journal of Machine Learning Research* 17.153 (2016), pp. 1–43.

[34] Robert Tibshirani. “Regression shrinkage and selection via the lasso”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1996), pp. 267–288.

[35] Yu Wang, Wotao Yin, and Jinshan Zeng. “Global convergence of ADMM in nonconvex nonsmooth optimization”. In: *arXiv preprint arXiv:1511.06324* (2015).

[36] Zaiwen Wen and Wotao Yin. “A feasible method for optimization with orthogonality constraints”. In: *Mathematical Programming* 142.1-2 (2013), pp. 397–434.

[37] Xiaojing Zhu. “A Riemannian conjugate gradient method for optimization on the Stiefel manifold”. In: *Computational Optimization and Applications* 67.1 (2017), pp. 73–110.