5-2009

On the Applicability of Genetic Algorithms to Fast Solar Spectropolarimetric Inversions for Vector Magnetography

Brian J. Harker
Utah State University

ON THE APPLICABILITY OF GENETIC ALGORITHMS TO FAST SOLAR SPECTROPOLARIMETRIC INVERSIONS FOR VECTOR MAGNETOGRAPHY

by

Brian J. Harker

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

of

DOCTOR OF PHILOSOPHY

in

Physics

Approved by:

Jan J. Sojka Major Professor	K.S. Balasubramamiam Committee Member
Eric Held Committee Member	Lie Zhu Committee Member
Warren Phillips Committee Member	Byron R. Burnham, Dean of Graduate Studies

UTAH STATE UNIVERSITY
Logan, Utah

2008

ABSTRACT

On the Applicability of Genetic Algorithms to Fast Solar Spectropolarimetric Inversions for Vector Magnetography

by

Brian J. Harker, Doctor of Philosophy

Utah State University, 2008

Major Professor: Dr. Jan J. Sojka
Department: Physics

The measurement of vector magnetic fields on the sun is one of the most important diagnostic tools for characterizing solar activity. The ubiquitous solar wind is guided into interplanetary space by open magnetic field lines in the upper solar atmosphere. Highly-energetic solar flares and Coronal Mass Ejections (CMEs) are triggered in lower layers of the solar atmosphere by the driving forces at the visible "surface" of the sun, the photosphere. The driving forces there tangle and interweave the vector magnetic fields, ultimately leading to an unstable field topology with large excess magnetic energy, and this excess energy is suddenly and violently released by magnetic reconnection, emitting intense broadband radiation that spans the electromagnetic spectrum, accelerating billions of metric tons of plasma away from the sun, and finally relaxing the magnetic field to lower-energy states. These eruptive flaring events can have severe impacts on the near-Earth environment and the human technology that inhabits it.

This dissertation presents a novel inversion method for inferring the properties of the vector magnetic field from telescopic measurements of the polarization states (Stokes vector) of the light received from the sun, in an effort to develop a method that is fast, accurate, and reliable. One of the long-term goals of this work is to develop such a method that is capable of rapidly-producing characterizations of the magnetic field from time-sequential data, such that near real-time projections of the complexity and flare-productivity of solar

active regions can be made. This will be a boon to the field of solar flare forecasting, and should help mitigate the harmful effects of space weather on mankind's space-based endeavors. To this end, I have developed an inversion method based on genetic algorithms (GA) that have the potential for achieving such high-speed analysis.

(256 pages)

ACKNOWLEDGMENTS

First and foremost, I would like to thank my mother, Judy, and my sister, Pamela, for always being supportive and encouraging. Without them, I would not be where I am today. Thank you Aunt Karen and Uncle Ace, Aunt Susan and Uncle George for your constant support. To my grandfather, Gerald, and grandmother, Mary: thanks for watching over me all these years.

I owe a great deal of gratitude to my major advisors, Dr. Jan J. Sojka and Dr. K.S. Balasubramaniam, for their constant input, encouragement, and guidance. Thanks are also due to my other advisory committee members, namely, Dr. Eric Held (particularly for ushering me into the great wide world of Fortran), Dr. Lie Zhu, and Dr. Warren Phillips. Of course, without the support of Dr. J. Steven Hansen of the Space Dynamics Laboratory and the Tomorrow Fellowship in Solar Physics, none of the last five years would have been possible, so I would especially like to thank him.

I would also like to thank the Department of Physics Tracking Committee for their indispensable perspectives on balancing teaching with research. Along the same lines go my thanks to James Coburn for allowing me access to the demonstration store room and for his advice on teaching/demonstrating basic principles of physics and astronomy. My heartfelt thanks go to the ladies in the Physics Department Office: Karalee Ransom, Shelley Williams, Sharon Pappas, Melanie Oldroyd, and Shawna Johnson. Without them, I would (a) never find Dr. Sojka and (b) never remember to register for the next semester (I'm looking at you, Karalee)! Thanks also go to John James for opening my eyes to the evils of Micro$oft and introducing me to Linux. To Andrew Auman: thank you for providing your LaTeX template when my own, aesthetically-pleasing template was shot down for not being up to USU standards.

Many thanks go to the Director of the National Solar Observatory, Dr. Steve Keil, not only for allowing me to spend time at one of the world's premier solar observation facilities, but for making sure I have a place to live and an office in which to work. My time in Sunspot

has been some of the most enjoyable and productive over the last five years, in part due to the residents, so I would like to thank Han Uitenbroek and Craig Gullixson (for the thrice-weekly volleyball games), Mike Bradford, Chris Berst, Ali Tritschler (for providing several very nice datasets to work with), Kit and Jan Richards, Satya Balasubramaniam, Jackie Diehl, Rebecca Coleman, Brady Jones (especially for rescuing me from the occasional Linux-based emergency), Dave Dooling, Tim "Tex" Henry (for many IDL tutorials), Thomas Rimmele, and Jose Marino (for providing much-needed respite from my frustrations on several occasions). Ah! I cannot forget the cafeteria staff: Kathy Plum, Ramona Elrod, and Lou Ann Gregory for providing such siesta-worthy lunches every day.

In a more professional tone, I'd like to officially thank the *Hinode* science team for making their data publicly available. Hinode is a Japanese mission developed and launched by ISAS/JAXA, with NAOJ as domestic partner and NASA and STFC (UK) as international partners. It is operated by these agencies in co-operation with ESA and NSC (Norway).

Thanks also to Jeff Frisby, Matt "TR" Astill, Brian Quick and Ryan Bolton for befriending the new kid on the block, and to Trae Arnold and Chandler Durney for always being there when I needed a break. To anyone I may have forgotten to mention explicitly, Thank You!

Brian J. Harker

CONTENTS

LIST OF TABLES

LIST OF FIGURES

GLOSSARY

ANN: An Artificial Neural Network, a subclass of artificial intelligence/machine-learning algorithms.

AO: Adaptive Optics, a real-time feedback control mechanism based on a small, deformable mirror used to remove atmospheric distortion from ground-based telescope images.

AR: An Active Region, a grouping of magnetically-active sunspots.

ASP: The Advanced Stokes Polarimeter, a dedicated instrument of the Dunn Solar Telescope, used for recording the polarization states of observed light, as a function of wavelength.

ATST: The Advanced Technology Solar Telescope, a next-generation solar observation facility to be built (hopefully) in the not-too-distant future, which will incorporate state-of-the-art instrumentation for observing solar phenomena on never-before-seen spatial and temporal scales.

BBSO: Big Bear Solar Observatory, located on Big Bear Lake in California, USA, and operated by the New Jersey Institute of Technology.

CME: Coronal Mass Ejection, an eruption of solar plasma from the Chromospheric level out into interplanetary space, typically preceded by a solar flare.

CNO Cycle: The Carbon-Nitrogen-Oxygen cycle, a fusion process in the cores of heavy stars whereby Hydrogen is fused into Helium by using Carbon, Nitrogen, and Oxygen as catalysts for the reactions.

DELO: The Diagonal Lambda Operator method for iteratively solving the equations of Polarized Radiative Transfer as a function of optical depth.

DLSP: The Diffraction-Limited Spectro-Polarimeter, an instrument for solar polarimetry located at the National Solar Observatory in Sunspot, NM. It has better spatial resolution than the ASP.

DNC: Dynamic Niche Clustering, a strategy used in genetic algorithms for the optimization of multimodal functions.

DST: The Dunn Solar Telescope, located at the National Solar Observatory in Sunspot, NM. It is the largest solar telescope in the world.

EPC: Embarassingly Parallel Computation, a class of algorithm which naturally lends itself to a parallel implementation, so much so that it would be embarassing to run it in a serial fashion.

ES: The Echelle Spectrograph, an instrument located at the National Solar Observatory in Sunspot, NM, which is used in concert with the UBF to allow observations at any wavelength in the visible spectrum.

GA: Genetic Algorithm, a class of optimization algorithm under the umbrella classification of Evolutionary Strategies, where principles of biological evolution are interpreted in a computational sense in order to solve problems.

HAO: The High Altitude Observatory, located in Boulder, CO, is a division of the Earth and Sun Systems Laboratory within the National Center for Atmospheric Research.

LM: The Levenberg-Marquardt algorithm, a traditional adaptive gradient ascent/descent method for optimization.

LTE: Local Thermodynamic Equilibrium, a state of a gas or plasma, which is dominated by collisions between the atoms or molecules. Local energy input is quickly thermalized with the immediate surroundings.

MHD: Magnetohydrodynamics, a field of physics dedicated to the study of how gases and plasmas interact with magnetic fields.

MPI: The Message Passing Interface, a suite of computational libraries which form the framework for sending and receiving information between independently-executing programs or processes.

NCP: Net Circular Polarization, an integrated measure of the Stokes V circular polarization.

NGA: Niching Genetic Algorithm, an improvement on the traditional GA which incorporates a strategy called fitness sharing for multimodal function optimization.

NLTE: Non-Local (or Not Limited-to) Thermodynamic Equilibrium, a state of a gas or

plasma where radiative effects are as important as collisional effects. Local energy input may be quickly thermalized with the immediate surroundings, or may be transported to spatially-distant regions via radiation.

NOAA: The National Oceanic and Atmospheric Administration, is America's oldest national scientific agency, which focuses on arenas of research from the surface of the Sun to the depths of the ocean floors, and everything in between.

NSO: The National Solar Observatory, located at Sunspot, NM, and run by the Associated Universities for Research in Astronomy. It is home to the Dunn Solar Telescope, Evans Coronagraph, and the Optical Solar Patrol Network, and as such, is one of the premier solar observation facilities in the world.

PCA: Principle Components Analysis, an algorithm for pattern-matching/pattern-classification based on similarities between objects in a known database and the object to be matched or classified.

PRTE: The Polarized Radiative Transfer Equations, which describe the attenuation of the intensity of light as well as its polarization states as it propagates through a medium.

SGA: Simple Genetic Algorithm, a minimal combination of selection, crossover, and/or mutation operators that perform an evolutionary optimization.

SRTE: The Scalar Radiative Transfer Equation, which describes the attenuation of only the intensity of light as it propagates through a medium.

TRACE: Transition Region and Coronal Explorer, a NASA Small Explorer satellite mission to image the transition region and corona in EUV and X-Ray wavelengths at high angular and temporal resolutions.

UBF: Universal Birefringent Filter, an instrument located at the National Solar Observatory used in combination with the Echelle Spectrograph to allow observations to be made at any wavelength in the visible spectrum.

VTT: Vacuum Tower Telescope, the previous name of the Dunn Solar Telescope. The two are typically used interchangeably, although there is an observatory properly named the Vacuum Tower Telescope located at Tenerife in the Canary Islands, operated by the

Kiepenheuer Institute of Solar Physics.

CHAPTER 1
INTRODUCTION AND MOTIVATION

New results in solar physics are being published every day using the data from current solar observation facilities (Dunn Solar Telescope (DST), Vacuum Tower Telescope (VTT), Dutch Open Telescope (DOT), etc.). Despite their tremendous data-gathering capabilities, they are ultimately limited simultaneously in spatial, spectral, and temporal resolution. Because of these limitations, many interesting phenomena presumed to exist at small scales are unobservable. Next-generation observation facilities (e.g., Advanced Technology Solar Telescope (ATST)) will alleviate some of these limitations, and open up new windows through which never-before-seen solar phenomena can be observed. For instance, the maximum spatial resolution of the DST at the solar surface is approximately 0.1 arcseconds in the visible spectrum, increasing to about 0.33 arcseconds in the infrared. However, because of turbulence in the Earth's atmosphere, features of this scale are typically not observable, and in practice the resolution of the DST is closer to, but typically less than, 1 arcsecond. For comparison, the ATST is expected to be constructed with a 4-meter primary mirror, and thus will have a maximum spatial resolution of approximately 0.03 arcseconds on the photospheric surface. Furthermore, with the use of multi-conjugate adaptive optics at the ATST, atmospheric turbulence will be mitigated, and practical resolution will be very close to the theoretical diffraction-limited case. This will give ATST the ability to observe phenomena on linear scales of approximately 30 km. Such a drastic increase in resolution will supply solar astronomers with a flood of data orders of magnitude greater than can be obtained from current instrumentation. As a result, I expect that the most widely-used polarimetry data reduction techniques will produce a bottleneck in the flow of information. Current polarimetry datasets taking X hours to computationally reduce and interpret will be minuscule compared to the datasets expected to be obtained from future solar instrumentation. Using the same computational techniques, these future observations will take Y hours to reduce on the same computer architecture, where $Y \gg X$. A rough estimate for

X and Y can be obtained as follows: the pixel scale for the Advanced Stokes Polarimeter (ASP, located at the National Solar Observatory, Sunspot, NM) is about 0.525" $\times$ 0.36". At the Earth's orbital distance, one arcsecond corresponds to a linear scale of about 725 km on the visible surface of the Sun. Since the expected spatial resolution for the ATST is roughly 30 km (diffraction-limited), this represents a factor of ~ 24 improvement in spatial resolution, ultimately yielding an increase by a factor of $\sim 24^2$ in the number of pixels required to cover the same field-of-view imaged by the ASP. The data and results to be presented in the following chapters show that a full field-of-view ($\sim 68,700$ pixels) processing of ASP data with the techniques developed in this dissertation requires roughly 2.5 hours. To process the same field-of-view using ATST polarimetry data would therefore require ~ 60 days! Therefore, in order to obtain relevant results in any reasonable amount of time, new and potentially faster techniques must be developed for handling such large amounts of data.

This dissertation presents the exploration of a technique used to infer magnetic field strength and orientation on the solar surface from observed linear and circular polarization profiles. This type of problem is an example of an *inverse* problem, wherein the desired model parameters must be extracted from the data, taking into account the interaction(s) of the physical system being studied with the measurement apparatus (in this case, the telescope and its associated instrumentation). Conversely, the *forward* problem is quite easily solved; given a set of model parameters, it is trivial to calculate the intensity and polarization profiles resulting from the physical system specified by those parameters. It is, however, not easy to infer the model parameters from the intensity and polarization profiles due to, for example, instrumental noise, atmospheric turbulence, and limited resolution. The *inversion* method presented in this work seeks to solve the inverse problem by effectively utilizing the solution(s) to the forward problem in an efficient and straightforward way with a class of evolutionary optimization algorithms known as *genetic algorithms*. The technique is almost optimally-suited for parallelization across a large number of independent processing nodes in a cluster, and is expected to scale in a nearly perfect linear fashion with the number

of CPUs while showing little sign of an execution speed plateau. Other current inversion codes utilize inherently serial iterative algorithms which cannot be easily parallelized and have long execution times. The merit of these techniques is their accuracy; there is a trade-off between long execution times and the large precision of their results. On the other hand, the technique presented herein sacrifices 8-decimal-place accuracy for speed while still obtaining reasonably accurate results. Accuracy is always of key importance, but for space weather prediction efforts, I believe that attaining precision in the last two decimal places of a parameter describing a magnetic atmosphere at the expense of drastically-increased execution time is not worth the effort, and obtaining accurate but slightly less precise results much faster is far more efficient and productive. The potential speed-up factor of this technique should allow it to efficiently reduce large sets of data, with the ultimate goal of (near) real-time magnetic field "measurements." As such, I believe the technique shows promise for reducing the information bottleneck to which I alluded earlier.

In order to understand the promises and pitfalls of such a technique, one must first understand the structures to be observed. This dissertation begins with a review of the star Sol (our Sun) and its structure, from its innermost regions to the boundaries of its "sphere of influence", with particular emphasis on solar active regions (ARs), which are areas of concentrated magnetic fields that give rise to much (if not all) of the interesting behavior associated with the sun. Our tour of the Sun begins with a general overview of its structures and properties.

1.1 The Sun, from the Core Out to Infinity

Our star, the sun, is a G2-class star that is approximately 4.6 billion years old. It represents a typical "middle-age" star, as can be seen in Figure 1.1. The figure shows the *Hertzsprung-Russell diagram*, which catalogues the evolutionary track of stars, from their birth as hot, luminous young stars to their main sequence lifetimes and finally to stellar death. The horizontal axis represents surface temperature, while the vertical axis usually denotes the difference in apparent magnitude of the star as seen in two different

filters (wavelength bandwidths), or the luminosity (brightness) of the star. The "normal" path of stellar evolution seen moving from the upper left (hot, young stars) to the lower right (cooler, older stars) is the *main sequence*. White dwarf stars (the remnants of type I supernovae) are seen in the lower left corner of the diagram, while red giant stars (stars that have exhausted the supply of hydrogen in their cores and are fusing helium in a shell around the core) are seen in the upper right. The sun is expected to become a red giant in approximately 7.6 billion years, at which time its expansion will take its boundaries out past the current orbit of Mars. Luckily, this is not projected in the near future.

Our sun has a mean mass of 1.99×10^{30} kg and a mean radius of 6.96×10^{5} km. It currently produces an average luminosity (energy emitted per unit time) of 3.83×10^{26} W, although this has been determined to vary by as much as 0.4% by Wilson (1980). The vast majority of its mass is composed of Hydrogen and Helium, respectively denoted by the fractional abundances, X and Y, which take the values 0.73 and 0.25. This is interpreted as saying the sun is composed of approximately 73% Hydrogen and 25% Helium, with the remaining 2% being the sum of heavy metals (all elements heavier than Helium).

The sun rotates around an axis that is roughly perpendicular to the plane of the solar system. However, being a large sphere of (ionized) gas, it does not rotate as a solid body. Instead the sun possesses what is called *differential rotation*, with the equator executing one full rotation every ~ 26 days, while the regions nearer to the poles rotate once every ~ 30 days. This differential rotation plays a vital role (as will be seen later) in the regeneration of solar magnetic fields and their associated dynamical behavior.

Since the re-discovery of the telescope by Galileo Galilei in the 17^{th} century, man has been able to observe visible signs of solar magnetic fields, though at the time it was unknown that magnetic fields even existed, much less that they caused the dark areas on the solar disc now known as sunspots. Nevertheless, the numbers of visible sunspots were recorded, and over the centuries a pattern began to emerge. This pattern consists of a regular oscillation of the number of visible sunspots with a roughly constant period of ~ 11 years. This pattern has come to be known as *the sunspot cycle*, and the data describing it (all

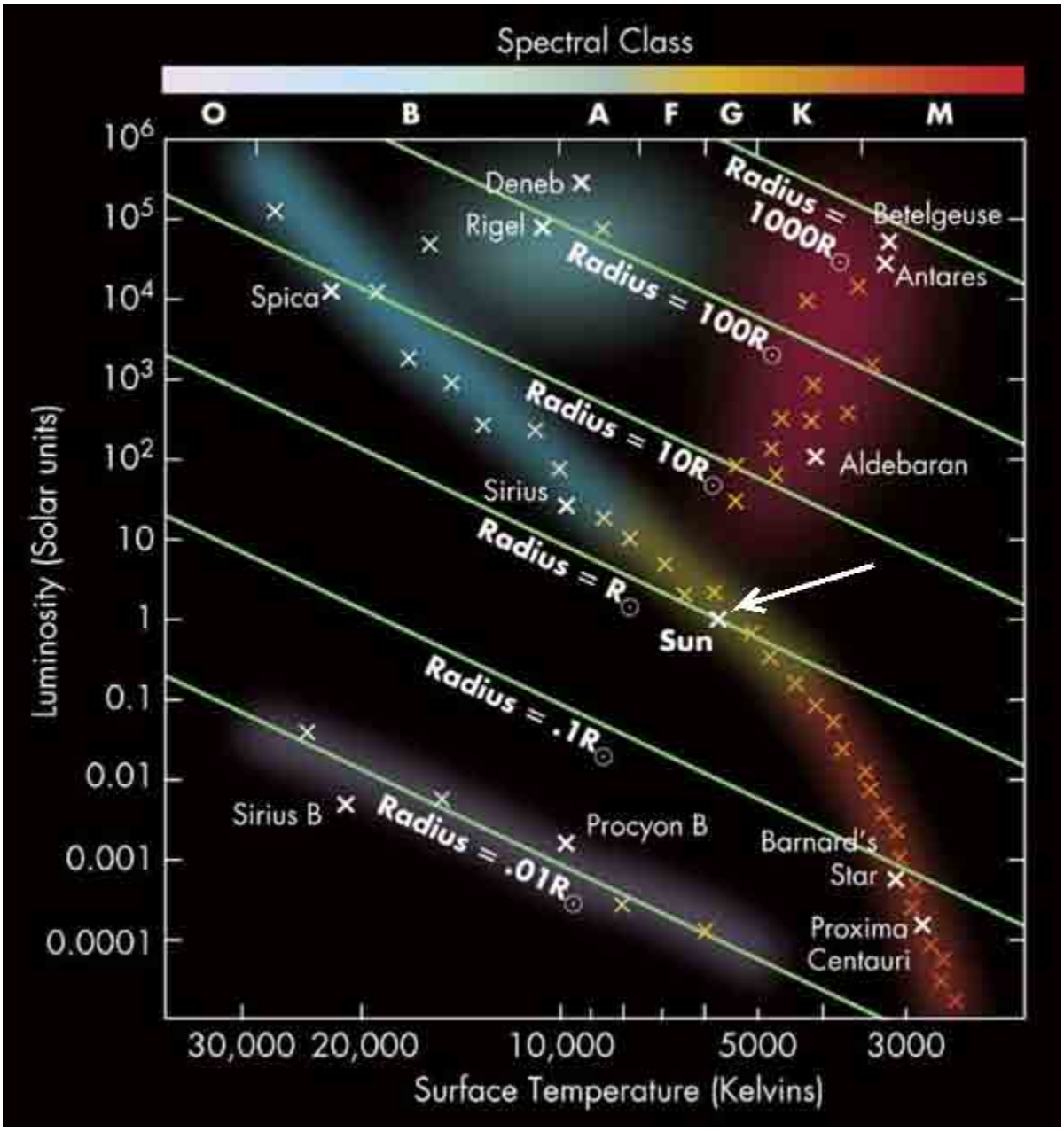

FIG. 1.1 The Hertzsprung-Russell diagram. This chart shows the path of stellar evolution, from birth to death. The position of our Sun is located just to the right of center, indicated by the arrow. From Arny and Schneider, *Explorations: An Introduction to Astronomy.*

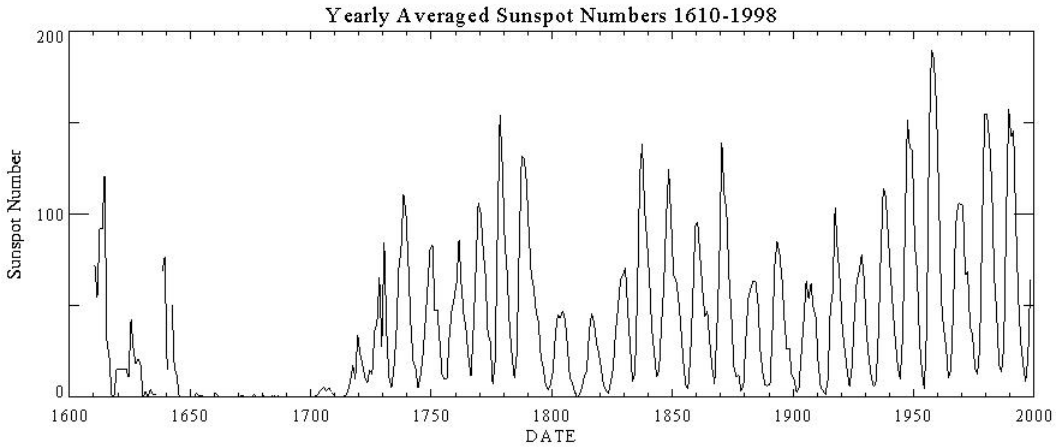

FIG. 1.2 Yearly averaged sunspot numbers, 1600–2000 A.D. Note the roughly constant period, but the widely-variable amplitude of the oscillations. Of particular interest is the period of little-to-no sunspots from 1645–1715 A.D., known as the Maunder minimum. Conversely, are we approaching a maximum period where a large number of sunspots will exist for many years? We will have to wait to find out. Courtesy of NSO.

the way back to the 17^{th} century) can be seen in Figure 1.2. The observational difference between solar maximum and solar minimum is shown in Figure 1.3. To date, it is still not precisely known what causes the sunspot cycle, although several dynamo theories seem promising. These dynamo theories attempt to describe the feedback mechanisms between the growth of solar magnetic fields, solar convection, and solar rotation. The feedback processes lead to oscillatory solutions to the equations of MHD that approximate the behavior of the solar magnetism, to zeroth order only. The details and fine structure of the real sun and its magnetic field(s) are still somewhat beyond the capabilities of these theories, although as computational techniques/power and observational resolution are improved they are expected to draw closer to reality.

A successful dynamo theory about the (re-) generation of solar magnetic fields not only must be able to explain the solar sunspot cycle, but must also reproduce several empirically-determined "laws" about the behavior of sunspots and their associated magnetic fields. These laws are as follows:

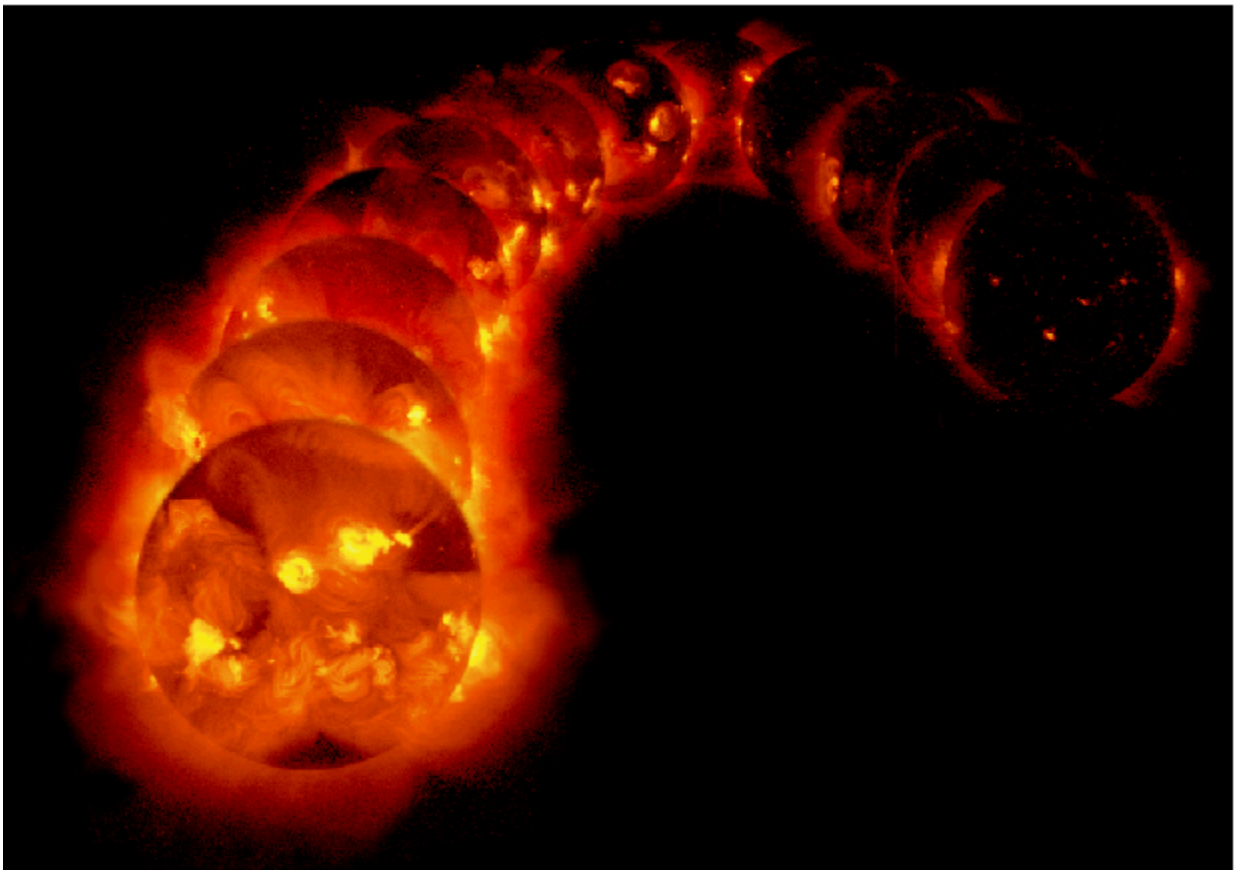

FIG. 1.3 A mosiac of full-disc X-ray images between solar maximum and solar minimum. The figure shows the progression from solar maximum *(left)* to solar minimum *(right)*, observed by the Yohkoh Soft X-Ray Telescope imager during the end phase of solar cycle 23. The images show the x-ray activity of the sun, with bright areas corresponding to regions where magnetic energy in the photosphere is being converted to thermal energy of the coronal plasma.

(1) Joy's Law: The line joining the centers of the opposite polarities of a bipolar sunspot pair typically show some inclination relative to the east-west direction of the solar equator. The amount of tilt is roughly constant, although it does sometimes show behavior that suggests the amount may be a weak function of solar latitude. This phenomenon is shown in Figure 1.4, and is presumed to be a result of the Coriolis force acting on a horizontal flux tube as it rises through the convective zone.

(2) Hale's Law: In opposite hemispheres, the leading sunspots (leading in the direction of the solar rotation) of a bipolar sunspot pair typically have opposite polarities. At

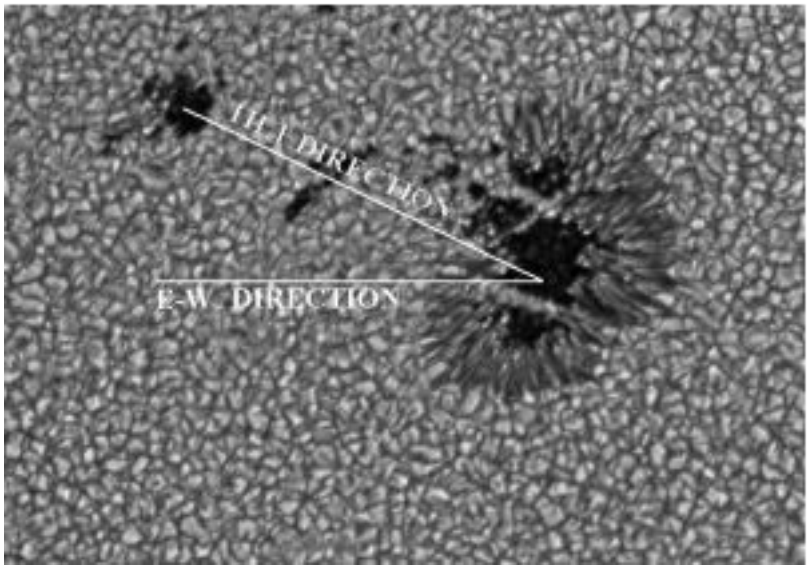

FIG. 1.4 Joy's Law for sunspot equatorial inclination. The inclination of the sunspot pair (in both hemispheres) is such that the trailing spot lies further away from the solar equator, i.e., has a larger solar latitude. Courtesy of NSO.

the start of a new sunspot cycle, the leading polarities are reversed. This is shown in Figure 1.5 and is presumed to be caused by a combination of the stretching action of the differential rotation and the buoyant rise of the stretched magnetic fields.

(3) Spörer's Law: As the 11-year sunspot cycle progresses, sunspots and active regions tend to form in latitudinal bands that draw closer to the equator. That is, at the start of a new cycle, sunspots are more likely to be found in bands centered around $\pm 45^{\circ}$ latitude. Towards the end of the cycle, these formation bands are most likely to be found near the equator. The spatio-temporal behavior described by Spörer's Law is shown in Figure 1.6, typically called a *butterfly diagram*, for obvious reasons.

Figure 1.7 shows a cartoon flow-chart schematic of one such proposed dynamo mechanism. The action of the Coriolis force on a rising flux tube is typically called *the alpha-effect*, while the solar differential rotation is dubbed *the* Ω *effect*. Therefore, the feedback scenario shown in the figure has come to be known as the $\alpha\Omega$ dynamo.

Of course, the surface magnetism that we observe is inexorably intertwined with the

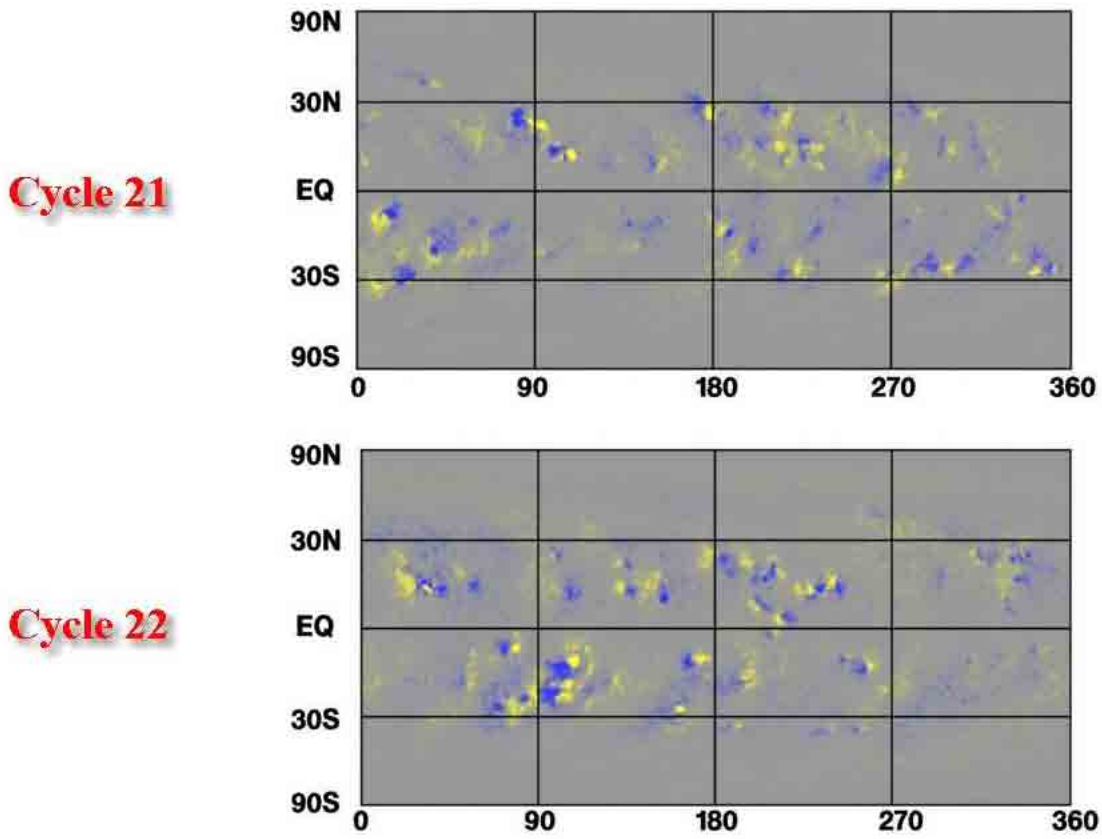

FIG. 1.5 Hale's Law for sunspot polarities. The image displays the Carrington rotation maps (planar projections) of surface magnetic fields in two consecutive sunspot cycles. Positive-polarity magnetic fields are shown in blue, and negative-polarity fields in yellow. The direction of the solar rotation is left to right, and the opposite polarity of the leading spots in opposite hemispheres is clearly visible. From solarscience.msfc.nasa.gov.

behavior of the solar plasma and magnetic fields *below* the surface layers that we can readily observe. In the following subsections, I will summarize the properties and behaviors of the most important regions within the solar interior and exterior atmosphere.

1.1.1 The Core

The sun's core occupies a spherical volume that extends out to $\sim 20\%$ of the full solar radius. Within this extremely hot and dense region with temperatures over 15 million Kelvin, the gravitational contraction of the overlying layers allows thermonuclear fusion to take place, (re-)generating the vast amounts of energy that are required to keep the

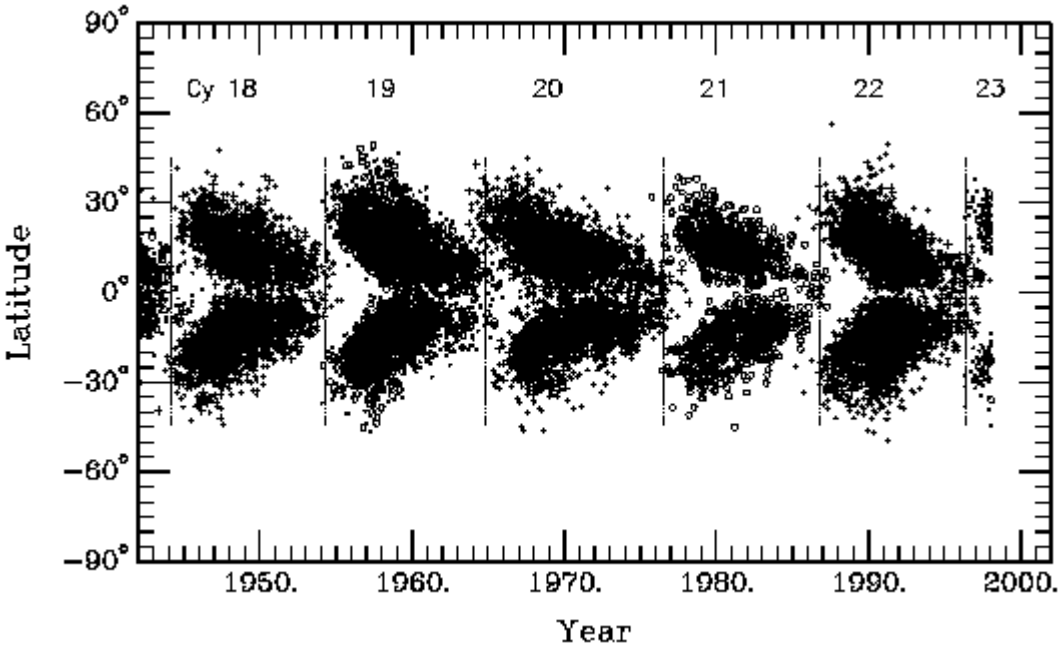

FIG. 1.6 Spörer's Law for sunspot formation latitudes. The migration of the latitudinal sunspot formation bands toward the solar equator is evident. Each cycle starts over, with sunspots first appearing at mid-latitudes, while as the cycle progresses, sunspots form at subsequently lower latitudes. Courtesy of Lowell Observatory.

sun shining. This energy is the free energy resulting from the fusion of light nuclei into subsequently heavier nuclei. An elementary example of this process follows. Assume 2 protons and 2 neutrons are fused together to form a Helium nucleus. As it turns out, the mass of the Helium nucleus is *not* the sum total of the proton and neutron masses, and, in fact, is slightly less than the total mass of the constituents. This difference is mass can be directly translated to energy via Einstein's famous relation, $E = mc^2$, and this extra little bit of mass is converted directly into radiant energy in the fusion process. This is a somewhat simplified view of nuclear fusion. In fact, the fusion processes occurring in the sun's core are significantly more complicated, involving many intermediate stages.

The fusion process responsible for producing the vast majority of the solar luminosity is called the *P-P chain*. In this process, protons are successively fused together to form Helium-4. This is not a single-path process, however, since Helium-4 can be produced by

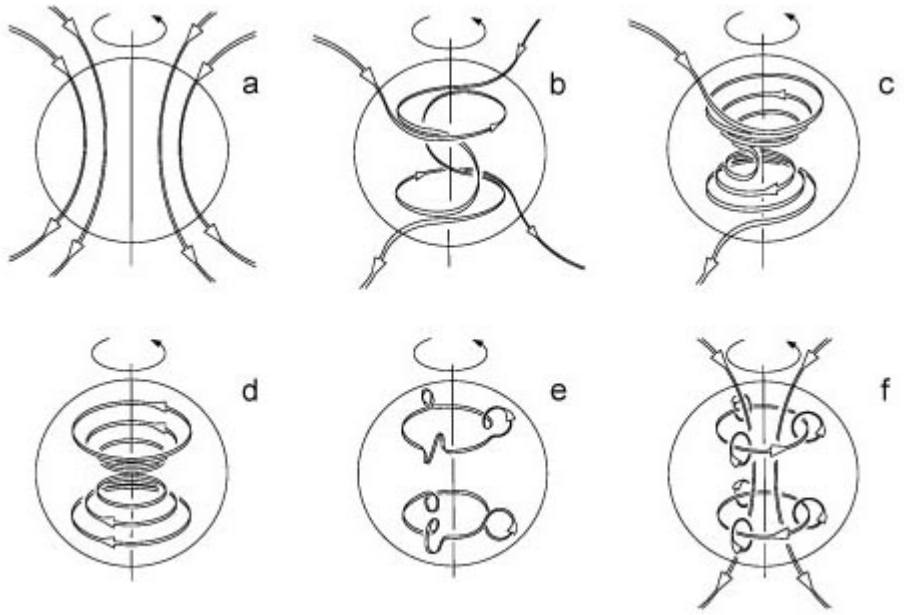

FIG. 1.7 The $\alpha\Omega$ solar dynamo. The figure shows an initially poloidal seed-field (perhaps the magnetic field of the Milky Way galaxy) as it is stretched by the differential rotation to obtain a toroidal component (frames (a)-(d)). This toroidal field becomes unstable and rises through the convective zone, as in frame (e). During its travel time, the flux tube is buffeted by convective cells that rotate due to the Coriolis force. This has the effect of twisting the rising flux tube so that by the time it reaches the surface, it is no longer parallel to the equator, but inclined by some amount relative to it. Once the flux tube has pierced the surface, it's footpoints are interpreted as a bipolar sunspot pair (frame (f)). As the sunspots decay and their flux carried toward the poles by the meridional circulation, the global (dipolar) magnetic field reverses, and the dynamo process repeats with the sunspot pairs reversing their leading polarities, in qualitative agreement with observations of consecutive sunspot cycles. From Love (1999).

three (3) different fusion mechanisms, all occurring with distinct probabilities:

$$ {}^{1}_{1}H + {}^{1}_{1}H \longrightarrow {}^{2}_{1}H + e^{+} + \nu_e \tag{1.1} $$

$$ {}^{2}_{1}H + {}^{1}_{1}H \longrightarrow {}^{3}_{2}He + \gamma \tag{1.2} $$

$$ 2\,{}^{3}_{2}He \longrightarrow {}^{4}_{2}He + 2\,{}^{1}_{1}H \tag{1.3} $$

This branch (branch I) of the P-P chain has an occurrence probability of 69%. An alternate

branch (branch II) of the P-P chain may occur approximately 31% of the time, as seen below.

$$^{3}_{2}He +^{4}_{2} He \longrightarrow\ ^{7}_{4}Be + \gamma \tag{1.4}$$

With 99.7% probability, the following reactions take place using the product of branch II:

$$^{7}_{4}Be + e^{-} \longrightarrow\ ^{7}_{3}Li + \nu_e \tag{1.5}$$

$$^{7}_{3}Li +\ ^{1}_{1}H \longrightarrow\ 2\ ^{4}_{2}He \tag{1.6}$$

With 0.3% probability, the following reactions take place using the product of branch II:

$$^{7}_{4}Be +\ ^{1}_{1}H \longrightarrow\ ^{8}_{5}B + \gamma \tag{1.7}$$

$$^{8}_{5}B \longrightarrow\ ^{8}_{4}Be + e^{+} + \nu_e \tag{1.8}$$

$$^{8}_{4}Be \longrightarrow\ 2\ ^{4}_{2}He \tag{1.9}$$

1.1.2 The Radiative Zone

The radiant energy (photons, neutrinos) produced by fusion reactions in the core must travel outwards towards the solar surface, or else the sun would not appear bright. The radiative zone is a relatively low-density (compared to the core), low-opacity region extending from the core out to about 70% of the solar radius, where the dominant energy transport mechanism is direct radiation, as opposed to thermal conduction or convection. The dominant source of opacity in the outer layers of the sun is H^{-} photoabsorption, but at the high temperatures characteristic of the radiative zone, the extra (weakly) bound electron cannot be retained by the ion, decreasing its contribution to the opacity. Therefore, photons generated in the core can easily travel through the radiative zone, almost unimpeded.

Helioseismic analyses of l- and p-mode oscillations present in the sun (essentially sound waves) have indicated that the radiative zone does not rotate as in the rest of the sun (Basu and Antia (2003)). Instead of the differential rotation, the radiative zone *rotates like a*

solid body. Current theories suggest the presence of a "fossil" magnetic field, essentially the poloidal seed-field used in the $\alpha\Omega$ dynamo scenario, as strong as 1 milligauss, which, when coupled to the low opacity of the radiative zone plasma, inhibits the differential rotation (Kichatinov and Rüdiger (1996)). However, until greater resolution in the helioseismic kernel and/or observational techniques is achieved, definite conclusions about some of the properties of the radiative zone are difficult to produce.

1.1.3 The Overshoot Region (Tachocline)

Lying on top of the radiative zone is a very narrow layer of highly-sheared plasma called the overshoot region, or the tachocline. The plasma in this layer is sub-adiabatically stratified, and therefore is stable against the onset of convective motion. It is this layer that is theorized to store strong magnetic fields that, through instability mechanisms, become buoyant and rise to the surface to produce the visible magnetic phenomena observed on the solar surface. It is also theorized to be intimately linked with the oscillatory nature of the large-scale solar magnetic field, in what has come to be known as the *solar dynamo*. Very little is precisely known about this region, shown in Figure 1.8, and it remains a topic of much debate among the solar physics community.

1.1.4 The Convective Zone

Extending upward from the tachocline to the solar surface is the region known as the convective zone. As the name implies, this a region of strong plasma convection where the energy produced in the core is transported outward much more efficiently by thermal convection than by radiation, as is the case in the radiative zone.

Because of the high opacity in the tachocline, radiant heat tends to "pile up" there, effectively turning it into a heated slab. In this thin boundary layer, heat conduction plays a more important role, because the vertical temperature gradient is large, the scale-length, d, is much smaller than in the base of the convective zone. The Rayleigh number, Ra,

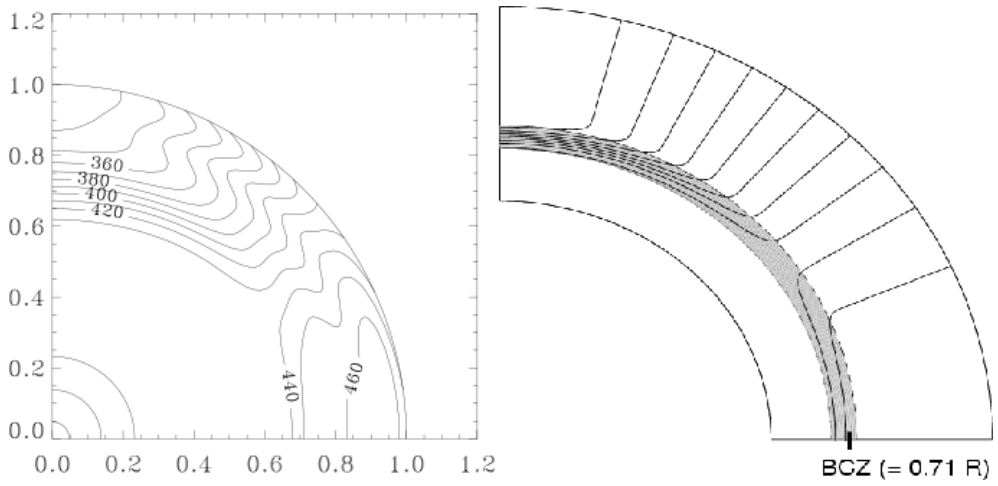

FIG. 1.8 A visualization of the solar tachocline. *(Left frame):* A map of the subsurface differential rotation profile inferred from a 2 year helioseismological observing campaign (Charbonneau et al. (1999)). The contours denote surfaces of constant rotational velocity, labeled by their p-mode frequencies in nHz. Note the thin layer of strongly-sheared plasma at the base of the convective zone (dashed line), and the qualitative agreement with the simplified model for mid-to-low latitudes. *(Right frame):* A simplified model of the subsurface differential rotation from Chatterjee et al. (2004), typical of tachocline simulation studies, showing that the rotation velocity is not constant on cylindrical surfaces, but rather on conical surfaces.

describes the relative importance of convection to conduction, and is given by:

$$Ra \equiv \frac{g\alpha\Delta T d^3}{\nu\kappa}, \tag{1.10}$$

where g is the local gravitational acceleration, α is the thermal expansion coefficient, ν is the viscosity, and κ is the thermal diffusivity. The temperature difference between the fluid element and a point far-removed from the fluid is denoted by ΔT. The onset of turbulent motion (convection) occurs at a critical Rayleigh number:

$$Ra_c = \frac{\pi^4(1+d^2)^3}{d^2}. \tag{1.11}$$

Therefore, at a point outside the boundary layer between the tachocline and the convective zone where the stratification is just such that $Ra \geq Ra_c$, convective upwelling motions begin, and the logarithmic temperature gradient,

$$\nabla \equiv \frac{dlogP}{dlogT} \tag{1.12}$$

surpasses the adiabatic temperature gradient,

$$\nabla_{ad} \equiv \left(\frac{dlogT}{dlogP} \right)_{ad}. \tag{1.13}$$

Therefore, convective cells begin to appear, in which the hot, expanding plasma moves upwards, away from the heat source of the tachocline. When it is sufficiently far away, the radiant flux on the plasma decreases, and hence the plasma begins to cool, become denser, and falls back toward the tachocline. This whole process then repeats *ad infinitum* if the heat source is constant. The convective cells do not travel all the way to the surface, however, and the cells themselves have a range of statistical sizes/volumes centered around some typical value that is dependent on the properties of the plasma (e.g., pressures, temperatures). Due to an energy/vorticity/enstrophy cascade, as one approaches the surface where the plasma density decreases, the convective cell sizes also decrease. At the surface of the convective zone, the convective cells typically have a scale length of ~ 1 Mm (~ 1000 km). These convective cells have highly irregular boundaries, and cannot be described by simple Rayleigh-Bénard convection since their vertical extents are typically of the same order as their horizontal extents. The inclusion of magnetic fields also complicates the convection, as they are swept around and merged by the upwelling and downfalling motions of the convective cells themeselves.

1.1.5 The Photosphere

The "surface" of the sun is a thin layer on the order of ~ 100 km thick, called the photosphere. The quotation marks are meant to warn that the sun does not strictly have a solid surface, although we tend to think of the photosphere as the visible surface of the sun, since this is approximately the layer we can see in a "white-light" image integrated over the visible part of the electromagnetic spectrum. Figure 1.9 shows a large, complex sunspot, surrounded by the photospheric granulation. The tops of the convective cells are readily apparent, as is the dark intercell (or intergranule) network, where cooling plasma is settling back into the convective zone proper. The dynamical timescale of the convection cells is of the order of 10–15 minutes, during which the convective plasma is fully overturned and

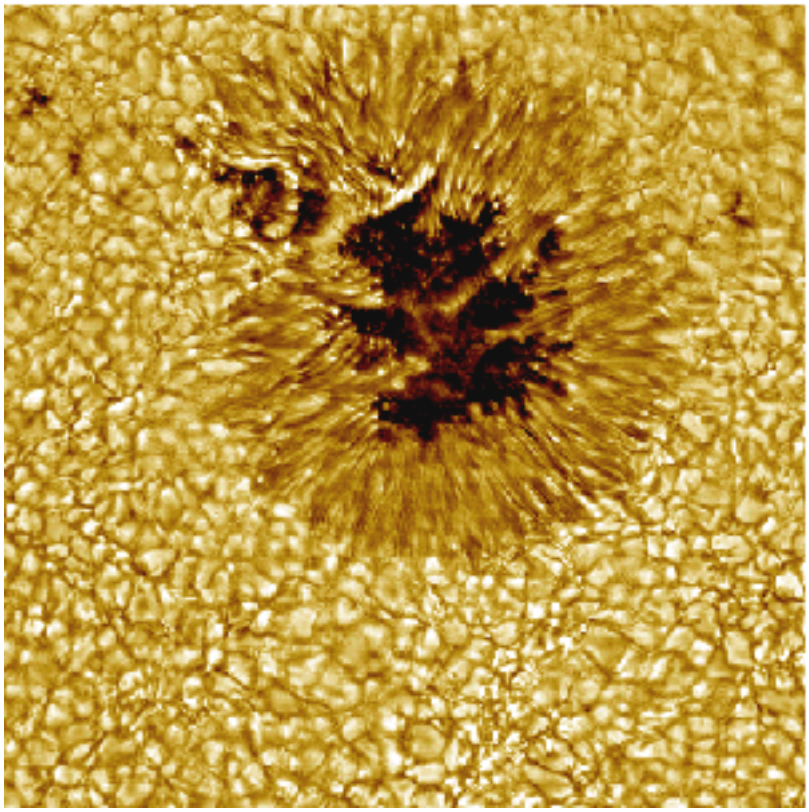

FIG. 1.9 A typical sunspot. The figure shows a white-light (integrated visible spectrum) image of a sunspot and the surrounding convective cells. The dark umbral and filamentary penumbral structures are clearly seen, as is the contrast of the solar granulation. Courtesy of NSO.

replaced by more upwelling plasma.

Figure 1.10 shows a full-disc, white-light image in which many sunspot groups can be seen. Also apparent in the figure is the phenomenon of *limb-darkening*, which can be explained as follows: the Eddington approximation states that the integrated emission along the line-of-sight has its largest contribution from layers at or above an optical depth $\tau = 2/3$. Since the optical depth along a ray-path is an integrated measure of the transparency of the medium, lines-of-sight that pass through more of the medium will have greater optical depth than those that pass through a lesser amount of the same medium. As the line-of-sight moves from disc-center out toward the limb, it passes through more photospheric plasma, due to the spherical curvature of the solar surface, and so towards the limb, the geometrical

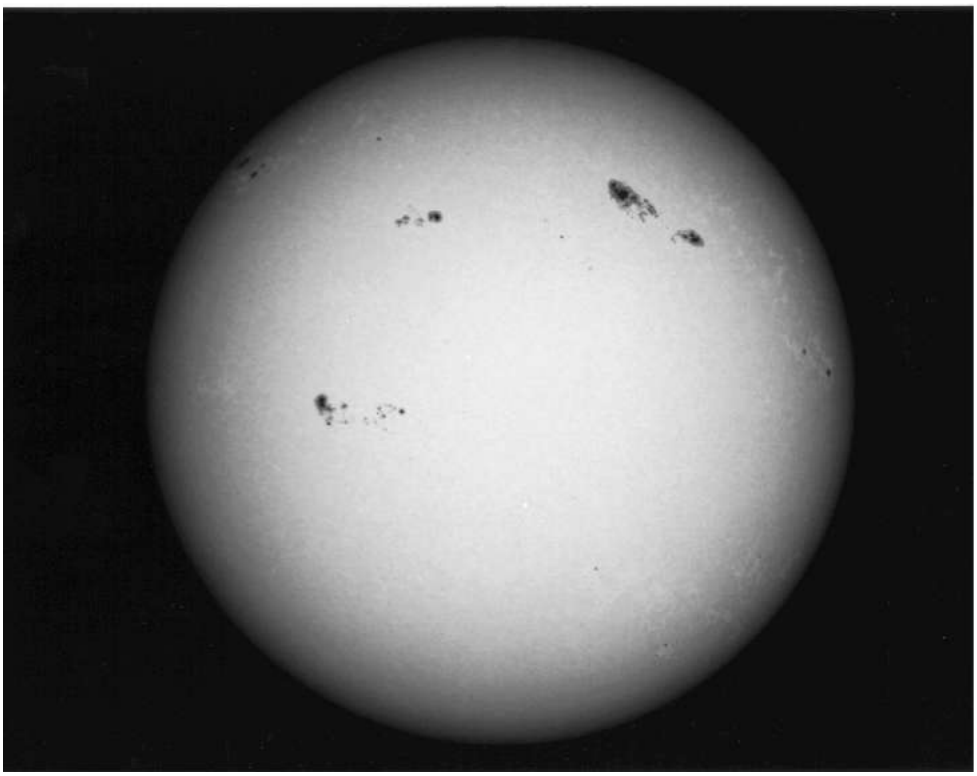

FIG. 1.10 A full-disc image of the solar surface. The figure shows a white-light image of the visible "surface" of the sun during an active period in the sunspot cycle. For comparison, the spatial extents of the sunspot groups in this image are many times the size of the Earth. Also note the limb-darkening effect, as described above.

depth at which $\tau = 2/3$ is less than that at disc-center. Therefore, when observing the limb, one is actually seeing higher altitude, cooler layers of the photosphere. Because of this, the natural thermal emission over the visible spectrum is less than at disc-center, producing the decrease in observed white-light integrated intensity as a radial function of position on the solar disc.

1.1.6 The Transition Region and Chromosphere

Above the photosphere lies the transition region, which, as the name suggests, is a thin atmospheric layer in which the photosphere abruptly transitions into the chromosphere. Within this region, which has a thickness of roughly 1000 km, the temperature begins to sharply rise from about 4000 Kelvin up to the characteristic temperature of the chromo-

sphere, which is on the order of 25,000 Kelvin. Because of the sharp increase in temperature and drop in plasma density, the chromospheric plasma is not strongly absorptive, and instead spectral *emission* lines are observed in the chromosphere.

Chromospheric diagnostics in Hα and the Ca H and K lines are important for tracing the vertical structure of the magnetic fields that are line-tied (footpoint-anchored) to the photosphere. The Hα ribbons, swirls, and brightenings all indicate the presence of strong, tangled magnetic fields that shape the emission structures of the predominantly hydrogen chromosphere. These brightenings are highly dynamic and evolve on rapid timescales of a few minutes. Figure 1.11 displays full-disc views of the sun, current on the day of this writing, in both integrated optical light (white-light) and Hα light, as well as a chromospheric view of the atmosphere above an active region sunspot.

1.1.7 The Corona

The corona is perhaps the most mysterious region of the sun, despite the fact that it is easily observable. The hot, tenuous corona lies above the chromosphere, and rises to temperatures greater than 10 million Kelvin, as inferred by the widths of emission lines and the presence of highly-ionized atoms. How such a hot plasma with weak magnetic fields (of order 10 Gauss) is produced and sustained above the photosphere has been a mystery for many decades, and is usually called the *coronal heating problem.* A good candidate for explaining the coronal heating are Alfvén waves propagating upwards from the subphotosphere that deposit their energy in the corona to heat it (Moriyasu et al. (2004)), but others have argued that the observed rate of energy deposition is not high enough to account for such high temperatures in the corona, and instead suggest rapid, "nano-scale" magnetic reconnection in the coronal plasma.

The corona is the source of the solar wind. Although TRACE and Yohkoh have observed closed-loop and cusp structures in the corona which are line-tied (anchored) to the photosphere (see Figures 1.12 and 1.13) by imaging in x-ray wavelengths, the dark regions correspond to regions of open magnetic field lines that extend out into interplanetary space,

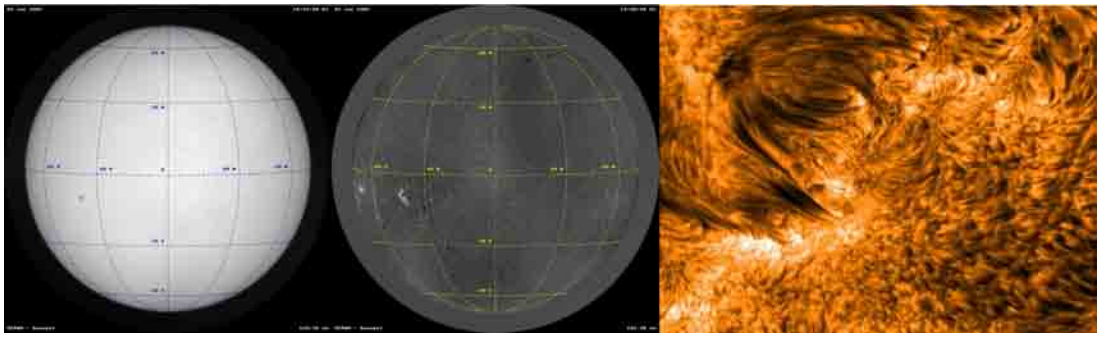

FIG. 1.11 Chromospheric appearance in white-light and Hα. *Left:* The solar disc observed in white light on 28 Jun 2007, showing the photosphere. Note the small, symmetric sunspot at $40°E$, $10°S$. Courtesy of OSPAN (NSO). *Center:* The solar disc, seven minutes later, observed in Hα light. The sunspot is still visible, as is the surrounding magnetic structure, which could not be seen in white light. These include the Hα filaments seen surrounding the spot, as well as the bright arcade structure which has a footpoint just to the east of the spot, and at the same latitude. Also visible in Hα are the plasma filament structures in the northern hemisphere at 50° latitude and in the southern hemisphere at 30° latitude. Courtesy of OSPAN (NSO). *Right:* The textured appearance of the chromosphere when viewed in Hα is due to the presence of *spicules*, which are produced when the constant stream of acoustic waves from the solar interior leak into the less-dense lower chromosphere, giving rise to shock waves that travel relatively small distances while heating up the chromospheric plasma. The superposition of the acoustic waves leaked into the lower chromosphere produces vertically-expanding shock-fronts that achieve maximum heating and compression along their areas of contact, leading to thin, wiry, mostly vertical emission structures like those seen in the right-lower side. Credit: SST, Royal Swedish Academy of Sciences, LMSAL.

turning radial after a few solar radii. The open field lines serve as guides for accelerating solar particles away from the sun at several hundred kilometers per second, via a mechanism whereby the low-pressure footpoint is taken to be at infinity, and the high temperatures of the coronal plasma lead to large velocities which exceed the escape velocity of the Sun.

Coronal magnetometry techniques for measuring the magnetic fields in the corona are still in their infancy. The Zeeman approximation is most often used in this young discipline, whereby the center-separation of the two lobes of the Stokes V circular polarization signal is proportional to the magnetic field strength. However, this approximation is sensitive to the Stokes V peak separation in a non-linear fashion when the field strength is low (Socas-Navarro (2005)). This consistently underestimates the field strengths in the weak-

FIG. 1.12 Magnetic arcades observed by TRACE. The figure shows an image of the solar corona near the limb, taken by the Transition Region and Coronal Explorer (TRACE) satellite. Note the loop arcades, which are magnetic field lines (thin flux tubes) that contain hot plasma constrained to move along the field lines, thereby highlighting them against the background of empty space.

field regime. However, the Zeeman effect is sensitive to magnetic fields only above some critical value (typically $\sim$ 100 G or so), and this is much larger than the presumed strength of the coronal fields ($\sim$ 10 G), and so in practice it cannot be used as a diagnostic tool for coronal magnetism. Both the Hanle effect and the Paschen-Back effect are sensitive to magnetic fields between 1 milligauss and 1000 Gauss (roughly) and so might serve as better magnetic field diagnostics in the corona. Nevertheless, an appropriate model for radiative transfer through the corona is required for interpretation of the Hanle and Paschen-Back effects. This is somewhat complicated, since at temperatures in excess of 10 million Kelvin and in the tenuous coronal plasma, collisions are infrequent, and radiative effects nullify local thermodynamic equilibrium conditions.

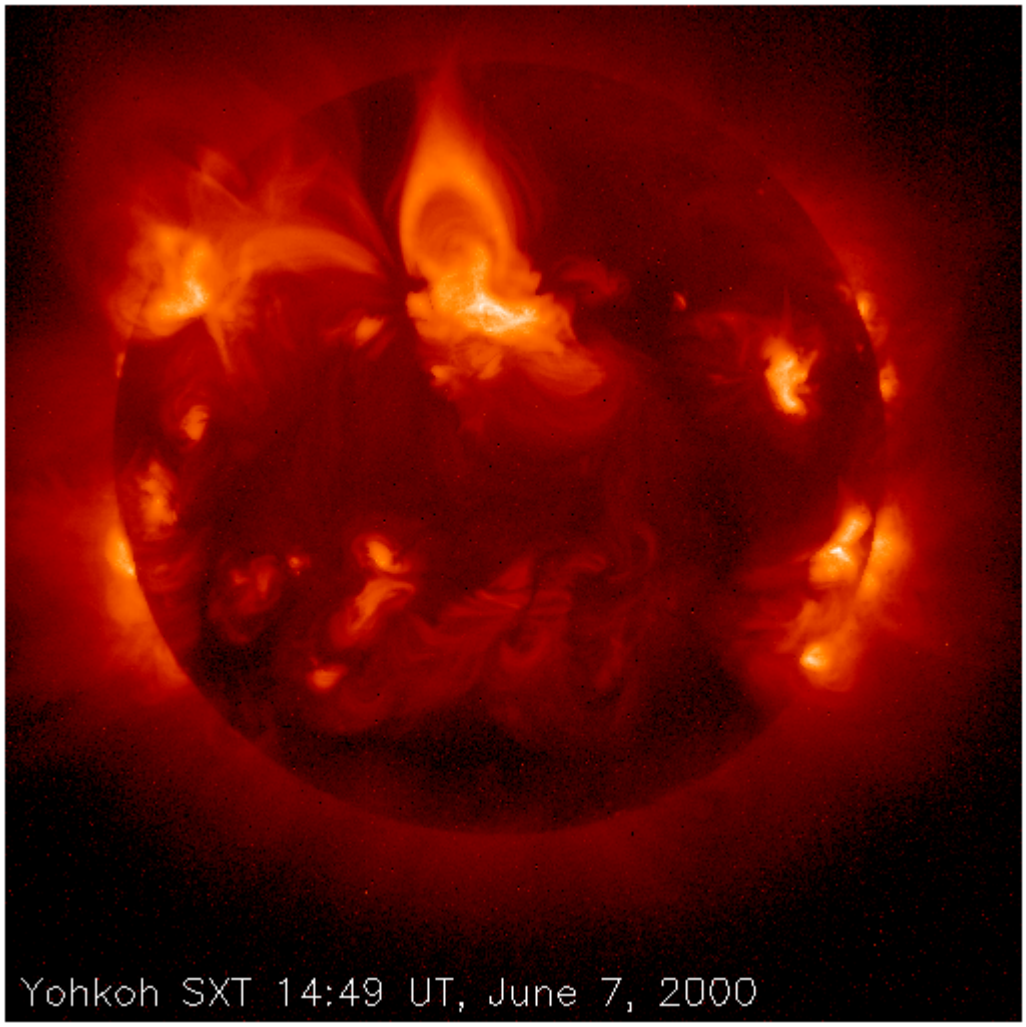

FIG. 1.13 The solar corona, as observed by the Yohkoh satellite. The figure shows an excellent full-disc image taken by the Soft X-Ray Telescope (SXT) onboard the Yohkoh satellite. Note the perfect cusp structure just above center. This shows very plainly the X-type transition region between the closed coronal field lines, as seen in Figure 1.12, and the open coronal field lines that extend outwards into interplanetary space. It should be noted that all the bright structures in these x-ray images are regions of closed field line arcades. Darker areas are historically called *coronal holes*, although they are anything but empty. These coronal holes are regions of open magnetic field which do not return to the sun like the arcades, but sweep out into space, and are responsible for guiding the solar wind outward, away from the sun.

1.1.8 The Heliosphere

The heliosphere is the sun's "range of influence". Since the sun is continuously emitting streams of matter and radiation in the form of the solar wind, the heliosphere is defined by the interaction of the solar wind with the flows present in interstellar space. That is, the *solar* wind travels only so far before encountering the *stellar* winds permeating space. The interaction of the solar wind with interstellar flows produces a bow shock at the heliopause, where the kinetic energy of the solar wind goes to zero. The geometry of the solar system and heliosphere is shown in Figure 1.14, and the position of the Voyager 1 probe (as of 2005) is marked.

1.2 Solar Magnetism

This work aims at inferring the magnetic field strength and geometry at the photospheric level, and it would be naive to assume that the only manifestation of solar magnetism are sunspots. In this section and the subsequent subsections, I present a brief overview of the various magnetic structures to be found at the photospheric level.

1.2.1 Small-Scale Magnetic Elements

The tops of convective cells can readily be seen in white light at the photospheric level. Due to the high plasma β (ratio of gas to magnetic pressure) in the quiet sun, any present magnetic fields will be swept along and carried with the plasma motions. Because of the convective overturning, this tends to concentrate quiet-sun magnetic flux in the intergranular network, the dark lanes that appear between the bright convective cell-tops, as can be seen in Figure 1.15. These intergranular magnetic fields typically have strengths on the order of 100 G, although Spruit (1979) put forth the idea of "convective collapse" to explain the observation of intergranular fields which seemed to have field strengths on the order of 1000–1500 G, about the same strength one would expect in the penumbral regions of a mature sunspot.

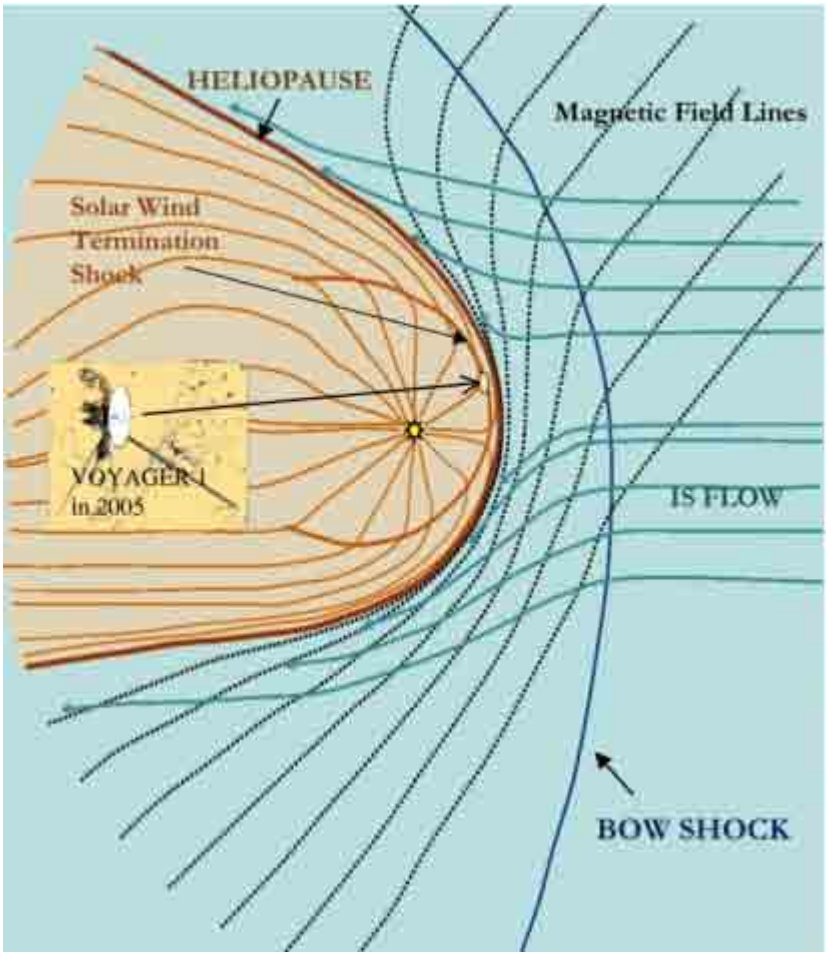

FIG. 1.14 A schematic of the heliosphere. The figure displays a zoomed-out image of our solar system, showing the boundaries of the sun's region of influence (the heliosphere), as well as the current positions of the Voyager 1 spacecraft, which is just now venturing into what can be properly called the boundaries of interstellar space. The sun is moving to the right in this image, into the interstellar winds. From Pogorelov and Zank (2004).

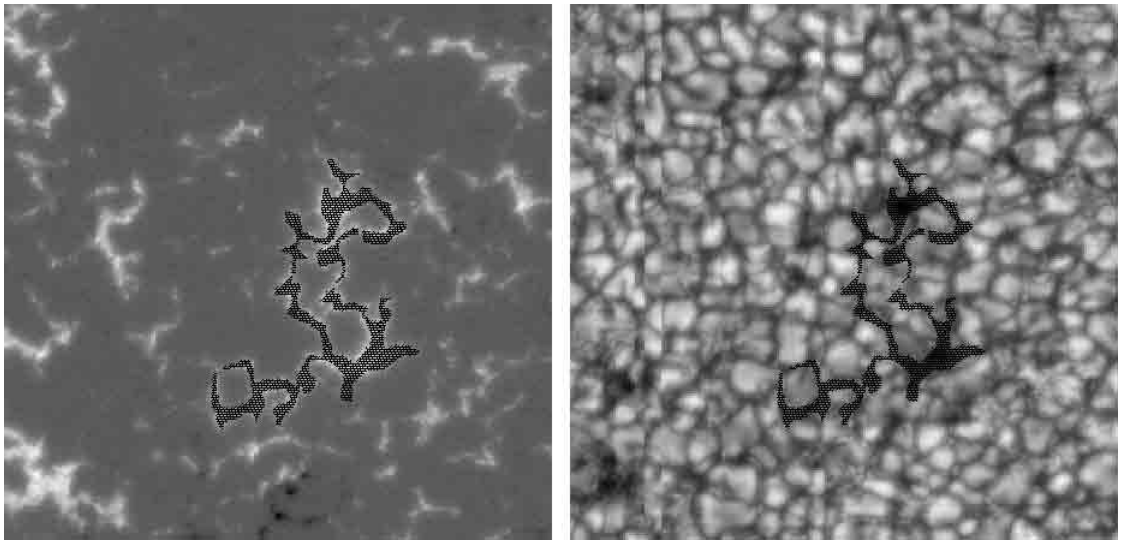

FIG. 1.15 Small-scale magnetic elements in the solar granulation. *(Left Frame):* Integrated circular polarization over a patch of quiet-sun. The hashed region covers an area containing significant circular polarization signal indicative of the presence of magnetic fields not corresponding to a sunspot. *(Right Frame):* The same region showing quiet-sun convection cells. Note that the regions containing a net circular polarization signal correspond to the dark intergranular lanes where flux has been swept into by the convective overturning.

1.2.2 Moving Magnetic Features

Moving magnetic features (MMFs) are readily observed in time-series observations of active region sunspots. Surrounding a sunspot is a radial flow directed away from the main spot, the so-called *moat flow*, which seems to carry small, coherent magnetic structures away from the sunspot. This has been interpreted as field lines or flux tubes that have detached from the main "trunk" of sunspot magnetic fields. The intersections of these flux tubes with the photospheric surface appear as small-scale magnetic features that are then subject to the flow conditions in and around the penumbra, leading to the appearance of MMFs. Furthermore, it has been suggested (Kubo et al. (2007a,b)) that these MMFs could be important for sunspot decay. Notably, Ravindra (2006) has observed MMFs which originate within the penumbra of a sunspot and proceed to move outside the boundary of the spot, as in Figure 1.16. Although the interpretation of free-floating magnetic flux tubes that have detached from the main sunspot lends itself to the idea that MMFs have the same

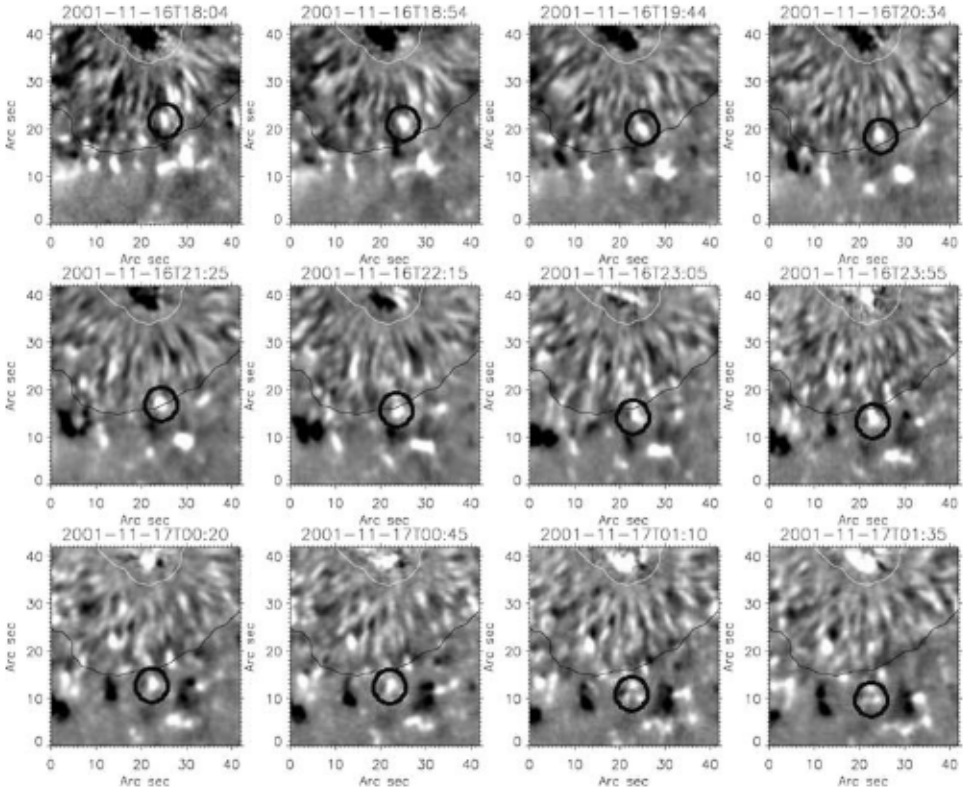

FIG. 1.16 Moving magnetic features (MMF). Over a period of roughly 7 hours, a small MMF (in black circle) was seen to detach from penumbral fields and be carried outside the boundary of the sunspot. From Ravindra (2006).

polarity as their host sunspot, MMFs with opposite polarity, as well as bipolar MMFs, have been identified.

1.2.3 Pores

Pores are typically considered to be miniature sunspots, with some very important differences. Most notable is their size; pores tend to have roughly the same diameter as that of a convective cell ($\sim 1000km$) but can be larger. They most likely are the result of accumulation of magnetic flux into organized flux tubes by the sweeping motions of their surrounding convective cells. Despite their small stature, they can display field strengths on the order of 2000 G (Leka and Skumanich (1998)). Another notable difference is the lack of a penumbra; pores display umbral structures only, as can be seen in Figure 1.17. Finally,

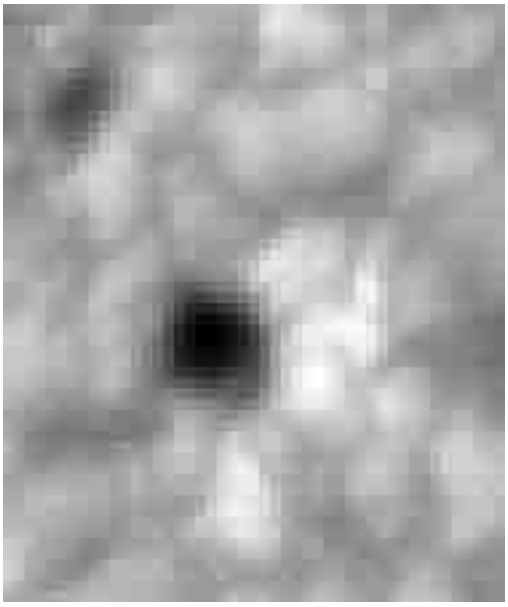

FIG. 1.17 A solar pore. The figure shows a small region of accumulated flux that has become strong enough to inhibit the underlying convection, thus appearing dark. Note that the linear size is not much different than the surrounding convective cell-tops.

as opposed to the radially outward moat flow around sunspots, Roudier et al. (2002) and Lagg et al. (2006) have found inward and downward flows in the regions surrounding pores, perhaps from radiative cooling of the surrounding material into the relatively evacuated flux tube.

1.2.4 Plage

Diffuse magnetic flux is constantly emerging into the solar atmosphere, even in regions where no sunspot or pore activity is present. This emerging flux has no signature in intensity, because it is not yet strong enough to significantly inhibit the convection from bringing hot material up from the subsurface. Despite this fact, it can be detected in the circular and linear polarization signals it presents. The net circular polarization (NCP):

$$NCP = \int |V(\lambda)| d\lambda \tag{1.14}$$

where $V(\lambda)$ is the Stokes V net circular polarization profile as a function of wavelength, is integrated around a magnetically-sensitive spectral line will suddenly show signs of this

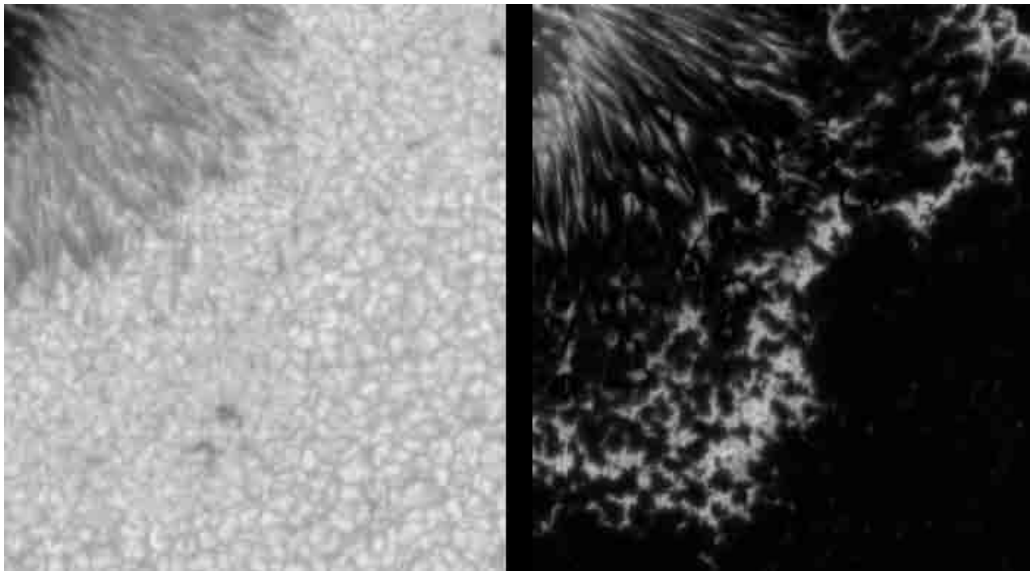

FIG. 1.18 Plage magnetic fields. *(Left Frame):* Continuum (white-light) image of a region surrounding a sunspot. *(Right Frame):* Net circular polarization, integrated around the Fe I 6301.5Å absorption line. This measure is related to the longitudinal magnetic field strength, clearly showing the presence of magnetic fields where no corresponding reduction in continuum intensity is observed.

diffuse *plage* field. Figure 1.18 shows a small subregion around a sunspot observed with the Hinode satellite, including part of the penumbra, and the corresponding integrated circular polarization measure.

1.2.5 Active Regions and Sunspots

Perhaps the most interesting structures accessible to human observation are sunspots. These aggregated regions of concentrated magnetic field are responsible for producing the most energetic events in the solar system: solar flares and coronal mass ejections. The complex magnetic topology and evolution of sunspots and sunspot groups permits the rapid conversion of the energy present in the magnetic field to the kinetic and thermal energy of the plasma, via the poorly-understood process of *magnetic reconnection*. Although the reconnection and subsequent flaring occur in the chromosphere and corona, the footpoints of this activity are the photospheric-level magnetic fields of sunspots. Recent helioseismological data has hinted at the sub-surface structure of sunspots, for instance in Figure 1.19 showing

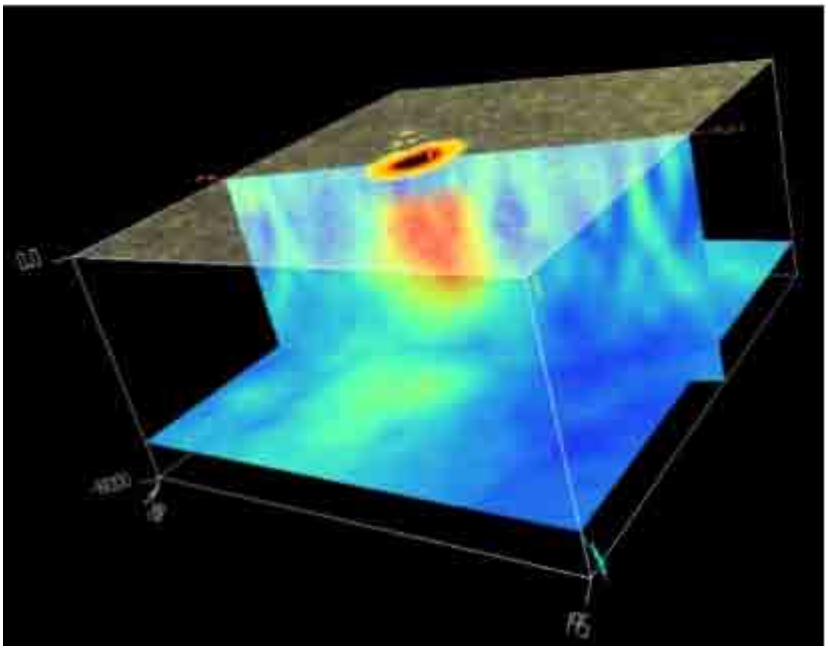

FIG. 1.19 Sub-surface structure of a sunspot. The figure is derived from time-distance helioseismological techniques by the SOHO Michelson Doppler Imager (MDI). Sound speed is plotted in two orthogonal planes, showing the dependence on position within the sunspot. Blue represents slower sound speed, with red indicating a high sound speed. The sound speed variations translate into density and pressure variations withing the sub-sunspot plasma.

the sound speed below a sunspot in planes perpendicular and parallel to the flux tube axis. The magnetic structure in the same region generally appears as in Figures 1.20 and 1.21; that is, below the surface, the flux tube looks like a tree, with the umbra representing the main trunk, and the penumbra being formed by the branching magnetic fields. The main features of these structures (penumbra, umbra) are presented in the following subsections.

The Umbra

A sunspot umbra is a region of intense magnetic field strength, where the magnetic field lines pierce the surface more or less vertically. Because of the strong magnetic pressure inside the sunspot, convection is inhibited in the umbra. Therefore, hot plasma from subsurface regions cannot be drawn to the surface by convection. Heating by conduction is not a major contribution to the total heat transfer in stellar atmospheres, and as a consequence

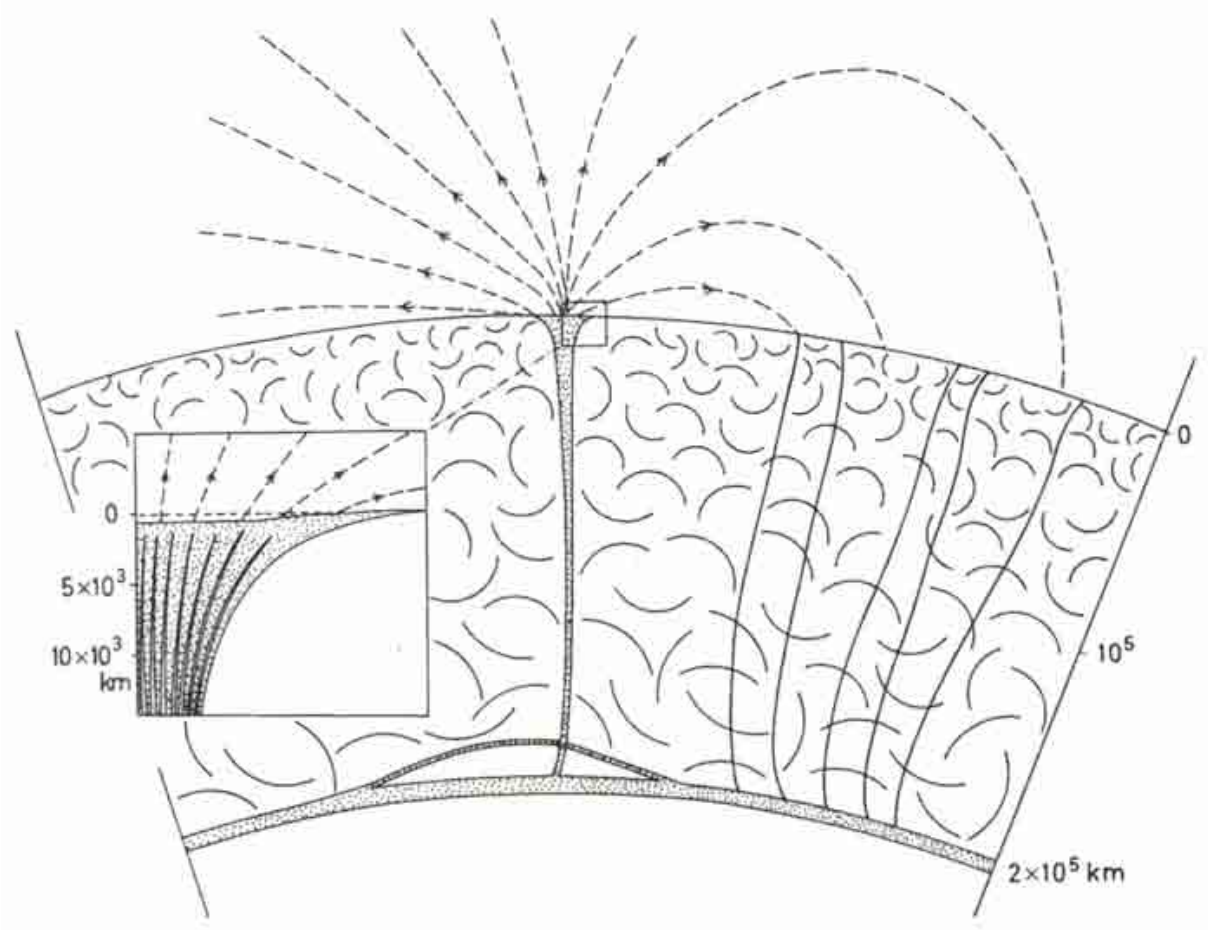

FIG. 1.20 A cartoon schematic of the vertical structure of a sunspot. The figure shows an edge-on schematic view of the surface and sub-surface structure of a typical sunspot, highlighting the vertical geometry of their magnetic fields. From unknown source.

the umbra appears dark against the surrounding photospheric granulation. A characteristic temperature range in a sunspot umbra is of the order of 3000–5000 Kelvin. Because of this low temperature (compared to the quiet-sun), the convective gas forces outside the sunspot (temperature of 6000–6500 Kelvin) compete with the magnetic forces in the sunspot. In the absence of any large-scale magnetic field structures, this is exactly what drives the quiet solar granulation to keep small-scale (e.g., plage) fields relatively confined to the dark intergranular lanes. However, because of the nearly vertical, strong fields (usually of the order of 2000–4000 Gauss) in the umbra, any compression of the sunspot structure by being

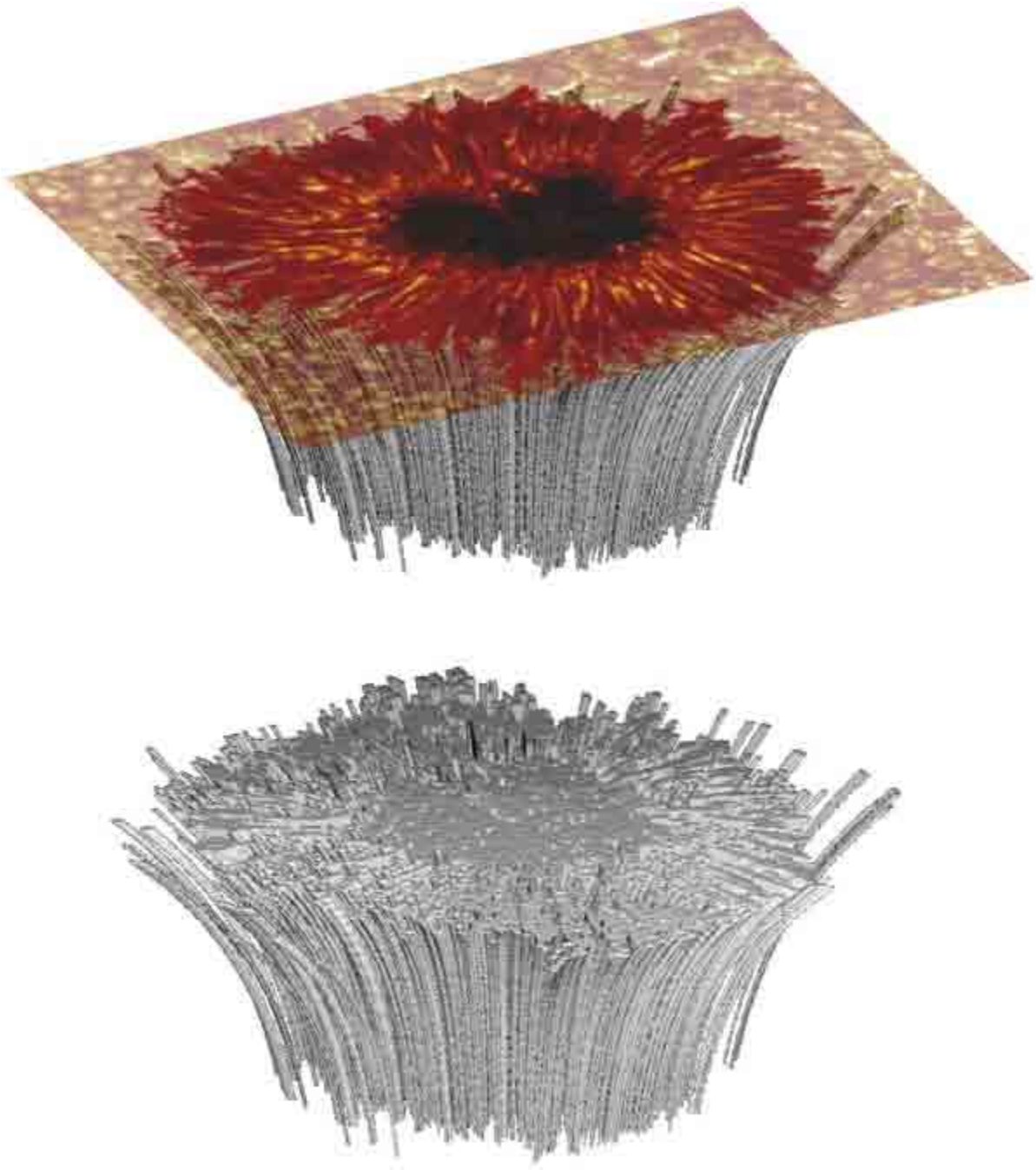

FIG. 1.21 Flux-tube structure of a sunspot. The figure shows the results of a simulation about how sunspots are formed from self-bound collections of individual thin flux tubes. The thin, inclined flux tubes branching out from the main trunk pierce the surface to form the penumbra. From unknown source.

buffeted by external convective cells increases the outward magnetic pressure of the umbral fields, stabilizing the sunspot against the convective forces. However, the umbra *is* situated slightly lower in the photosphere than the penumbra and quiet sun, in what has historically been called the *Wilson depression*. Closer to the umbral boundary with the penumbra, umbral magnetic fields tend to transition from nearly vertical to an $\sim 20°$–$30°$ inclination from vertical (Bellot Rubio et al. (2004)).

The umbrae of most sunspots appear uniformly dark, giving one the false impression that the field strength is relatively uniform over the umbra. As can be seen in Figure 1.22, this is not the case, and in fact is just the result of the contrast between the dark umbra and the bright quiet regions. The umbra of a sunspot contains significant substructure, most notably the *umbral dots*, regions of lower field strength, which therefore appear brighter than the rest of the umbral surroundings. The lower field strengths in these umbral dots does not inhibit convection as strongly as in the surroundings, so hotter material is able to upwell between the flux bundles that make up the umbra. An example of these umbral dots is displayed is Figure 1.23.

The Penumbra

In contrast to the umbra, the magnetic fields in a sunspot penumbra are highly-inclined to the local solar normal. However, due to the "uncombed" structure of the penumbra (Solanki and Montavon (1993)), the average penumbral magnetic field is not horizontal. The alternating light and dark filamentary structure is clearly evident, and highly reminiscent of flux tube structures. The individual filaments tend to be brighter on the side nearer the umbra, then darken the further one goes from the umbra, and evolve on timescales of tens of minutes. This is related to the so-called *penumbral grains*, which appear as the bright "heads" of the filaments near the umbra/penumbra boundary (Figure 1.24). They are the signatures of hot upflows that eventually turn horizontal, giving rise to the Evershed effect (c.f. next subsection), and observations as well as simulations have shown that they migrate inward toward the umbra (Schlichenmaier et al. (1998)) after a steady outflow

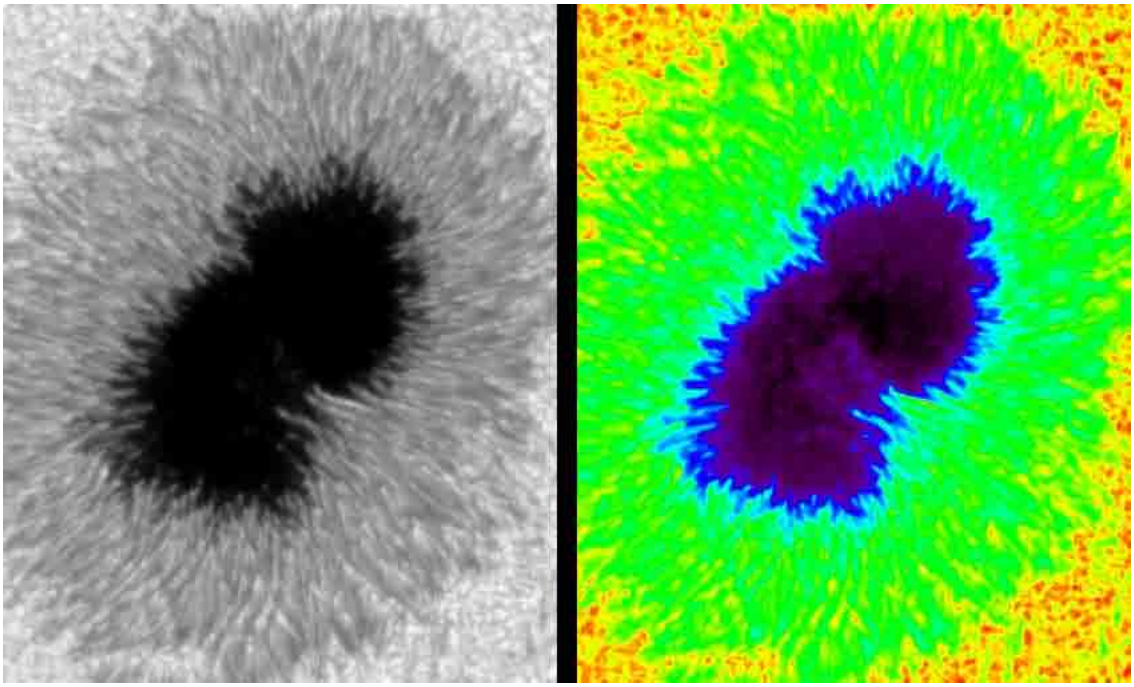

FIG. 1.22 Contrast-enhanced umbral structure. *(Left Frame):* Grayscale continuum intensity. *(Right Frame):* A slightly different color scheme to accentuate the substructure of the umbra that is washed out in the grayscale continuum image.

state has been achieved. The field strengths in the penumbra are typically lower than in the umbra, and in fact, take values $\sim$ 2000 Gauss near the umbra/penumbra boundary. This typically decreases radially to $\sim$ 500 Gauss near the penumbra/quiet-sun boundary. Characteristic temperatures in the penumbra are only slightly lower than the quiet-sun (5500–6000 Kelvin), causing it to appear much brighter than the umbra, but still dimmer than the photospheric granulation in the quiet-sun.

As mentioned previously, helioseismological analysis has revealed the existence of two zonal flows in the subsurface structure of a sunspot that are consistent with current theories on how sunspots are formed. The presence of circulating downflows that weaken closer to the photospheric surface allow thin flux tubes to be bound together, rooted deeper in the surface, while the weaker return flow allows the spreading of flux tubes to create the penumbra. The inclined penumbral field lines do not spread out as the umbral fields do. Considering the penumbral field lines as thin flux tubes embedded in a large sunspot flux tube, the plasma

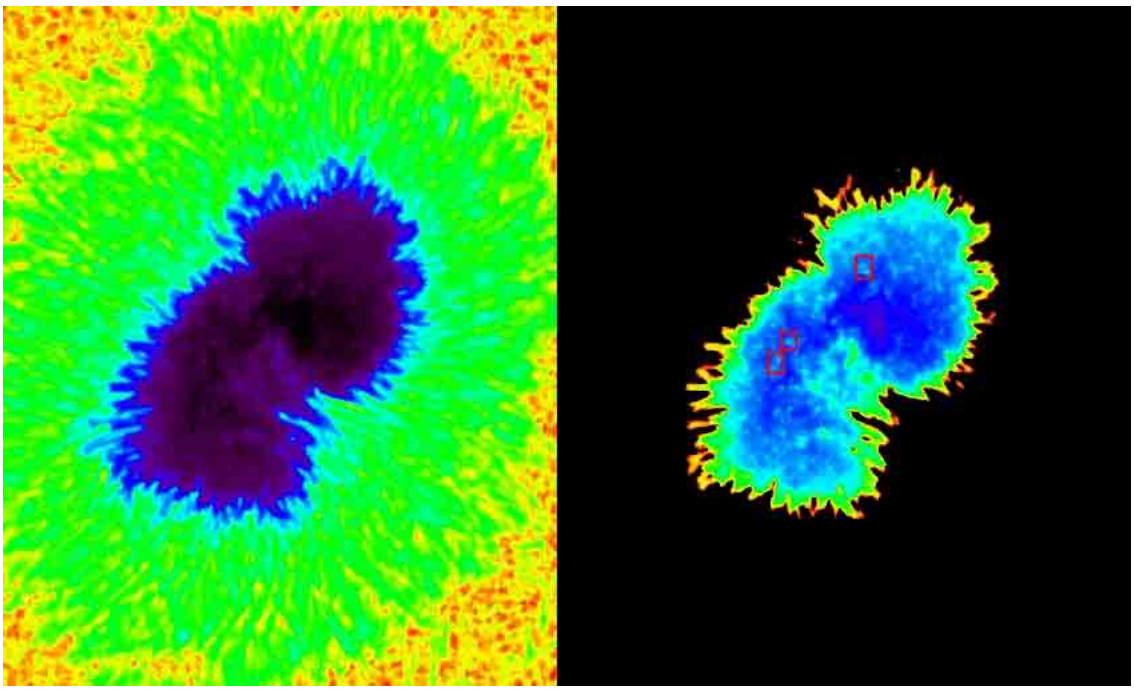

FIG. 1.23 Umbral dots. *(Left Frame):* Colorized continuum intensity. *(Right Frame):* Isolating the umbra even more by masking out the other areas of the image reveals the previously obscured umbral dots (outlined in the red boxes), where field strengths are lower and temperatures are higher than the surrounding umbral areas.

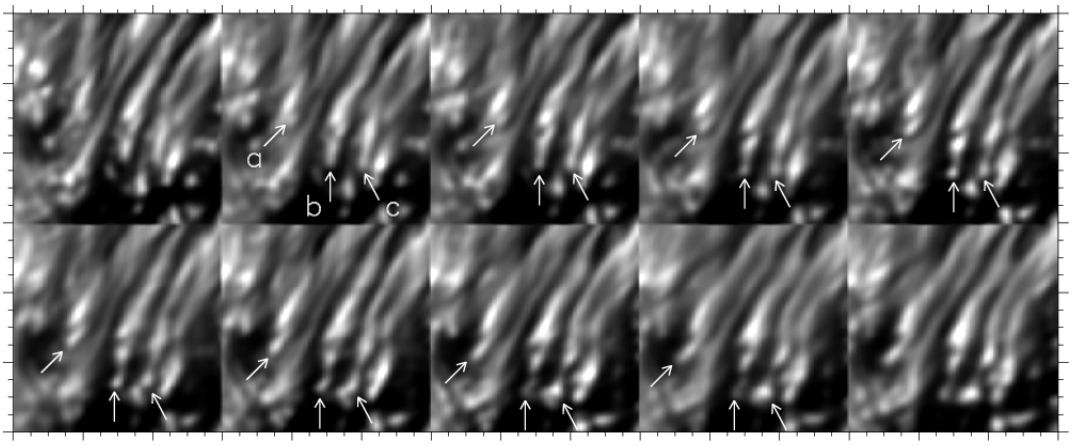

FIG. 1.24 Migration of penumbral grains. The penumbral grains are seen as the bright points on the umbra-side of penumbral filaments. The arrows point to three penumbral grains that consistently migrate toward the umbra, in agreement with simulations of penumbral flux tubes by Schlichenmaier. From Hirzberger et al. (2005).

in the thin flux tube is radiantly heated until the element becomes buoyantly unstable. The flux tube rises, guiding the interior plasma, to the surface, where it suddenly encounters the sub-adiabatically stratified plasma of the photosphere. Unable to continue with the upwards convection of the plasma in the thin flux tube, the plasma instead spreads out radially away from the umbra, while simultaneously dragging the flux tube with it (frozen-in field). This further increases the inclination of the penumbral magnetic fields, creating more horizontal fields.

The Evershed Effect

The buoyant instability of the embedded thin flux tube in the previous subsection was responsible for creating the magnetic structure of the highly inclined fields in the penumbra. Recall that hot plasma from deep beneath the surface was guided up the surface by the embedded thin flux tube. As the flow pierces the surface, the plasma elements in the tube begin to cool off, radiating away their excess heat as it comes into thermal equilibrium with the ambient medium. This cools the plasma element, which decreases the temperature and increases the density of the plasma inside the tube. Therefore, the tube not only sinks back down toward the surface (effectively creating the penumbra/quiet-sun boundary), but the decrease in gas pressure inside the tube sets up a pressure gradient from the inner footpoint outward, which drives a diverging siphon flow in the visible parts of the penumbra. This is called the *Evershed effect*, after its namesake who first hypothesized in 1909 that certain observed wavelength shifts could be interpreted as material flows perpendicular to the line-of-sight, and it is commonly observed in photospheric images of large sunspots. An extensive review of penumbral fine structure can be found in Schlichenmaier (2003).

The so-called penumbral grains therefore have the natural interpretation of being the bright footpoints of the hot upflow channels created by the penumbral flux tubes. These are easily seen near the umbra/penumbra boundary in Figure 1.24, which shows some interesting structure, in that the grains are brighter on the upper and left sides of the boundary, which is a line-of-sight effect. It could be that the subsurface downflows have

captured a fragmented branch of the main trunk of the flux tube, and because it is deeper, the upwelling plasma in the Evershed flow is coming from a hotter atmospheric layer, and hence appears brighter. Furthermore, penumbral grains are sometimes observed in the outer penumbra moving radially outward. These may be interpreted as the arches of newly-emerging penumbral flux tubes that merge into the structures already present, or they may be manifestations of the moving magnetic features presented in an earlier subsection.

1.3 Motivation: Importance of Magnetic Field Diagnostics

The most violent, eruptive phenomena in the solar system occur when local parts of the solar magnetic field become incredibly intertwined and tangled. Field lines of opposite polarity that are brought close to one another in a plasma create current sheets whose high conductivity allows the magnetic fields to break and reconnect, forming a new, relaxed field topology in a lower energy state. This sudden conversion of the magnetic energy in the tangled fields into the thermal and kinetic energy of the local solar plasma is termed a *magnetic reconnection event*, and releases electromagnetic radiation that spans the entire spectrum as well as accelerates solar particles out into interplanetary space. The release of broadband electromagnetic radiation is typically referred to as a *flare*, while the corresponding acceleration of massive particles, e.g., protons, is called a coronal mass ejection (CME). The two are frequently grouped under the umbrella term "solar flare", but the distinction is quite important.

Solar flares and CMEs present unique dangers, not only to man-made constructs like satellite constellations, but to all space-faring humans. The unshielded radiation dose received from a large flare or CME could be career-ending for an astronaut, if not fatal. The need for rapid, accurate, and reliable techniques for forecasting space weather effects produced and influenced by the sun will become ever more important as our civilization increases our foothold in the near-Earth environment.

The potential for flare prediction via observation of an as-yet unknown characteristic precursor is highly debated in the solar physics community. Being able to simultaneously

watch for brightenings *and* complex magnetic fields should bolster current flare prediction methods, which usually only rely on brightening, as well as *a posteriori* analysis of sunspot magnetic fields, extrapolating backwards and looking for some sign of imminent flare onset. Models for definite signatures that a flare is imminent have yet to be developed, and current forecasting techniques rely mostly on the past experience and intuition of the forecaster.

However, a member of the solar physics community, Ake Nördlund, once said,"no problem can withstand the assault of sustained thought." I believe that unique flare precursors *do* exist, but due to our limited spatial, spectral, and temporal resolutions, the current instrumentation is just not intelligent enough to resolve and observe them right now. However, with the right tools and the next generation of advanced solar observational equipment, we will extract even more information about sunspot magnetic fields and their behavior on ever-smaller scales. Flare initiation is presumed to occur at the upper-photospheric and/or lower chromospheric level, due to the complex tangling motions from convection and other photospheric flows. The mapping of sunspot magnetic fields at the photospheric level is important for analysis of any flare event, since the sunspots are the footpoints of the reconnecting field lines, and it is generally accepted that the plasma motions in and around sunspots are ultimately responsible for creating a flaring configuration. They therefore provide important observational information, but can also be used as realistic boundary conditions in simulations of coronal magnetic fields and flaring magnetic reconnection.

This dissertation describes a novel genetic algorithm method for recovering the photospheric magnetic structure of sunspots from the observed intensity and polarization profiles. The method is shown to be very robust over the active region structures, and yields uncertainties in the model parameters that are quite acceptable ($\sim$ few percent in all cases). The inferred magnetic fields and other parameters are uniquely determined using a variety of observational constraints, and show structure that is indicative of a typical symmetric sunspot, reinforcing the belief that the method works. Moreover, the method has the potential to vastly reduce the computation time for producing such maps, a feature that I believe will be absolutely essential for the interpretation of large spectropolarimetric datasets from

planned next-generation telescopes, specifically the Advanced Technology Solar Telescope (ATST).

The format of this dissertation is as follows. Chapter 2 addresses the issues of polarization and radiative transfer through the magnetized solar atmosphere, specifically how they allow the determination of magnetic field characteristics and some of the thermodynamic parameters of the solar plasma. Chapter 3 outlines the data and observations used in this study, and the instrumentation used to obtain them. Chapter 4 presents a tutorial (of sorts) on genetic algorithms and their operation, presents some of the improvements I have made to the basic genetic algorithm, and discusses the implementation in the inversion procedure along with several test cases. Chapter 5 displays the preliminary results of this method when applied to a variety of distinct sunspots and active regions, and presents several MHD-related quantities that are derived from the basic inversion results. Chapter 6 addresses the parallelization of the genetic inversion method, presents results on timing and scalability to a cluster computing environment, and Chapter 7 summarizes the conclusions drawn from this work and outlines the direction of future work in this area.

CHAPTER 2
POLARIZED RADIATIVE TRANSFER IN A MAGNETIZED ATMOSPHERE

2.1 Introduction

The propagation and attenuation of the light emitted from a distant source, in this case the integrated emission from a small patch of solar photospheric surface, needs to be described in terms that easily lend themselves to extending the analysis to cover the presence of magnetic fields. This chapter presents the mathematical foundations used in this work to describe light polarization and its propagation/attenuation, as well as the mathematical model describing the observable polarization signals at the photospheric surface. Unless otherwise indicated, all wavelengths are given in Angstroms ($1\text{Å} = 10^{-10}\text{m}$).

2.2 The Stokes Description of Polarized Light

2.2.1 Maxwell's Equations

Maxwell's equations describe the behavior of electric, $\vec{E}$, and magnetic, $\vec{B}$, fields in vacuum as well as in matter (dielectric). Explicitly, the set of coupled equations is given by (following *Classical Electrodynamics* by Jackson (1998)):

$$\vec{\nabla} \cdot \vec{E} = \frac{\rho_c}{\varepsilon_0} \tag{2.1}$$

$$\vec{\nabla} \cdot \vec{B} = 0 \tag{2.2}$$

$$\vec{\nabla} \times \vec{E} = -\frac{\partial \vec{B}}{\partial t} \tag{2.3}$$

$$\vec{\nabla} \times \vec{B} = \mu \vec{J} + \mu\varepsilon \frac{\partial \vec{E}}{\partial t}, \tag{2.4}$$

where ρ_c and $\vec{J}$ are source terms representing the charge density and current density, respectively. The quantities μ and ε are the permeability and permittivity, respectively, of the dielectric medium through which the waves are traveling. If we assume the medium is a stationary plasma, the requirements of global charge neutrality and no free currents allows one to set $\rho_c = 0$ and $\vec{J} = \vec{0}$. With the *a priori* knowledge that light is a self-sustaining

electromagnetic wave, we look for plane wave solutions to the coupled set of equations. Assuming a harmonic time-dependence of the form $e^{-i\omega t}$ for solutions to Maxwell's equations, the following relations are obtained:

$$\vec{\nabla} \cdot \vec{E} = 0 \tag{2.5}$$

$$\vec{\nabla} \cdot \vec{B} = 0 \tag{2.6}$$

$$\vec{\nabla} \times \vec{E} = -\frac{\partial \vec{B}}{\partial t} \longrightarrow i\omega \vec{B} \tag{2.7}$$

$$\vec{\nabla} \times \vec{B} = \mu\varepsilon \frac{\partial \vec{E}}{\partial t} \longrightarrow -i\omega\mu\varepsilon \vec{E}. \tag{2.8}$$

Applying the curl operator to equations (2.7) and (2.8) yields:

$$\vec{\nabla} \times \vec{\nabla} \times \vec{E} = i\omega \vec{\nabla} \times \vec{B} \tag{2.9}$$

$$\vec{\nabla} \times \vec{\nabla} \times \vec{B} = -i\omega\mu\varepsilon \vec{\nabla} \times \vec{E}. \tag{2.10}$$

Using the well-known identity:

$$\vec{\nabla} \times \vec{\nabla} \times \vec{A} = \nabla^2 \vec{A} - \vec{\nabla} \left(\vec{\nabla} \cdot \vec{A} \right), \tag{2.11}$$

the following equations are obtained as:

$$\nabla^2 \vec{E} + \mu\varepsilon\omega^2 \vec{E} = 0 \tag{2.12}$$

$$\nabla^2 \vec{B} + \mu\varepsilon\omega^2 \vec{B} = 0. \tag{2.13}$$

These are harmonic equations for the spatial variation of the electric and magnetic fields, so plane waves of the form $e^{i\vec{k}\cdot\vec{x} - i\omega t}$ may be substituted, ultimately yielding the condition that:

$$\vec{k} \cdot \vec{k} = \mu\varepsilon\omega^2, \tag{2.14}$$

where $\vec{k}$ is the wavenumber vector, and is oriented along the propagation direction of the wave ($\vec{k} = k\hat{n}$). From this it can easily be seen that the phase velocity of the light wave is:

$$v_{ph} = \frac{\omega}{k} = \frac{1}{\sqrt{\mu\varepsilon}}. \tag{2.15}$$

If the equations describing the electric and magnetic fields are given by

$$\vec{E} = \vec{E}_0 e^{i\vec{k}\cdot\vec{x} - i\omega t} \tag{2.16}$$

$$\vec{B} = \vec{B}_0 e^{i\vec{k}\cdot\vec{x} - i\omega t} \tag{2.17}$$

then the zero-divergence conditions requires that

$$\vec{\nabla} \cdot \vec{E}_0 = 0 \tag{2.18}$$

$$\vec{\nabla} \cdot \vec{B}_0 = 0. \tag{2.19}$$

Since, under the plane-wave assumption, the operator $\vec{\nabla}$ may be replaced with $i\vec{k}$, we obtain

$$\hat{n} \cdot \vec{E}_0 = 0 \tag{2.20}$$

$$\hat{n} \cdot \vec{B}_0 = 0, \tag{2.21}$$

which describes a transverse wave whose components ($\vec{E}$ and $\vec{B}$) are always orthogonal to the propagation direction, $\hat{n}$. Substituting the plane-wave form into the curl equations further yields the relationship between the field amplitudes, $\vec{E}_0$ and $\vec{B}_0$:

$$\vec{B}_0 = \sqrt{\mu\varepsilon}\frac{\vec{k} \times \vec{E}_0}{k}. \tag{2.22}$$

Defining the right-handed coordinate system ($\hat{\epsilon}_1,\ \hat{\epsilon}_2,\ \hat{n}$) such that the electric (magnetic) field vector lies in the direction of $\hat{\epsilon}_1$ ($\hat{\epsilon}_2$), we have now developed a full spatio-temporal description of the propagation of plane light waves along the direction $\hat{n}$.

2.2.2 Polarization and Its Representation

The *polarization* of a beam/ray of light is defined as the direction of its electric field vector, $\vec{E}$. If one rotates the coordinate axis, introduced at the end of the last section, around the $\hat{n}$-direction, by 90°, we arrive at a different description of the polarization. The electric field now points in the $\hat{\epsilon}_2$-direction, while the magnetic field points in the $-\hat{\epsilon}_1$-direction. The direction of propagation is unchanged, as is the wave itself, so is the beam polarized along $\hat{\epsilon}_1$ or $\hat{\epsilon}_2$? One can avoid such geometrical ambiguities by dealing with the

superposition of the two cases. Let the two independent plane-wave electric field vectors be given by

$$\vec{E}_1 = \hat{\epsilon}_1 E_1 e^{i\vec{k}\cdot\vec{x}-i\omega t} \tag{2.23}$$

$$\vec{E}_2 = \hat{\epsilon}_2 E_2 e^{i\vec{k}\cdot\vec{x}-i\omega t}. \tag{2.24}$$

Then the most general plane-wave propagating along $\hat{n}$ can be written as

$$\vec{E}(\vec{x},\ t) = \left[E_1\hat{\epsilon}_1 + E_2\hat{\epsilon}_2\right] e^{i\vec{k}\cdot\vec{x}-i\omega t}, \tag{2.25}$$

where now the amplitudes $E_{1,2}$ may be complex (i.e., have phase differences) and can be written in the form "amplitude $\times$ phase factor":

$$E_j = A_j e^{i\delta_j} \qquad j = 1, 2. \tag{2.26}$$

Remember that the actual electric and magnetic fields are the *real* parts of the quantities we are dealing with here. If E_1 and E_2 have no relative phase difference, then the wave is said to be *linearly polarized* along the direction $tan^{-1}E_2/E_1$. For the case where a phase-difference exists, there exist two possibilities: (1) $E_1 = E_2$ or (2) $E_1 \neq E_2$. A wave of the first type with a phase difference (mod 2π) of 90° between the two components is said to be *circularly polarized*, while an arbitrary wave of the second type with a phase difference $> 0^\circ$ is *elliptically polarized*.

In terms of circular polarization, the components are out of phase by 90° but are of equal magnitude (E_0), so the wave takes the form:

$$\vec{E}(\vec{x},\ t) = E_0\left(\hat{\epsilon}_1 \pm i\hat{\epsilon}_2\right) e^{i\vec{k}\cdot\vec{x}-i\omega t} \tag{2.27}$$

The sign conventions denote the direction of rotation of the electric field vector as seen by an observer looking *into* an oncoming light wave. For the (+) sign, the direction of rotation is counterclockwise, and the wave is called *left-circularly polarized*. Conversely, for the (–) sign, the electric field vector rotates clockwise and is said to be *right-circularly polarized*.

One can imagine that the proper vector superposition of a left and right circularly polarized wave can yield a linearly polarized state, much like superposing two oppositely-traveling waves on a string can lead to a standing, or stationary, wave. Therefore, it seems

that we should be able to describe some arbitrary polarization state by not only using the coordinate system defined above, but also by switching to a new set of basis vectors describing circular polarization. Defining orthonormal basis vectors $\hat{\epsilon}_+$ and $\hat{\epsilon}_-$ as complex linear combinations of the linear basis vectors gives

$$\hat{\epsilon}_\pm = \frac{1}{\sqrt{2}} \left(\hat{\epsilon}_1 \pm i\hat{\epsilon}_2 \right). \tag{2.28}$$

The most general circularly polarized wave can then be written as:

$$\vec{E}(\vec{x}, t) = (E_+\hat{\epsilon}_+ + E_-\hat{\epsilon}_-)e^{i\vec{k}\cdot\vec{x} - i\omega t}, \tag{2.29}$$

where $E_\pm$ are new complex amplitudes with the following properties:

1. $E_\pm$ can be described by an amplitude and a phase. Formally, $E_\pm = A_\pm e^{i\delta_\pm}$.

2. If E_+ and E_- have the same phase ($\delta_+ = \delta_-$), but different amplitudes, the beam is elliptically polarized with its major axis parallel to the x-axis.

3. If E_+ and E_- have a relative phase difference ($\delta_+ \neq \delta_-$), the axes of the polarization ellipse are rotated counterclockwise from the x-axis by an amount equal to half the phase difference.

4. If $E_-/E_+ = \pm 1$ and there is no relative phase difference, a linearly polarized state is recovered. Note that this is exactly the result expected from the standing wave analogy made earlier.

In this dissertation, I choose to work within the circular polarization basis since, as will be discussed in a subsequent chapter, magnetic fields on the surface of the sun are invariably linked with the net circular polarization of the observations.

2.2.3 The Stokes Parameters

To begin, note that the circular polarization basis vectors can easily be shown to satisfy the following relations, from the properties of the imaginary unit, i:

$$\hat{\epsilon}_\pm^* \cdot \hat{\epsilon}_\mp = 0 \tag{2.30}$$

$$\hat{\epsilon}_{\pm}^{*} \cdot \hat{\epsilon}_{\pm} = 1 \tag{2.31}$$

$$\hat{\epsilon}_{\pm}^{*} \cdot \hat{n} = 0 \tag{2.32}$$

By taking the dot-product of each (complex-conjugated) basis vector with the general expression for the plane-wave electric field, one obtains the independent amplitudes of the radiation along each basis vector direction:

$$\hat{\epsilon}_{+}^{*} \cdot \vec{E} = \ amplitude\ of\ left\ circular\ polarized\ component \tag{2.33}$$

$$\hat{\epsilon}_{-}^{*} \cdot \vec{E} = \ amplitude\ of\ right\ circular\ polarized\ component. \tag{2.34}$$

The corresponding intensities are then simply the squares of the amplitudes, in accordance with established principles of optics. Since we are dealing with complex quantities, we might be interested in their magnitudes, as well as their real and imaginary parts. The *Stokes parameters* in the circular polarization basis are therefore defined as:

$$I = |\hat{\epsilon}_{+}^{*} \cdot \vec{E}|^2 + |\hat{\epsilon}_{-}^{*} \cdot \vec{E}|^2 \tag{2.35}$$

$$Q = 2\ Re\left[(\hat{\epsilon}_{+}^{*} \cdot \vec{E})^{*}(\hat{\epsilon}_{-}^{*} \cdot \vec{E})\right] \tag{2.36}$$

$$U = 2\ Im\left[(\hat{\epsilon}_{+}^{*} \cdot \vec{E})^{*}(\hat{\epsilon}_{-}^{*} \cdot \vec{E})\right] \tag{2.37}$$

$$V = |\hat{\epsilon}_{+}^{*} \cdot \vec{E}|^2 - |\hat{\epsilon}_{-}^{*} \cdot \vec{E}|^2 \tag{2.38}$$

or

$$I = |E_{+}|^2 + |E_{-}|^2 \tag{2.39}$$

$$Q = 2\ Re\left[E_{+}^{*} E_{-}\right] \tag{2.40}$$

$$U = 2\ Im\left[E_{+}^{*} E_{-}\right] \tag{2.41}$$

$$V = |E_{+}|^2 - |E_{-}|^2. \tag{2.42}$$

Substituting the real amplitude and phase description of the complex coefficients, $E_{\pm}$, one obtains:

$$I = A_{+}^2 + A_{-}^2 \tag{2.43}$$

$$Q = 2A_+A_- \cos(\delta_- - \delta_+) \tag{2.44}$$

$$U = 2A_+A_- \sin(\delta_- - \delta_+) \tag{2.45}$$

$$V = A_+^2 - A_-^2. \tag{2.46}$$

Now, since the square of the amplitude is physically interpreted as the intensity of the beam, one can make a few observations about the parameter set $\{I, Q, U, V\}$:

1. The parameter I represents the total intensity of the beam, which as the reader recalls is actually a general superposition of two arbitrary plane waves. This is a physical observable.

2. The parameters Q and U give information on the *relative* phase difference between the two plane waves. The phase is a mathematical tool, and as such does not have a strict physical interpretation. However, the parameters Q and U represent the intensity of linearly polarized light along two orthogonal directions, and this *is* a physical observable.

3. The parameter V represents the difference in intensity between the positive and negative helicity components. Put another way, it gives the prevalence of left-circularly polarized light over right-circularly polarized light and is a physical observable. This *net circular polarization* will play a pivotal role in the inference of magnetic fields on the surface of the sun.

The four Stokes parameters are not independent from each other, in fact it can be shown from simple algebraic manipulation that the following relationship holds:

$$I^2 = Q^2 + U^2 + V^2. \tag{2.47}$$

However, this is strictly only for monochromatic light, with a single-valued wavenumber ($k = \frac{2\pi}{\lambda}$), which is impossible to achieve in practice. In the real world, the observations typically contain a range of wavelengths centered about some particular wavelength, in which case the

Stokes parameters are replaced by their integrals over the observation window. In practice, then, the quasi-monochromatic radiation satisfies:

$$I^2 \geq Q^2 + U^2 + V^2. \tag{2.48}$$

Completely unpolarized, natural light has $Q = U = V = 0$, while for an arbitrary polarization state, the degree of polarization is defined as:

$$P = \sqrt{\frac{Q^2 + U^2 + V^2}{I^2}}. \tag{2.49}$$

The Stokes parameters are typically represented by a column 4-vector, $\vec{I}$, given explicitly by:

$$\vec{I} = \begin{pmatrix} I \\ Q \\ U \\ V \end{pmatrix}. \tag{2.50}$$

From this representation, it can be seen that the Stokes vector of an arbitrary beam of light can be decomposed as follows:

$$\vec{I} = \begin{pmatrix} I \\ Q \\ U \\ V \end{pmatrix} = \begin{pmatrix} I - \sqrt{Q^2 + U^2 + V^2} \\ 0 \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} \sqrt{Q^2 + U^2 + V^2} \\ Q \\ U \\ V \end{pmatrix}. \tag{2.51}$$

The first term on the right-hand-side of the above equation represents natural (unpolarized) light while the second term is the Stokes vector of a perfectly elliptically-polarized beam with a degree of polarization of unity. Therefore, it is seen that any beam of light can be represented by the superposition of a completely unpolarized state and a completely polarized state. This decomposition will play a vital role in trying to identify magnetic fields embedded in a non-magnetic background.

The Stokes vector will be an essential tool in characterizing the polarization properties induced in the observed solar light by the presence of magnetic fields at the photospheric

level. Now that we have built a firm mathematical foundation describing the instantaneous properties of a beam of light, the next section describes the propagation of light through some arbitrary dielectric medium. In subsequent sections and chapters, the notion of this "arbitrary" medium will be replaced with a characterization of the solar plasma present in the line-formation layer of a neutral iron absorption line.

2.3 The Physics of Radiative Transfer

Light carries energy and momentum of a definite value, and travels at a definite speed. One might ask, how much energy is carried through some area by light passing through the bounding curve of that surface? To answer this question, consider a bundle of light rays occupying an infinitesimal solid angle $d\Omega$ and containing quasi-monochromatic light of frequencies in the range $[\nu,\ \nu + d\nu]$. This bundle of rays passes through an area element dA. What is the change in energy crossing dA in a time interval dt? If we increase the area through which the energy flows, the energy flux should increase, so dE should be proportional to dA. Similarly, if the bundle is "opened" up by increasing $d\Omega$, rays from more oblique angles can cross the surface, increasing the flux. Larger intervals in frequency allow a higher spread of photon energies, contributing more to the change in energy, and finally increasing the windowing time dt allows a larger net count of photons to cross the surface. These arguments suggest that the amount of energy in frequency range $d\nu$ crossing the area dA in time dt can be calculated as:

$$dE = I_\nu dA d\Omega d\nu dt, \tag{2.52}$$

where the coefficient of proportionality I_ν is defined as the *specific intensity*, *brightness*, or *intensity*. For this work, this quantity will be referred to as the intensity of the beam, as was derived in the previous section.

But what if the ray bundles are not crossing some imaginary area *in vacuo*, but are traveling through some medium with a specific mass density, ρ? In passing through matter, the ray bundle may have energy added to or removed from it by emission and absorption, respectively. The photons that make up these hypothetical rays may be scattered out of the

beam itself, and emission from the medium may contribute external photons to the beam, further modifying the energy content and intensity. The derivation of radiative transfer equations in the following subsections follows the textbook by Rybicki and Lightman (1979).

In emission, the ray bundle increases in intensity as the background emits photons from its own physical processes. The emission coefficient, j_ν, is a unit rate-of-change measure of the emitted energy, given by:

$$dE = j_\nu dV d\Omega d\nu dt, \tag{2.53}$$

which can also be written in terms of the emissivity ε_ν (emitted energy per unit frequency per unit time per unit mass) as:

$$dE = \varepsilon_\nu \rho dV \left(\frac{d\Omega}{4\pi}\right) d\nu dt. \tag{2.54}$$

This gives the relationship between the spontaneous emission coefficient and the emissivity of the medium within the beam solid angle, $d\Omega$:

$$j_\nu = \frac{\varepsilon_\nu \rho}{4\pi}, \tag{2.55}$$

which fully characterizes the emission in an arbitrary medium of mass density, ρ. Rewriting equation (2.54) in terms of the emission coefficient then gives

$$dE = j_\nu dV d\nu dt, \tag{2.56}$$

but since the volume element dV is simply the product of the beam cross-section dA with the distance ds traveled in time dt, even this can be rewritten as:

$$dE = (j_\nu ds) dA d\Omega d\nu dt. \tag{2.57}$$

Therefore, by comparing this expression with the general form of the specific intensity given in equation (2.52), the increase in beam intensity due to spontaneous emission from the medium is given by:

$$dI_\nu = j_\nu ds. \tag{2.58}$$

This is physically intuitive, since this equation states that the background emission modifies the intensity of the beam by an amount proportional to the distance travelled through the

medium. A longer travel-distance implies a larger amount of accumulated photons taken from the medium.

In absorption, the intensity of the beam is decreased as the atoms in the medium absorb the energy of incident photons. The absorption coefficient, α_ν, represents the loss of intensity as the beam travels a distance ds. Unlike the background emission, the strength of the absorption depends on the intensity of the beam, since the more photons that are present, the more can be absorbed by the medium. Assume the medium has a number density, n. Each particle has an effective absorption cross-section of σ_ν, and so the total absorption area is $n\sigma_\nu dAds$. The decrease in intensity must rely on the distance traveled, the original intensity of the beam, and on the total number of absorbers in the volume element $dV = dAds$. Therefore, we may write:

$$dI_\nu = -n\sigma_\nu I_\nu ds, \tag{2.59}$$

where the (-) sign indicates an explicit decrease in intensity. More generally,

$$dI_\nu = -\alpha_\nu I_\nu ds \tag{2.60}$$

where α_ν is the absorption coefficient, and is sometimes characterized by the macroscopic mass density and opacity coefficient of the medium, as opposed to the microscopic quantity, σ_ν:

$$\alpha_\nu = n\sigma_\nu = \rho\kappa_\nu. \tag{2.61}$$

The opacity of the medium describes the "difficulty" of light in penetrating the medium, and is generally a function of frequency, as denoted by the "ν" subscript.

To account for photon scattering in and out of the beam, a somewhat different approach is needed. Under the assumption of isotropic scattering (direction-independent) as well as coherent scattering (energy emitted = energy absorbed), the emission coefficient for scattering is obtained as:

$$j_\nu^{(s)} = \sigma_\nu J_\nu^{(s)} \tag{2.62}$$

where σ_ν is the *scattering absorption coefficient* and:

$$J_\nu^{(s)} = \frac{1}{4\pi}\int I_\nu d\Omega \tag{2.63}$$

represents the mean intensity, averaged over solid angle. The source function for scattering, defined as the ratio of emission to absorption, is then:

$$S_\nu = \frac{j_\nu^{(s)}}{\sigma_\nu} = J_\nu^{(s)}, \tag{2.64}$$

which is intuitive, since scattering can be treated as a "random" process, and as such, should be characterized by some average behavior of the beam. When treated simultaneously with absorption, this leads to a source function that is a weighted average:

$$S_\nu = \left(\frac{\alpha_\nu}{\alpha_\nu + \sigma_\nu}\right) B_\nu(T_e) + \left(1 - \frac{\sigma_\nu}{(\alpha_\nu + \sigma_\nu)}\right) J_\nu^{(s)}, \tag{2.65}$$

which significantly complicates the formal solution to the polarized radiative transfer equations (PRTE) because it is then an *integro-differential equation* for I_ν. For now, we neglect the effects of scattering on the source function.

2.3.1 The Scalar Radiative Transfer Equation

With the contributions to the change in intensity of a propagating beam given in the last section, we can assemble them into a single descriptive equation. The total change in beam intensity due to emission, absorption, and scattering can then be written as a differential:

$$dI_\nu = dI_\nu^{(e)} + dI_\nu^{(a)} + dI_\nu^{(s)}, \tag{2.66}$$

or more explicitly as:

$$dI_\nu = j_\nu ds - \alpha_\nu I_\nu ds + Z^{(s)} ds, \tag{2.67}$$

where $Z^{(s)}$ is a generic scattering term to be derived. Note also that the scattering has two components: into the beam and out of the beam, so for now the sign of $Z^{(s)}$ is indeterminate. Dividing through by the infinitesimal path length ds gives a (integro-) differential equation for the beam intensity as a function of path length, called the scalar radiative transfer equation (SRTE):

$$\frac{dI_\nu}{ds} = j_\nu - \alpha_\nu I_\nu + Z^{(s)}. \tag{2.68}$$

It is convenient to redefine the scale of the problem in terms of *optical* depth, instead of the geometrical depth, s. The optical depth of a (point in the) medium describes the transparency of the medium with respect to the wavelength(s) that compose the light traveling through the medium. It takes a value of zero (0) at the surface layer of the medium, and increases as one goes deeper into the medium. It is traditionally defined at a reference wavelength $\tau_{5000} = 5000\text{Å}$, and the relationship between geometrical depth and optical depth is linear:

$$d\tau_\nu = \alpha_\nu ds \quad \longrightarrow \quad \tau_\nu(s) = \int_0^s \alpha_\nu(s')ds'. \tag{2.69}$$

Under this transformation, the SRTE takes the form:

$$\frac{dI_\nu}{d\tau_\nu} = S_\nu - I_\nu + \frac{Z^{(s)}}{\alpha_\nu}, \tag{2.70}$$

where S_ν is defined as the *source function* and is simply equal to the ratio of emission to absorption, as in the scattering case,

$$S_\nu = \frac{j_\nu}{\alpha_\nu}. \tag{2.71}$$

The formal solution to the SRTE can be obtained by multiplying by the integrating factor e^{τ_ν}, applying the chain rule for differentiation, renaming some variables and finally formally integrating the reduced differential equation. Ultimately, for the propagation of the beam, this formal solution of the SRTE gives:

$$I_\nu(\tau) = I_\nu(0)e^{-\tau} + \int_0^\tau S_\nu(\tau')e^{-(\tau-\tau')}d\tau' + \int_0^\tau \frac{Z_\nu^{(s)}(\tau')}{\alpha_\nu(\tau')}e^{-(\tau-\tau')}d\tau'. \tag{2.72}$$

This equation is the sum of three terms, with the following descriptions:

(1) The first term represents the initial beam intensity, exponentially diminished by absorption along the beam path.

(2) The second term represents the integrated emission along the beam path, also moderated by absorption.

(3) The third and final term represents the amount of scattering in and out of the beam, attenuated by absorption. This term requires slightly more interpretation. When α_ν is small, there is little absorption, so photons scattered into the beam are just as likely to be scattered back out, instead of being destroyed by absorption before it can reach the observer. This makes the scattering term important in the limit of low absorption by the medium. Conversely, when α_ν is large, the medium is heavily-absorbing, and a photon scattered into the beam will most likely be absorbed before reaching the observer. Therefore, scattering should not be important in the limit of high absorption coefficients, as the integrand of the scattering term is small.

We will disregard the further treatment of scattering and its effects on the polarization states of the beam as outside the scope of this dissertation, and will take $Z_\nu^{(s)} = 0$ from here on, unless stated otherwise. Note, however, that the framework for dealing with a specific functional form of $Z_\nu^{(s)}$ has been established.

2.3.2 The Polarized Radiative Transfer Equations (PRTE)

Presently, we have described the physics of how the *intensity* of a beam of light is modulated as it travels through some medium. We have made no mention of how the other characteristics of the beam, e.g., polarization, are affected. If the beam is in a polarized state, then the attenuation of the intensity (Stokes I) has effects on the polarized components as well, since they contribute directly to the beam intensity. It seems then, that a modification to the basic scalar radiative transfer equation is needed in order to fully describe the effects of the medium on the state of the light beam.

Since the parameters of interest are Stokes I, Q, U, and V, all of which represent some kind of intensity, it seems logical that a simple vectorization of the radiative transfer equation is in order, with appropriate modifications to the emission and absorption coefficients. Specifically, the emission coefficient will become the *emission vector* while the absorption coefficient becomes the *absorption matrix*, and the PRTE will then describe the propagation of polarized light through the medium.

Recalling the form of the Stokes vector from previous sections, we may then write:

$$\frac{d\vec{I}_\nu}{d\tau} = \vec{S}_\nu - \mathbf{K}_\nu \vec{I}_\nu, \tag{2.73}$$

where $\vec{S}_\nu$ is the source function *vector*, and $\mathbf{K}_\nu$ is the absorption *matrix*. Much like the emission coefficient was used to describe the increase in beam intensity as it traveled through some emitting medium, the emission vector builds on this by also describing the changes in beam polarization states due to the emitting medium. Similarly, the absorption matrix replaces the absorption coefficient, in order to describe the cross-talk (or interaction) between the Stokes parameters. For example, a change in beam intensity (Stokes I) may induce a change in the other Stokes parameters if the lost intensity was that from the preferential absorption of a particular linearly polarized state, for example. Therefore, the 4×4 matrix $\mathbf{K}_\nu$ acts to mix up the polarization states of the beam so that we will not have 4 independent equations for 4 independent quantities, but rather 4 *coupled* equations describing the intensity and polarization states of the beam. By extension, then, the formal solution to the PRTE is given as:

$$\vec{I}_\nu(\tau) = \vec{I}_\nu(0)e^{-\tau} + \int_0^\tau \mathbf{K}_\nu(\tau)\vec{S}_\nu(\tau)e^{-(\tau-\tau')}d\tau'. \tag{2.74}$$

The only fundamental difference between the PRTE and the SRTE is the effect of the absorption matrix, $\mathbf{K}_\nu$, on the polarization states, which are now coupled to the beam intensity through absorption processes.

2.3.3 The Geometry of Radiative Transfer

Many radiative transfer codes (e.g., Maselli et al. (2003), Rijkhorst et al. (2005), Baron and Hauschildt (2007)) exist for the calculation and analysis of stellar structure and evolution, including radiative output based on the PRTE. For a fully 3-D simulation of what an observer would see, the equations must be recast in spherical geometry, as is shown in Figure 2.1. The angular difference between the local normal to the surface (or more generally, an atmospheric layer) and the direction to the observer (line-of-sight) is

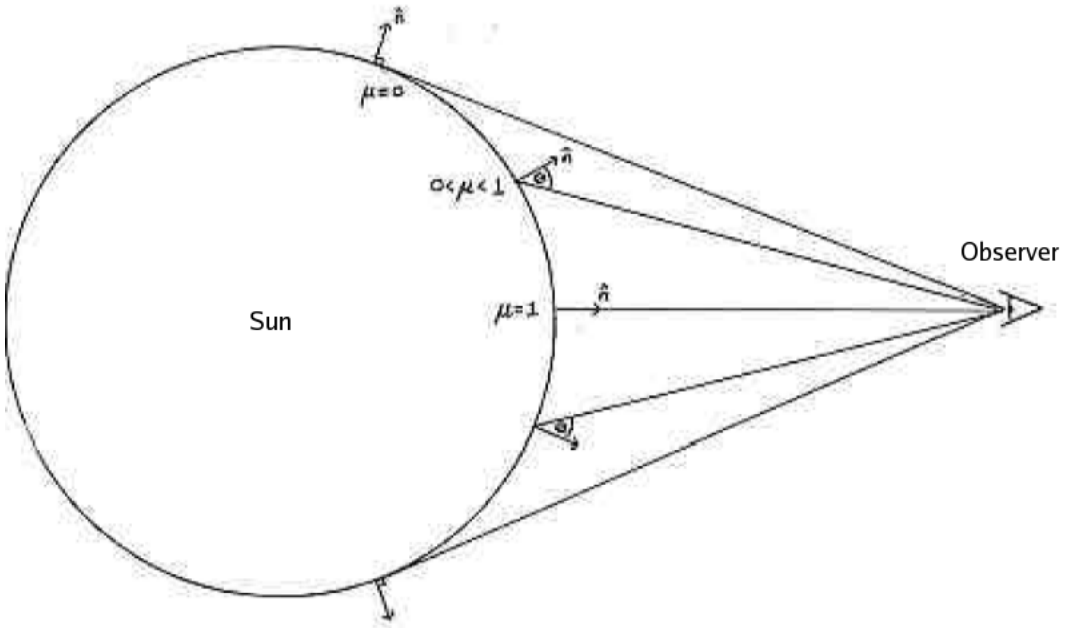

FIG. 2.1 Radiative transfer in spherical geometry. The spherical curvature of the surface within an SRTE/PRTE transfer problem may play an important role, for example, when the observer is relatively close to the source, or for full-disk observations of the source surface from a greater distance.

denoted by θ, and is a function of the direction to the observer, as can be seen in Figure 2.1. The PRTE in this case is modified to read:

$$\mu \frac{d\vec{I}_\nu}{d\tau} = \vec{S}_\nu - \mathbf{K}\vec{I}_\nu, \tag{2.75}$$

where $\mu = cos\theta$.

However, since the actual physical phenomena we are trying to study are (relatively) small structures at the level of the visible surface of the sun, the degree of spherical curvature is small. A typical solar active region (AR) is roughly the size of the Earth itself. For comparison, the width of the formation layer of the magnetically-active FeI λ6301.5Å spectral line is of the order of 100 km (Balasubramaniam, K.S., personal communication), a paltry distance compared to the diameter of the Earth. Therefore, the gentle sloping of the atmospheric layers may be well-approximated by a "stack" of planar atmospheric layers. This is called a *plane-parallel atmosphere*, and is the approach used throughout this work.

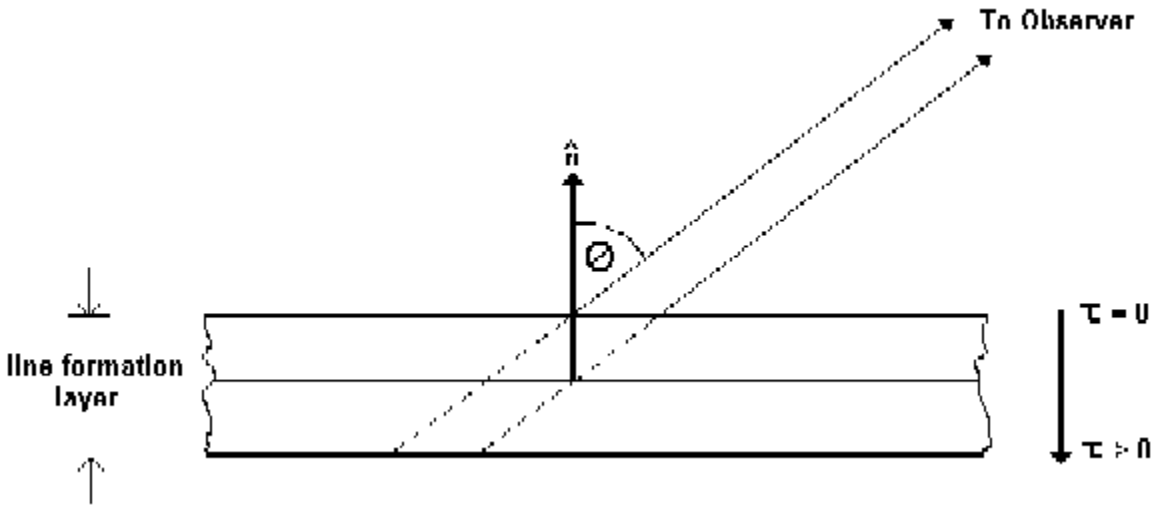

FIG. 2.2 Radiative transfer in plane-parallel geometry. The geometry of radiative transfer in a plane-parallel atmosphere contains a planar formation layer for some arbitrary spectral line. The dotted lines represent the integrated nature of the emission and polarization profiles along the observer's line-of-sight. When the vertical extent of the formation layer is relatively small compared with, say, the scale size of the atmospheric structures (surface convective cells, pores, sunspot structures), one may neglect the variation of the atmospheric parameters with position along the line-of-sight (to first order), for observations sufficiently far from the stellar/solar limb. The observer is assumed to be very distant from the source.

The geometry of the atmosphere and the beam is shown in Figure 2.2. Again, the quantity $\mu = cos\theta$ characterizes the orientation of the atmosphere relative to the observer, and in the plane-parallel case assumes a constant value for a given observation. For positions near the solar limb, $\theta = 90^\circ$ and therefore $\mu = 0$. Regions near disc-center have $\mu \leq 1$. For regions within $\approx 15^\circ$ of disc-center, an SRTE/PRTE analysis can safely set $\mu = 1$ with very little effect on the final results; only near the limb is μ appreciable enough to include explicitly.

2.4 Radiative Transfer in the Presence of Magnetic Fields

2.4.1 The Zeeman Effect

The *Zeeman effect* occurs when propagation of the beam of light through a plasma is perturbed by the presence of an external (background) magnetic field that permeates the plasma. This causes the precession of an electron's magnetic moment vector around the

direction of the external magnetic field. The breakdown in the symmetry of the problem comes from the different magnitudes of the projections of the angular momentum states along the field direction. In optics, the Zeeman effect can be observed by the splitting of a spectral line into three distinctly-lobed components in the case of the *normal* Zeeman effect, and is shown in Figure 2.3. In the *anomalous* Zeeman effect, the pattern can be more complex, due to asymmetries between the level populations of the emitting atom and therefore the necessity of solving the equations of statistical equilibrium for the atomic energy levels (Auer et al. (1977a)). We treat the normal Zeeman effect in the case of the FeI λ6301.5Å spectral line.

One may recall that a quantum system (an electron in this case) can be described by several quantum numbers. The quantum number n specifies the electronic energy levels, while the quantum number l is related to the orbital angular momentum. Specifically, the orbital angular momentum of an electron with quantum number l is

$$L = l^* \frac{h}{2\pi} = l^* \hbar, \tag{2.76}$$

where $l^* = \sqrt{l(l+1)}$. The orbital magnetic moment, μ_L, is then given by

$$\mu_L = -\frac{e}{2m_e c} L = -l^* \frac{eh}{4\pi m_e c}. \tag{2.77}$$

There is another kind of angular momentum at the quantum scales, that of *spin* angular momentum. In many ways, it behaves similarly to the orbital angular momentum, and so we can treat it in generally the same manner. However, this treatment of the Zeeman effect relies heavily on the *coupling* between the spin and orbital angular momenta, shown schematically in Figure 2.4. The angular momenta, $\vec{S}$ and $\vec{L}$, vectorially add to yield the total quantum angular momentum, $\vec{J}$. In a similar vein, the vector magnetic moments add vectorially, yielding a generalized magnetic moment, $\vec{\mu}_{LS}$, that is not co-aligned with the angular momentum, $\vec{J}$. This leads to the precession of the coupled magnetic moments around the magnetic field.

We can use projections of the various magnetic moments to define the influence of the various quantum states on the splitting of the energy levels. The total angular momentum

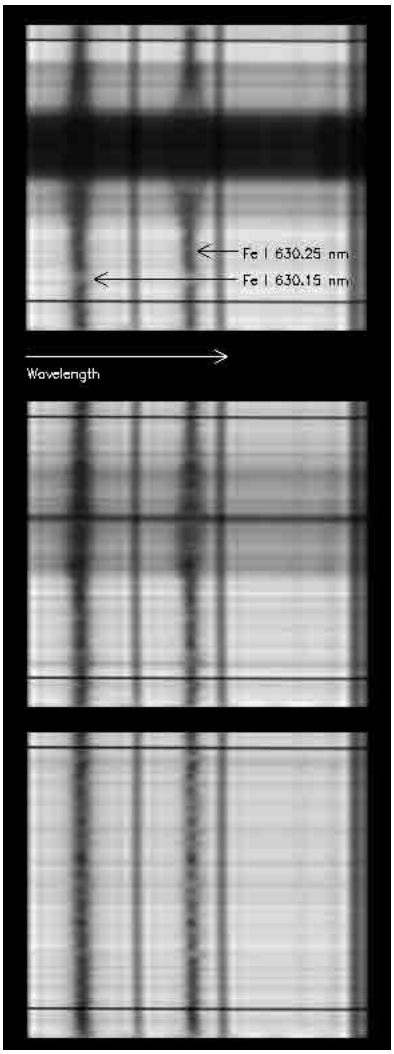

FIG. 2.3 Neutral Iron multiplet #816 in the umbra, penumbra, and quiet-sun. *(Top frame):* representative FeI $\lambda\lambda$6301.5, 6302.5 spectral lines as imaged by the ASP for a position within the umbra of AR9240. Note the clear splitting of the spectral line into multiple components, induced by the high magnetic field strengths in the umbra. *(Middle frame):* the same iron doublet, but for a position in the penumbra. *(Bottom frame):* FeI doublet in a quiet-sun region. The horizontal direction represents the spectral observation window, with each pixel having a spectral width of 11 mÅ. The vertical direction denotes the spatial position along the spectrograph slit entrance.

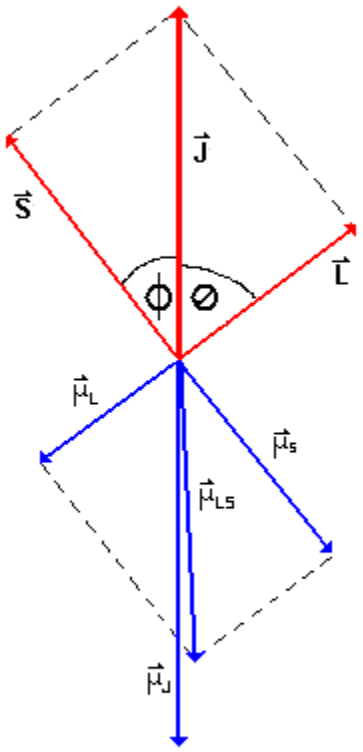

FIG. 2.4 Coupling between orbital ($\vec{L}$) and spin ($\vec{S}$) angular momenta. This coupling produces a quantized total angular momentum, $\vec{J}$.

vector is dependent on the angular momentum quantum number, $j = l \oplus s$, and assumes the form

$$\|\vec{J}\| = j^*\hbar = \sqrt{j(j+1)}\hbar. \tag{2.78}$$

Imposing the quantum condition that the component of the total angular momentum, $\vec{J}$, along the magnetic field direction is quantized as:

$$J_B = m_j\hbar, \tag{2.79}$$

where

$$m_j = \pm\frac{1}{2},\ \pm\frac{3}{2},\ \pm\frac{5}{2},\ \ldots\ ,\ \pm j. \tag{2.80}$$

The ratio of magnetic moment to angular momentum for both orbital and spin angular momentum are:

$$\frac{\mu_L}{L} = \frac{e}{2m_e c} = \frac{\mu_B}{\hbar} \tag{2.81}$$

$$\frac{\mu_S}{S} = 2\frac{e}{2m_e c} = \frac{2\mu_B}{\hbar}, \tag{2.82}$$

where μ_B is the constant Bohr magneton. The multiplying factor of 2 in equation (2.81) is a result of the fact that the spin angular momentum of a single electron is measured in terms of $\hbar/2$. The orbital and spin angular momenta satisfy the relations

$$\|\vec{L}\| = l^*\hbar = \sqrt{l(l+1)}\hbar \tag{2.83}$$

$$\|\vec{S}\| = s^*\hbar = \sqrt{s(s+1)}\hbar, \tag{2.84}$$

which, upon substitution, gives expressions for the magnetic moments, μ_L and μ_S:

$$\mu_L = l^*\mu_B \tag{2.85}$$

$$\mu_S = 2s^*\mu_B. \tag{2.86}$$

Now, the components of these magnetic moments along $\vec{J}$ are found by their relative angles (see Figure 2.4):

$$\mu'_L = l^*\hbar cos\theta \tag{2.87}$$

$$\mu'_S = 2s^*\hbar cos\phi. \tag{2.88}$$

The magnetic moment, μ_J of the total angular momentum is then given by their sum:

$$\mu_J = \left[l^* cos\theta + 2s^* cos\phi\right]\mu_B. \tag{2.89}$$

The quantity in brackets looks suspiciously like it could be proportional to the total angular momentum, in units of $\hbar$:

$$j^* = l^* + s^*, \tag{2.90}$$

so let us set

$$l^* cos\theta + 2s^* cos\phi = g_L j^*, \tag{2.91}$$

where g_L is a proportionality constant, historically called the *Landé g-factor*. Using some simple geometry and the law of cosines, one finds

$$l^{*2} = s^{*2} + j^{*2} - 2s^* j^* cos\phi \quad \longrightarrow \quad s^* cos\phi = \frac{j^{*2} - l^{*2} + s^{*2}}{2j^*} \tag{2.92}$$

$$s^{*2} = l^{*2} + j^{*2} - 2l^* j^* cos\theta \quad \longrightarrow \quad l^* cos\theta = \frac{j^{*2} + l^{*2} - s^{*2}}{2j^*}. \tag{2.93}$$

Substituting these into the expression explicitly containing the Landé g-factor gives

$$g_L = 1 + \frac{j^{*2} + s^{*2} - l^{*2}}{2j^{*2}} = 1 + \frac{j(j+1) + s(s+1) + l(l+1)}{2j(j+1)}. \tag{2.94}$$

Therefore we can write

$$\mu_J = g_L j^* \mu_B, \tag{2.95}$$

which, by Larmor's Theorem, means that the precessional frequency, ω_L, can be written as:

$$\omega_L = B\frac{\mu_J}{J} = g_L \frac{eB}{2m_e c}. \tag{2.96}$$

This is the angular frequency at which the magnetic moment precesses around the magnetic field, due to the torque exerted by non-alignment. If the electron is moving such that it has rotational kinetic energy in the *absence* of magnetic field, that energy is given by:

$$E = \frac{1}{2} I\omega^2, \tag{2.97}$$

where I is the rotational inertia, and ω is the angular frequency of the rotational motion, then in the presence of a magnetic field, Larmor's Theorem states that the total rotational energy is found from a simple superposition with the precessional frequency

$$E' = \frac{1}{2} I(\omega + \omega_L)^2. \tag{2.98}$$

The precessional energy is found from the difference of these two energies:

$$\Delta E = E' - E = \frac{1}{2} I\omega_L^2 + I\omega\omega_L, \tag{2.99}$$

but the precessional frequency is generally assumed to be smaller than that for the large-scale motions, and so only the second term above is important. Noting that the quantity

$I\omega$ is simply the angular momentum, we may write

$$\Delta E = (j^* \hbar cos\theta)\omega_L, \tag{2.100}$$

which contains the quantized angular momentum along the magnetic field direction. We may therefore write

$$\Delta E = g_L m_j B \mu_B, \tag{2.101}$$

where m_j is defined as before. Since this precessional energy in the presence of a magnetic field depends on the particular quantum state defined by g_L, transitions from one atomic energy state to another are split into $2j+1$ separate transitions that all contribute in their own way to the emission or absorption being observed.

Alongside the Zeeman effect is the energy shift due to spin-orbit interactions. Given without formal proof, this *fine-structure* correction is

$$\Gamma^{njls} = \frac{R\alpha^2 Z^4}{2n^3 l\left(l+\frac{1}{2}\right)(l+1)}\left[j(j+1) - l(l+1) - s(s+1)\right], \tag{2.102}$$

where R, α, and Z are the Rydberg constant, fine-structure constant, and atomic number, respectively. Therefore, the total energy shift due to the presence of an external magnetic field is given as:

$$\Delta E = g_L m_j B \mu_B + \Gamma^{njls}. \tag{2.103}$$

Not all transitions are allowed, however. There exist *selection rules* that allow only those transitions between the $2j+1$ Zeeman sublevels that satisfy

$$\Delta m_j = 0,\ \pm 1. \tag{2.104}$$

Figure 2.5 shows an example of the splitting of energy levels due to spin-orbit interactions and the presence of magnetic fields as well as the transitions allowed between the levels by the selection rules.

2.4.2 Faraday Rotation and Anomalous Dispersion

Up to now, we have described the propagation and attenuation of the intensity and polarization properties of a beam of light as it travels through some dielectric medium.

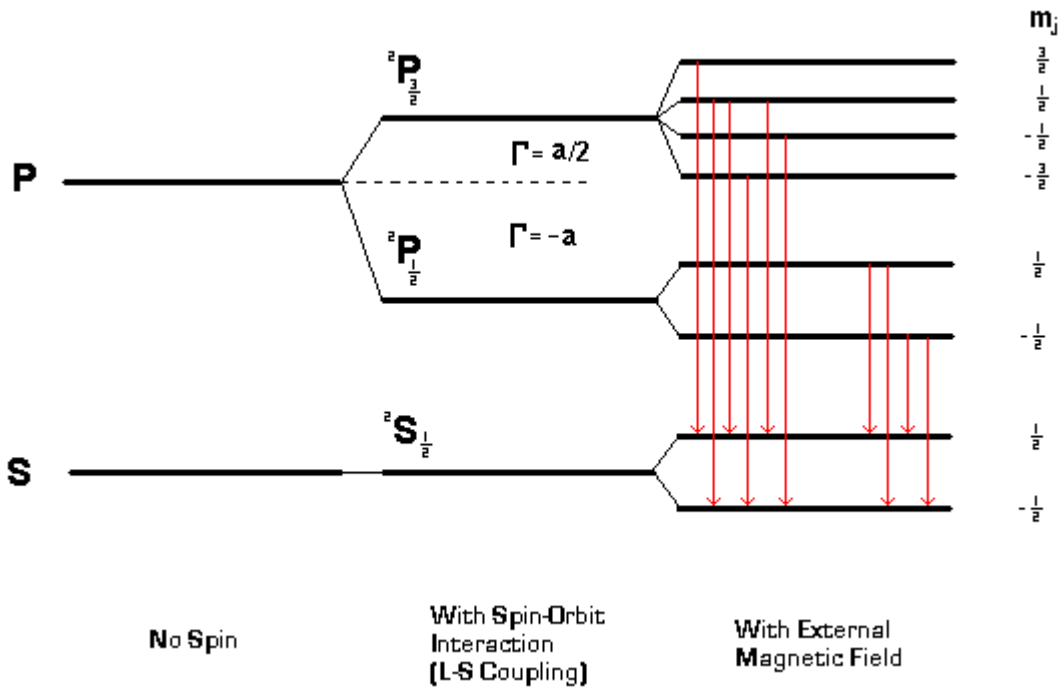

FIG. 2.5 Splitting of energy levels via the Zeeman effect. The splitting of an S-state ($l = 0$) and a P-state ($l = 1$) into multiple energy levels is caused by the inclusion of fine-structure corrections and the Zeeman effect. The red arrows indicate all the allowed transitions among the Zeeman sublevels that contribute to the ${}^2P_{\frac{3}{2}} \rightarrow {}^2S_{\frac{1}{2}}$ and ${}^2P_{\frac{1}{2}} \rightarrow {}^2S_{\frac{1}{2}}$ transitions. The constant a represents the factor multiplying the quantity in brackets in equation (2.102).

Since the purpose of this dissertation is to develop a modeling method that can handle radiative transfer in the highly-magnetized plasma of solar photospheric active regions, we must incorporate magnetic fields within the PRTE framework, and analyze their effects.

Assume there is a constant magnetic field $\vec{B}_0$ oriented parallel to the direction of propagation of the beam. That is, the beam travels *along* $\vec{B}_0$. What effects are observable in directions both parallel and orthogonal to the magnetic field lines? The answer to this question is that the plane of polarization of a linearly polarized wave undergoes a rotation as the beam travels through the medium.

Ultimately, this is due to the breakdown in directional symmetry within the problem itself. Prior to the inclusion of the magnetic field, the medium through which the beam

was propagating was homogeneous and isotropic. Now, we introduce a directional entity (a magnetic field), which has effects on the speeds at which the decomposed left- and right-circular polarized waves travel. This is a result of the slightly different indices of refraction for the $\sigma_{\pi,b,r}$ Zeeman components in the presence of a magnetic field (Jefferies et al. (1989)). Since we can represent any polarization state in the circularly-polarized basis of ($\hat{\epsilon}_+$, $\hat{\epsilon}_-$), a velocity difference translates into a phase difference upon superposition, which causes the plane of linear polarization to be $> 0^\circ$ and continually and smoothly changing as the beam propagates. As will be shown in a subsequent section, this Faraday rotation of the plane of linear polarization is intimately connected with the "anomalous dispersion" effects present in the mathematical framework of the PRTE.

2.5 Early Techniques

The first attempts at solving the PRTE yielded simplified, analytical expressions for the Stokes parameters under certain assumptions. The expressions became useful tools in analyzing the light received by solar telescopes and their respective external instrumentation. The problem of *inferring* the physical properties of a plasma through which a beam of light is being observed is typically referred to in the literature as an *inverse problem*. This is mainly because any quantity we want to "measure" must be derived from "second-order" information about the system, i.e., the propagating light beam, instead of direct *in-situ* data. The connotation of which is that, *given* the plasma parameters, it is generally an easier ordeal to directly calculate how the beam is attenuated by those parameters, which typically define the (non-)homogeneity and/or (non-)isotropy of the plasma. This is referred to as the *forward problem*. Not quite so straightforward is the situation where we are given the (already) attenuated beam, and wish to extract the properties of the plasma that performed the attenuation.

The first derivation of a solution to the forward problem was published by Unno (1956), under the assumptions of propagation through a collision-dominated plasma and a homogeneous magnetic field. These solutions were subsequently extended by Rachkovksy (1962) to

include the "anomalous dispersion" referenced earlier, which can be interpreted as extended σ_l, σ_π, and σ_r Zeeman components.

2.5.1 The Single Component Milne-Eddington Atmosphere

Unno (1956) first posited that an analytical solution to the PRTE could be obtained in the case of a *normal Zeeman triplet* under the assumption of local thermodynamic equilibrium, a depth-independent magnetic field and absorption matrix, and a source function that varied linearly with optical depth. A normal Zeeman triplet occurs when the field strength is sufficiently large so as to completely split the spectral line into three distinctly lobed components: the σ_b-component to the blue-side of the spectral line center, the σ_r-component to the red-side of line-center, and the π-component that resides at the line-center, as in Figure 2.3.

In this context, local thermodynamic equilibrium implies that the plasma is collision-dominated, meaning the collisions are frequent and therefore the atoms have very small collisional mean-free-paths, with internal radiative effects playing a less important role. A small collisional mean-free-path means that an atom or molecule travels a very small distance before interacting with another atom or molecule, small enough that the macroscopic properties of the plasma (temperature and pressure, for example) are essentially identical at the start- and end-points of the free path. A photon propagating through the plasma has a finite probability of being absorbed. If it is, the absorbing atom/molecule will typically travel, on average, only a collisional mean-free-path before the energy of the absorbed photon is redistributed through the plasma via collisional processes (Rybicki and Lightman (1979)). If this occurs on time-scales faster than the atom can re-radiate the photon, the plasma is said to be in local thermodynamic equilibrium (LTE). If, however, the collisionality of the plasma is low, the absorbing atom travels a significant distance before a collision, say, greater than some appropriate scale length (such as temperature or pressure), and therefore may or may not, probabalistically, have enough time to re-radiate the photon into the plasma at a different spatial location. This injection of the photon energy far from

the location of the photon's absorption is a non-local effect, as opposed to the more local conversion of photon energy to atomic kinetic energy in the case of LTE. Therefore, in this case, the plasma is said to be in non-local thermodynamic equilibrium (NLTE). Note that the probabilistic nature of re-radiation means that it is entirely possible for an atom to absorb a photon, travel a large mean-free-path *without* re-radiating it, collide with another atom, and re-distribute the photon energy as in LTE. Because of this property, the acronym NLTE is sometimes also interpreted as Not Limited to Thermodynamic Equilibrium.

Unno's original work was carried out in the linear polarization basis, as described in a previous section. He also defined the coordinate system so that $U = 0$ is always satisfied. The effects of "anomalous dispersion" (Faraday rotation) on the attenuation of the polarization state of the beam were neglected. The absorption matrix used by Unno had the form:

$$\mathbf{K}_\nu = \begin{pmatrix} 1+\eta_I & \eta_Q & 0 & \eta_V \\ \eta_Q & 1+\eta_I & 0 & 0 \\ 0 & 0 & 1+\eta_I & 0 \\ \eta_V & 0 & 0 & 1+\eta_I \end{pmatrix}. \tag{2.105}$$

With the assumption of the depth independence of the absorption matrix and a source function linear in optical depth

$$S_\nu = S_0 + S_1\tau_\nu, \tag{2.106}$$

Unno obtained the solutions

$$I(0) = S_0 + \mu S_1 \frac{1+\eta_I}{(1+\eta_I)^2 - \eta_Q^2 - \eta_V^2} \tag{2.107}$$

$$Q(0) = -\mu S_1 \frac{\eta_Q}{(1+\eta_I)^2 - \eta_Q^2 - \eta_V^2} \tag{2.108}$$

$$U(0) = 0 \tag{2.109}$$

$$V(0) = -\mu S_1 \frac{\eta_V}{(1+\eta_I)^2 - \eta_Q^2 - \eta_V^2}, \tag{2.110}$$

where the terms $\eta_{I,Q,V}$ are the absorption coefficients, which themselves depend on the atomic absorption coefficients per unit volume per unit solid angle for the π- and $\sigma_{b,r}$-

components of the normal Zeeman triplet:

$$\eta_I = \frac{\eta_\pi}{2} sin^2\psi + \frac{\eta_b + \eta_r}{4}(1 + cos^2\psi) \tag{2.111}$$

$$\eta_Q = \left(\frac{\eta_\pi}{2} - \frac{\eta_b + \eta_r}{4}\right) sin^2\psi \tag{2.112}$$

$$\eta_V = \frac{\eta_r - \eta_b}{2} cos\psi \ . \tag{2.113}$$

The terms $\eta_{\pi,b,r}$ represent the absorption profiles of the (π, b, r) Zeeman components, relative to the continuum absorption, and are given by the so-called Voigt function

$$H(a, v) = \frac{a}{\pi} \int_{-\infty}^{+\infty} \frac{e^{-y^2}}{(v - y)^2 + a^2} dy \tag{2.114}$$

as

$$\eta_\pi = \eta_0 H(a, v) \tag{2.115}$$

$$\eta_b = \eta_0 H(a, v - v_B) \tag{2.116}$$

$$\eta_r = \eta_0 H(a, v + v_B), \tag{2.117}$$

where η_0 is the continuum absorption coefficient in the absence of any magnetic field. The parameter a is the damping constant of the spectral line and is related to the collisionality of the plasma; v is a wavelength coordinate, relative to the spectral line center λ_0, normalized by the Doppler width of the spectral line

$$v = \frac{\lambda - \lambda_0}{\Delta\lambda_D}; \tag{2.118}$$

and v_B is the value of the wavelength split between the Zeeman components of the spectral line, also normalized by the Doppler width:

$$v_B = g_L \lambda_0^2 \frac{eB}{2m_e c \Delta\lambda_D}. \tag{2.119}$$

Note that the quantity $eB/2m_e c$ is simply the Larmor frequency for the precession of an electron around the direction of a magnetic field.

Rachkovsky (1962) extended Unno's approach to include the effects of Faraday rotation, which will couple the differential equations for the Stokes parameters and require the introduction of the magnetic field azimuthal angle χ, since we can no longer retain a reference

frame in which $U = 0$ always, due to the rotation of the plane of polarization caused by the presence of a magnetic field. The absorption matrix now reads:

$$\mathbf{K}_\nu = \begin{pmatrix} 1+\eta_I & \eta_Q & \eta_U & \eta_V \\ \eta_Q & 1+\eta_I & \rho_V & -\rho_U \\ \eta_U & -\rho_V & 1+\eta_I & \rho_Q \\ \eta_V & \rho_U & -\rho_Q & 1+\eta_I \end{pmatrix}. \tag{2.120}$$

where the extra terms $\rho_{Q,U,V}$ effect the coupling between the states of linear and circular polarization as the plane of the polarization is rotated by propagation through the plasma. The elements of this matrix are given by:

$$\eta_I = \frac{\eta_\pi}{2} sin^2\psi + \frac{\eta_b + \eta_r}{4}(1 + cos^2\psi) \tag{2.121}$$

$$\eta_Q = \left(\frac{\eta_\pi}{2} - \frac{\eta_b + \eta_r}{4}\right) sin^2\psi cos2\chi \tag{2.122}$$

$$\eta_U = \left(\frac{\eta_\pi}{2} - \frac{\eta_b + \eta_r}{4}\right) sin^2\psi sin2\chi \tag{2.123}$$

$$\eta_V = \frac{\eta_r - \eta_b}{2} cos\psi \ .$$

$$\rho_Q = \left(\frac{\rho_\pi}{2} - \frac{\rho_b + \rho_r}{4}\right) sin^2\psi cos2\chi \tag{2.124}$$

$$\rho_U = \left(\frac{\rho_\pi}{2} - \frac{\rho_b + \rho_r}{4}\right) sin^2\psi sin2\chi \tag{2.125}$$

$$\rho_V = \frac{\rho_r - \rho_b}{2} cos\psi, \tag{2.126}$$

which are straightforward extensions of equations (2.110)–(2.112). The newly-introduced terms, $\rho_{\pi,b,r}$, now represent the anomalous dispersion profiles, explicitly given by the Faraday-Voigt profile

$$F(a,v) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} \frac{(v-y)e^{-y^2}}{(v-y)^2 + a^2} dy, \tag{2.127}$$

as:

$$\rho_\pi = 2\eta_0 F(a,v) \tag{2.128}$$

$$\rho_b = 2\eta_0 F(a, v - v_B) \tag{2.129}$$

$$\rho_r = 2\eta_0 F(a, v + v_B). \tag{2.130}$$

The PRTE can then be solved analytically (e.g., by matrix inversion), much like Unno's solution, to give the full expression for the Stokes parameters at the surface, including Faraday rotation:

$$I(0) = S_0 + \frac{\mu S_1}{\Delta}(1+\eta_I)\left[(1+\eta_I)^2 + \rho_Q^2 + \rho_U^2 + \rho_V^2\right] \tag{2.131}$$

$$Q(0) = -\frac{\mu S_1}{\Delta}\left[(1+\eta_I)^2\eta_Q + (1+\eta_I)(\eta_V\rho_U - \eta_U\rho_V) + \rho_Q R\right] \tag{2.132}$$

$$U(0) = -\frac{\mu S_1}{\Delta}\left[(1+\eta_I)^2\eta_U + (1+\eta_I)(\eta_Q\rho_V - \eta_V\rho_Q) + \rho_U R\right] \tag{2.133}$$

$$V(0) = -\frac{\mu S_1}{\Delta}\left[(1+\eta_I)^2\eta_V + (1+\eta_I)(\eta_U\rho_Q - \eta_Q\rho_U) + \rho_V R\right], \tag{2.134}$$

where

$$\Delta = (1+\eta_I)^2\left[(1+\eta_I^2) - \eta_Q^2 - \eta_U^2 - \eta_V^2 + \rho_Q^2 + \rho_U^2 + \rho_V^2\right] - R^2 \tag{2.135}$$

$$R = \eta_Q\rho_Q + \eta_U\rho_U + \eta_V\rho_V. \tag{2.136}$$

It should be noted that equations (2.131)–(2.136) and the terms contained therein are all functions of wavelength, for which the explicit dependence was dropped for notational clarity. As a limiting case, the non-magnetic background may be obtained by setting $B = 0$. This gives $v_B = 0$, which leads to the conditions that $\eta_\pi = \eta_b = \eta_r$ and $\rho_\pi = \rho_b = \rho_r$. This gives the intensity of the spectral line in the absence of the magnetic field, since it is *not* the magnetic field that is responsible for *creating* the spectral line, but only for modifying it. Explicitly, the non-magnetic spectral line can then be calculated as:

$$I_{nm}(\lambda) = S_0 + \frac{\mu S_1}{1 + \eta_0 H(a, v)}, \tag{2.137}$$

where $H(a, v)$ is calculated as in the magnetic case, with the exception that the line-center wavelength may be different from that in the magnetic case. This can be used to account for the (expected) different Doppler shifts due to the presence of magnetic field-guided plasma flows. The observed profiles may actually be a superposition of the profile from a magnetic field that occupies a fractional pixel area, $\alpha \leq 1$ with that from the non-magnetic

TABLE 2.1 A summary of the parameters of the single-component Milne-Eddington model atmosphere.

Parameter	Description
$\|\vec{B}\|$	Magnitude (in G) of magnetic field.
ψ	Inclination (in degrees) of magnetic field from the line-of-sight.
χ	Azimuthal angle (in degrees) of external magnetic field.
α	Fraction of total pixel area occupied by magnetic field.
λ_0	Spectral line center (in Å)
$\Delta\lambda_D$	Doppler width (in mÅ) of spectral line.
a	Damping constant of spectral line.
η_0	Ratio of line-center to continuum opacity coefficients.
S_0	Source function (in units of intensity) at the surface ($\tau = 0$).
S_1	Source function gradient (in units of intensity) at the surface ($\tau = 0$).

background, which therefore occupies a fractional area of $1 - \alpha$. Therefore, the model polarization profiles may be mixed via the parameter, α:

$$\vec{I}(\lambda) = \alpha\vec{I}_m(\lambda) + (1 - \alpha)\vec{I}_{nm}(\lambda), \tag{2.138}$$

where $\vec{I}_{nm}(\lambda) = (I_{nm}(\lambda), 0, 0, 0)^T$.

Furthermore, the continuum intensity (the intensity of radiation at wavelengths far from the spectral line-center) can be obtained by setting $\eta_\pi = 0$, giving

$$I_c = S_0 + \mu S_1. \tag{2.139}$$

These equations describe the surface Stokes vector emerging from within a *Milne-Eddington* atmosphere, characterized by a constant absorption matrix, a linear source function, and a continuum intensity, I_c. The input parameters to the model and their physical interpretations are summarized in Table 2.1.

The Unno-Rachkovksy solution to the PRTE presents several conditions that can be used to ensure that a computational algorithm is functioning correctly. These solutions can

be found in Landi Degl'Innocenti (1976). Firstly, for a purely longitudinal magnetic field with $\psi = 0^\circ$ (field aligned with observer's line-of-sight), the following relation holds:

$$I(\tau, \lambda + \lambda_B) + V(\tau, \lambda + \lambda_B) = I_{nm}(\tau, \lambda). \tag{2.140}$$

For a purely transverse magnetic field with $\psi = 90^\circ$ and without loss of generality, $\phi = 0^\circ$ (orthogonal to observer's line-of-sight), the relevant identity is

$$I(\tau, \lambda) + Q(\tau, \lambda) = I_{nm}(\lambda). \tag{2.141}$$

Generalized Milne-Eddington Model

The assumption of a linear source function is used to simplify the formal solution to the PRTE, making it integrable. This is a simple case, however, and not all radiative transfer problems can be adequately addressed by a linear source function. This has led me to derive a general analytic expression for a Milne-Eddington atmosphere in which the source function is an arbitrary polynomial of order n. Granted, without a linear source function, the atmosphere cannot properly be called Milne-Eddington, but the rest of the Milne-Eddington assumptions remain identical. All polynomial forms are integrable in the formal solution to the PRTE, and I exploit this fact to derive the solutions. This may be a very handy approach for more complicated radiative transfer problems, since any function can be adequately approximated over some interval by some finite number of polynomial terms. The derivation follows.

Let us rewrite the PRTE in a more convenient form. Utilizing the notation of Stenflo (1994), the formal solution to the PRTE, giving the Stokes vector emerging from the surface, can be expressed as

$$\vec{I}_\nu(0) = \int_0^\infty e^{-\mathbf{K}_\nu \frac{\tau'}{\mu}} \vec{S}_\nu(\tau') d\tau'. \tag{2.142}$$

Integrating by parts, with the substitutions

$$\vec{u} = \vec{S}_\nu(\tau') \tag{2.143}$$

$$d\mathbf{v} = e^{-\mathbf{K}_\nu \frac{\tau'}{\mu}} d\tau' \tag{2.144}$$

gives the modified formal solution

$$\vec{I}_\nu(0) = \left[\mu(\mathbf{K}_\nu)^{-1} e^{-\mathbf{K}_\nu \frac{\tau'}{\mu}}\right]_0^\infty + \mu \int_0^\infty (\mathbf{K}_\nu)^{-1} e^{-\mathbf{K}_\nu \frac{\tau'}{\mu}} \left(\frac{d\vec{S}_\nu(\tau')}{d\tau'}\right) \frac{d\tau'}{\mu}. \tag{2.145}$$

Reducing the first term on the right-hand side yields

$$\vec{I}_\nu(0) = (\mathbf{K}_\nu)^{-1} \vec{S}_\nu(0) + \mu \int_0^\infty (\mathbf{K}_\nu)^{-1} e^{-\mathbf{K}_\nu \frac{\tau'}{\mu}} \left(\frac{d\vec{S}_\nu(\tau')}{d\tau'}\right) \frac{d\tau'}{\mu}. \tag{2.146}$$

Here, I propose a generalized polynomial form for the intensity source function:

$$\vec{S}_\nu = \left(\sum_{i=0}^{N} S_i \tau^i\right) \hat{\mathbf{e}} = (S_0 + S_1\tau + S_2\tau^2 + \ldots + S_N\tau^N)\hat{\mathbf{e}}, \tag{2.147}$$

where $\hat{\mathbf{e}} = (1, 0, 0, 0)^T$. Taking the derivative with respect to optical depth and substituting into equation (2.144), all integrand terms higher than linear in optical depth can be integrated by parts to finally give

$$\vec{I}_\nu(0) = \left[S_0 + \sum_{i=1}^{N} (-1)^i i \mu^i S_i (\mathbf{K}_\nu)^{-i}\right] \hat{\mathbf{e}}. \tag{2.148}$$

Neglecting coherent scattering and under the assumption of LTE, the source function can be written as:

$$\vec{S}_\nu = \mathbf{K}_\nu \begin{pmatrix} B_\nu(T_e) \\ 0 \\ 0 \\ 0 \end{pmatrix}, \tag{2.149}$$

where $B_\nu(T_e)$ is the Planck function at the local electron temperature, which assumes the optical depth dependence in the polynomial expansion above. The terms S_i in the polynomial source function can then be interpreted as being proportional to the i^{th} derivative of the Planck function, $B_\nu(T_e)$, with respect to optical depth, evaluated at the surface ($\tau = 0$). This is because equation (2.145) resembles a Taylor expansion of the exact source function at the surface.

2.5.2 Single Component Spectral Line Inversions

In this work, a *spectral line inversion* is defined as the determination of plasma properties based on the information content of observed light that falls within a particular wavelength range (an absorption spectral line) and its state of polarization.

Early spectral line inversion techniques (Landi Degl'Innocenti (1976), Auer et al. (1977b)) utilized the Unno-Rachkovsky solutions to match synthetic Stokes profiles to observational data with a non-linear least-squares Levenberg-Marquardt (LM) technique (Press et al. (1986)). The LM technique is an adaptive hill-climbing algorithm which systematically searches for improvements on the current parameter set, starting from some (presumably) educated initial guess. This is a relatively simple approach, based on the calculation of gradients of the χ^2 "goodness-of-fit" function with respect to the model parameters, which may be found analytically or calculated numerically. The χ^2 merit function is given by

$$\chi^2 = \frac{1}{4N - N_p} \sum_{\lambda} \left[\frac{\vec{I}_{obs}(\lambda) - \vec{I}_{model}(\lambda)}{\sigma_I(\lambda)} \right]^2 , \tag{2.150}$$

where $4N - N_p$ is the number of degrees of freedom inherent in the problem, and is the number of observations less the number of free parameters in the model. The number of observations comes from the four (4) Stokes polarization profiles, recorded in N identically-sized wavelength bins.

Landi Degl'Innocenti (1976) developed MALIP, a program for direct numerical integration of PRTE using the Unno-Rachkovsky solutions deep in the photosphere as boundary conditions, and a fourth-order Runge-Kutta method. In his 1976 paper, Degl'Innocenti gives a rather detailed description of the calculations carried out by his program, elaborating on several calculations of absorption coefficients, pressures, densities, etc., based on the detailed atomic physics contained in the references therein. However, no inversion results were presented, although the author states that several benchmark tests performed with the method returned results accurate to between one and ten percent.

Auer et al. (1977b) proposed an analytic inversion method that does not require integration in the case of negligible anomalous dispersion and a strong, constant magnetic field. Their absorption Mueller matrix therefore had the form of equation (2.104), and they took as a roughly representative model of conditions near the solar surface the Harvard Smithsonian Reference Atmosphere (Gingerich et al. (1971)). Synthetic spectra were generated with this model and known input magnetic fields, then the LM algorithm attempted to fit

the synthetic observations to the analytic surface Stokes profiles, recovering the four input parameters to within 5% in all cases. AHH demonstrated the applicability and reliability of a straightforward LM technique under the Milne-Eddington assumption. However, they made several simplifying assumptions, and as the amount of important physics in the problem increases, so does the mathematical and computational difficulty.

Landolfi and Landi Degl'Innocenti (1982) characterized the importance of magneto-optical (Faraday rotation, also known as anomalous dispersion and/or elliptical birefringence) effects, and their influence on the deduced free parameters, particularly the magnetic field azimuth. Unno's approach was to rotate to a frame of reference in which

$$U(\lambda) = 0 \quad \forall\lambda. \tag{2.151}$$

This is called the *preferred frame* of the problem, and the transformation itself has been used by other authors (e.g., Auer et al. (1977b)) to infer the azimuthal direction of the magnetic field. The coordinate system is rotated around the normal (beam-propagation) direction until the quantity

$$\sum_{\lambda} U^2(\lambda) \tag{2.152}$$

is minimized. However, even in the preferred frame, there occur the magneto-optical effects discussed earlier, generating non-zero $U(\lambda)$ signals. Therefore, the rotation of the coordinate system has erroneous effects on the precise determination of the field azimuth. LLD introduce an angle Φ as the "error angle" which represents the importance of magneto-optical effects. They found that, in the preferred frame, Stokes Q is most affected by magneto-optical effects, which leads to an error in azimuthal angle that is largest for small inclinations (field roughly aligned with line-of-sight) and a normalized Zeeman split $0.5 < v_B < 2.5$. In their tests and calculations, the error angle, Φ, typically assumed values between 10° and 20°, although errors as high as 40° were reported. They concluded, then, that magneto-optical effects are most important in regions of high magnetic field, and *must* be included in analyses of sunspot umbrae and penumbrae for precise determination of the magnetic field orientation.

Skumanich and Lites (1987) proposed several improvements on the AHH method, and added more relevant physics into the absorption matrix. Specifically, the analysis of AHH neglected the damping wings of the line, assuming the line-profile was Gaussian in shape. SL explicitly included the damping constant, Γ, and line-center position, λ_0, in the analysis, showing marked improvement over the fits to the spectral line data. They also included the field azimuth explicitly, opting not to perform the rotation into the preferred frame of the polarization, as well as incorporating the option of adding an extra "component" to the intensity that comes from either the non-magnetic patch of surface in which the magnetic field is embedded, or light scattered into the observations from surrounding regions.

Other (perhaps less useful) analytical solutions to the PRTE that could be used in a computational inversion procedure may be found in Hagyard (1971), Sidlichovsky (1975), and López Ariste and Semel (1999), although their derivations are much less transparent. For example, López Ariste and Semel (1999) base their derivation on an application of group theory to an ordering of seven operator-like terms. It is not clear to the author that any of these alternate derivations offer any particular advantages over a brute-force calculation directly from the PRTE.

2.6 Advanced Techniques

In a breakthrough paper by Rees et al. (1989), a technique for the formal integration of the PRTE was developed that allowed them to recover the variation of each model parameter with optical depth, over the entire line-formation region. The method has since been referred to as the DELO (Diagonal Lambda Operator) method, and is perhaps the most widely-used integration-based inversion technique to date, due to its accuracy, stability, and uniqueness. It utilizes a modified absorption matrix with zeroes on the diagonal to rewrite the PRTE in a more concise form, allowing a formal (analytical) integration between arbitrary optical depth nodes τ_k and τ_{k+1} that relates the Stokes vectors at those nodes. The procedure can then be applied iteratively until the full range of optical depth has been covered, ultimately yielding an expression for the emergent Stokes vector at the surface.

Landi Degl'Innocenti and Landi Degl'Innocenti (1977) first proposed a perturbative solution to the PRTE utilizing the concept of *response functions*, which serve to describe the modification of the surface Stokes profiles due to a perturbation in any of the model parameters at some arbitrary optical depth in the line-formation region. This approach is reminiscent of a Green's function, which describes the *response* of a linear system to a unit impulse or forcing term. The perturbative approach was subsequently extended by Ruiz Cobo and del Toro Iniesta (1992) to solve the PRTE using the explicit finite-difference and integration techniques of the DELO method at an arbitrary number of optical depth nodes. Their method also utilized an LM non-linear least-squares technique in which the response functions were shown to be related to the derivatives of the χ^2 merit function with respect to the model parameters at each optical depth node. By utilizing the DELO method, which yields, as an output, an approximation to the *matrix attenuation operator* (Landi Degl'Innocenti and Landi Degl'Innocenti (1985)), they were able to calculate the response functions in a very straightforward manner. The matrix attenuation operator, $\mathbf{O}(\tau, \tau')$, enters the formal solution to the PRTE, and physically describes the modification of the intensity and polarization of the light as it travels from an optical depth τ' up to τ. The operator must satisfy the following three conditions:

$$\frac{d}{d\tau}\mathbf{O}(\tau, \tau') = -\mathbf{K}(\tau)\mathbf{O}(\tau, \tau') \tag{2.153}$$

$$\mathbf{O}(\tau_0, \tau_2) = \mathbf{O}(\tau_0, \tau_1)\mathbf{O}(\tau_1, \tau_2) \tag{2.154}$$

$$\mathbf{O}(\tau, \tau) = 1. \tag{2.155}$$

With an approximation to the matrix attenuation operator given by the DELO method, Ruiz Cobo and Del Toro Iniesta were able to simultaneously calculate the response functions using a general expression first derived by Sánchez Almeida (1992). Explicitly, the response function for a perturbation in a parameter x at an optical depth τ, leading to a modification of the emergent Stokes vector is:

$$\vec{R}_x(\tau) = -\mathbf{O}(0, \tau)\left[\left(\frac{\partial}{\partial x}\mathbf{K}(\tau)\right)(\vec{I}(\tau) - \vec{S}(\tau)) - \mathbf{K}(\tau)\left(\frac{\partial}{\partial x}\vec{S}(\tau)\right)\right]. \tag{2.156}$$

These response functions enter the LM algorithm as derivatives of the χ^2 function, as well as explicitly generating the emergent Stokes vector:

$$\vec{I}(0) = \vec{I}_{DELO}(0) + \sum_j \int_0^\infty \vec{R}_j(\tau)\delta x_j d\tau. \tag{2.157}$$

This Stokes vector is then compared to the surface observational data to evaluate the χ^2 function and, depending on this comparison, the model parameters at each node are modified, leading to a new PRTE integration as well as the calculation of new response functions. This proceeds until the fractional decrease in χ^2 falls below some critical threshold.

Other, perhaps less-mainstream, approaches do exist for the inversion of solar Stokes profiles. Two such methods involve the use of Artificial Neural Networks (ANNs) and Principal Component Analysis (PCA), respectively. In the relatively new ANN approach (Carroll and Staude (2001), Socas-Navarro (2005)), a *neural network* accepts a spectral line profile as an input, while producing the set of physical quantities giving rise to the line as output. A neural network is nothing more than an interconnected set of *cells*, each of which holds a single scalar number, much like the memory cells in the RAM of a computer. The strengths and biases of the cell interconnections are determined by a process called *training*, whereby a (large) set of synthetic spectral lines with *a priori* known physical quantities is presented to the ANN. The goal of the training process is to iteratively adjust the network connections so that the ANN output is as close to the known physical parameters as possible. This adjustment is typically carried out by the so-called *back-propagation* algorithm described in Rumelhart et al. (1986). When the deviations of the output from the known parameters have been minimized, the network connections are frozen, and training is complete. Subsequently, new spectral lines not explicitly in the training set are presented to the ANN, to determine the quality of the inferred parameters. At its heart, the ANN approach is nothing more than a sophisticated interpolation routine which takes the spectral data and produces a set of parameters that is *representative* of the spectral line, based on what the ANN has "seen" in the training set. It should be noted that synthetic spectral lines are not the only means to generate a training set. A training

set composed of real solar observations that have already had a spectral line inversion performed on them may serve as a more realistic training set, due to the potential presence of more complex spectral line shapes. However, the major downfalls of this technique are that the generation of a sufficiently large training set may be prohibitively expensive (both in computing power and time), and the training itself may take a long time, depending on the ANN architecture.

The PCA technique relies on the decomposition of the observed spectral line into an orthonormal basis of "eigenprofiles", weighted by their respective eigenvalues (Rees et al. (2000), Eydenberg et al. (2005)). Again, this technique must compute a large database of synthetic profiles, from which an average profile and a covariance matrix are determined. The eigenvectors of this covariance matrix are utilized to form an orthonormal basis in which to expand the observations. If the spectral line, $\vec{I}$, composed of intensity measurements at N wavelengths, has the form

$$\vec{I} = (I_1, I_2, \ldots, I_N), \tag{2.158}$$

then the decomposition may be performed as:

$$\vec{I} = \overline{S} + \sum_{i=1}^{n} e_i \vec{s}_i, \tag{2.159}$$

where $\overline{S}$ is the average profile vector computed from the (large) synthetic profile database $\{\vec{S}_1, \vec{S}_2, \ldots, \vec{S}_M\}$. The covariance matrix, $\mathbf{C}$ is given by:

$$\mathbf{C} = \mathbf{X}\mathbf{X}^T, \tag{2.160}$$

where

$$\mathbf{X} = \left((\vec{S}_1 - \overline{S}), \ldots, (\vec{S}_M - \overline{S})\right) \tag{2.161}$$

is an $N \times M$ matrix. Therefore, the eigenvectors/eigenprofiles of the database may be obtained from

$$\mathbf{C}\vec{s}_i = \lambda_i \vec{s}_i. \tag{2.162}$$

The weighting coefficients, e_i, in the decomposition above are typically referred to as the *eigenfeature* vector

$$\vec{E} = (e_1, \ldots, e_n)^T. \tag{2.163}$$

The eigenfeature vector for *each* synthetic profile in the database can be computed, and the set of these vectors describes discrete samples of an n-dimensional surface called the *model manifold* (Rees et al. (2000)). The eigenfeature vector of the observed spectral line can then be compared to the model manifold, and the inversion process is then interpreted as finding the point on the model manifold that is a minimum distance from $\vec{E}_{obs}$. The "inversion" is not complete, however, since to date, only the first two leading orders in the decomposition can be physically interpreted. Skumanich and López Ariste (2002) found that the second- and third-order terms in the decomposition can be interpreted as the (scaled) first- and second-derivatives, respectively, of the first-order term. As such, they have shown how one can determine a "velocity" from the coefficients of the first derivative, and a "magnetic splitting parameter" from that of the second derivative in the weak-field approximation, where $\eta_0 \ll 1$. Although these parameters may arguably be the most important, their analysis of the information content of the PCA decomposition is far from a complete spectral line inversion.

2.6.1 Multi-Component Spectral Line Inversions

The presence of asymmetric or "anomalous" Stokes profiles hints at the existence of multiple distinct magnetic fields coexisting within the same resolution element (pixel). In the past, multi-component spectral line inversions were performed by hand, with no real automated inversion process. The inclusion of another magnetic component within the pixel is a straightforward extension of the single-component analytic solutions to the PRTE. By introducing yet another fill-factor, the emergent Stokes vector at the surface may be written as:

$$\vec{I}(0) = \alpha_1 \vec{I}_1 + \alpha_2 \vec{I}_2 + \alpha_{nm} \vec{I}_{nm}, \tag{2.164}$$

where the three fill-factors must satisfy the normalization condition

$$\alpha_1 + \alpha_2 + \alpha_{nm} = 1. \tag{2.165}$$

Note that α_{nm} may then be eliminated as a free parameter by writing the emergent Stokes vector as:

$$\vec{I}(0) = \alpha_1 \vec{I}_1 + \alpha_2 \vec{I}_2 + (1 - \alpha_1 - \alpha_2)\vec{I}_{nm}. \tag{2.166}$$

The modification to the number of free parameters in the two-component model is shown in the table below. Specifically, the addition of another set of magnetic field strength and direction parameters, as well as distinct spectral line centers for each component (to allow the possibility of each magnetic component harboring different line-of-sight flows) nearly doubles the size of the original single-component parameter set. It is assumed that the Doppler width, damping constant, and opacity coefficient of the spectral line are unchanged in the transition from the single- to the two-component model atmosphere. Descriptions of such inversion processes with multi-component magnetic fields can be found in Bernasconi and Solanki (1996), Leka (2001), and Borrero and Bellot Rubio (2002). As a final note, the advanced techniques described in a previous section are typically restricted to a single magnetic component, for computational tractability and efficiency.

2.6.2 Line-of-Sight Gradients

Coexisting magnetic elements are not the only explanation for anomalous Stokes profiles. Gradients along the line-of-sight in both large-scale plasma velocity and magnetic field have also been proposed as an alternate explanation. An analytical correction to account for these gradients may be obtained by a perturbative expansion of the PRTE, with the zeroth-order terms representing the model atmosphere in the absence of any gradients. Typically, this is usually adopted as the solution under the Milne-Eddington model atmosphere, although any suitable solution to the PRTE will suffice. The following derivation of the gradient corrections is a modification of the procedure found in Skumanich (2001). Writing the perturbations as

TABLE 2.2 A summary of the parameters of the two-component Milne-Eddington model atmosphere.

Parameter	Description
$\|\vec{B}_1\|$	Magnitude/strength (in G) of the first magnetic component.
$\|\vec{B}_2\|$	Magnitude/strength (in G) of the second magnetic component.
ψ_1	Inclination (in degrees) of first magnetic component.
ψ_2	Inclination (in degrees) of second magnetic component.
χ_1	Azimuthal angle (in degrees) of first magnetic component.
χ_2	Azimuthal angle (in degrees) of second magnetic component.
α_1	Fraction of total pixel area occupied by first magnetic component.
α_2	Fraction of total pixel area occupied by second magnetic component.
$\lambda_0^{(1)}$	Spectral line center (in Å) of sub-profile from first magnetic component.
$\lambda_0^{(2)}$	Spectral line center (in Å) of sub-profile from second magnetic component.
$\Delta\lambda_D$	Doppler width (in mÅ) of spectral line.
a	Damping constant of spectral line.
η_0	Ratio of line-center to continuum opacity coefficients.
S_0	Source function (in units of intensity) at the surface ($\tau = 0$).
S_1	Source function gradient (in units of intensity) at the surface ($\tau = 0$).

$$\delta\vec{I}_\lambda(\tau) = \vec{I}_\lambda(\tau) - \vec{I}^0_\lambda(\tau) \tag{2.167}$$

$$\delta\mathbf{K}_\lambda(\tau) = \mathbf{K}_\lambda(\tau) - \mathbf{K}^0_\lambda(\tau) \tag{2.168}$$

$$\delta\vec{S}_\lambda(\tau) = \vec{S}_\lambda(\tau) - \vec{S}^0_\lambda(\tau), \tag{2.169}$$

where the "0" superscript represents the zeroth-order quantities determined from the solution of the PRTE in the absence of gradients along the line-of-sight, the PRTE may be recast in terms of the perturbed quantities

$$\frac{d\delta\vec{I}_\lambda}{d\tau} = \mathbf{K}^0_\lambda\delta\vec{I}_\lambda + \delta\vec{e}_\lambda, \tag{2.170}$$

where

$$\delta\vec{e}_\lambda = \delta\mathbf{K}_\lambda\vec{I}^0_\lambda - \delta\vec{S}_\lambda. \tag{2.171}$$

The formal solution to this differential equation is mathematically identical to that for the zeroth-order PRTE itself:

$$\delta\vec{I}_\lambda(0) = -\int_0^\infty e^{-\tau'\mathbf{K}^0_\lambda(\tau')}\delta\vec{e}_\lambda(\tau')d\tau'. \tag{2.172}$$

Assuming the presence of a gradient in plasma velocity along the line-of-sight

$$v(\tau) = \overline{v} + (\tau - \overline{\tau})v'_{\overline{\tau}} \tag{2.173}$$

allows the perturbed absorption matrix to be expanded to first-order in the plasma velocity, ultimately yielding a functional form that is integrable in equation (2.172). Integration then yields

$$\delta\vec{I}_\lambda(0) = -\left[\left(\mathbf{K}^0_\lambda\right)^{-1}\vec{a}_0 + \left(\mathbf{K}^0_\lambda\right)^{-2}\vec{a}_1\right], \tag{2.174}$$

where $\vec{a}_{0,1}$ are constant vectors that are related to the unknown free parameters, $\overline{\tau}$ and $v'_{\overline{\tau}}$. These parameters can be determined by a low-dimensional least-squares fit to the *residual*

$$\vec{r}_\lambda = \vec{I}^{(d)}_\lambda - \vec{I}^0_\lambda, \tag{2.175}$$

where $\vec{I}^{(d)}_\lambda$ represents the observational data. By a similar procedure, corrections can be found for gradients in the magnetic field strength. Generally, I believe this procedure can be used to determine gradient corrections for *any* model parameter, no matter how the zeroth-order PRTE solution depends on that parameter. A generalized gradient correction may be forthcoming.

2.7 This Work

The inversion procedure developed in this work falls under what I would consider the category of non-mainstream techniques. I utilize a genetic-algorithm-based approach for its global exploration of the parameter space as well as its potential scalability to larger numbers of processors that is forseen to be needed to process the vast amounts of data

produced by current and next-generation solar observation equipment in any reasonable time.

In this work, I begin by performing spectral line inversions on the FeI λ6301.5 spectral line under the assumption of a single-component model. The spectral line under consideration originates in the photosphere, a layer of solar "surface" plasma that is collisionally-dominated and has historically been targeted by the design of polarization-measuring instrumentation. The next chapter will outline the observational data used in this work, the instruments used to collect the data and their operations, as well as various calibration considerations.

CHAPTER 3

DATA AND OBSERVATIONS

3.1 Introduction

The Advanced Stokes Polarimeter (ASP) is a spectropolarimetric instrument designed to work with the Dunn Solar Telescope (DST) located at the National Solar Observatory in Sunspot, New Mexico, USA. This chapter elaborates on the optical layout and imaging properties of the combination of the DST and the ASP, which is used to measure the linear and circular polarization profiles of select absorption spectral lines, and the data acquired by the ASP and used in this work are presented. Data obtained from the Diffraction-Limited Spectropolarimeter (DLSP) at NSO, as well as that from the Hinode satellite are also introduced after a brief introduction to their respective instrumentation.

3.2 The Dunn Solar Telescope (DST)

The National Solar Observatory (NSO) is currently located at Sunspot, NM, atop Sacramento Peak (elevation 9255 feet), and is home to one of the premier solar observational facilities, the Dunn Solar Telescope. The DST (shown in Figure 3.1) is one of the largest solar telescopes in the world, measuring approximately 135 feet from ground level to the top of the turret. However, much like an iceberg, only the tip is visible, as the telescope beam path plunges approximately 256 feet below ground, where the primary mirror is located.

Atop the turret is the entrance window leading to the optical path of the telescope. The window itself is 76.2 cm in diameter and is made of fused silica approximately 4 cm in thickness. Fused silica is used for its very small coefficient of thermal expansion, which means it can undergo large temperature changes with negligible expansion. This is indispensable in polarization studies, since the change in retardation (and its subsequent effect on polarization) due to thermal expansion is negligible over the course of an observing day. The turret has an external mount for holding sheet polarizers, which can be used to calibrate the polarization response of the telescope and optical elements. The primary mirror has a diameter of 152 cm, with an f-ratio of f/72, which forms an image of the solar

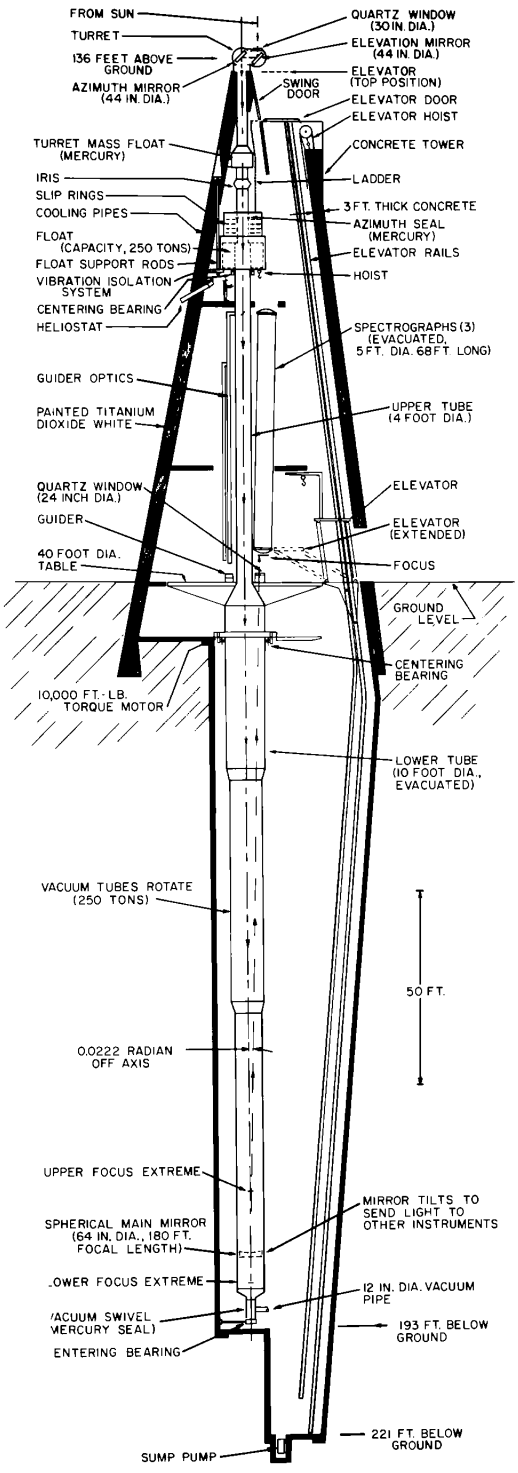

FIG. 3.1 The Dunn Solar Telescope (DST). This observational facility is located at the National Solar Observatory in Sunspot, NM. Since its construction, it has been one of the premier facilities for high-resolution solar observations.

disc just above ground level. The beam exits through a 20 cm diameter exit window, also made of fused silica, where it is then directed into one or more instrumentation setups. The entire optical path of the telescope is kept at a high vacuum, in order to eliminate any image motion due to thermal convective turbulence in the beam.

3.3 Calibration of the DST

Any telescopic system must be calibrated to ensure that the intensity signal that is measured is as close to the real intensity signal entering the telescope as is possible. This usually entails several different calibration steps, each one removing one source of spurious signal from the measurements.

Dark current calibration is performed to eliminate the natural thermal noise in the system. The thermally generated electrons within the CCD(s) used to record the observations leads to photoelectron counts that are larger than they should be, subsequently recording higher intensities than are present in the beam itself. To accomplish the dark current calibration, a *dark slide* is placed over the DST exit window, completely blocking the light from the DST from interacting with the remaining optical elements in the beam path. Reading out the CCD(s) under these conditions gives an image of the thermal electron generation within the CCD along with any registered thermal emission from the optical elements that is scattered into the CCD. Nominally, dark current calibration is performed many times, usually prior to the actual observations, then averaged to minimize statistical fluctuations. This average dark current image is then subtracted from the observations to eliminate the inherent thermal noise in the CCD.

Because of the slight pixel-to-pixel variability inherent in the CCD manufacturing process, each pixel will have a different response to the same amount of illumination. Thus, intensity is not measured uniformly across the entire CCD. The process of correcting this unavoidable defect is known as *flat-fielding*. The dark slide is placed over the DST exit window, and a uniform laboratory continuum light source is directed by a mirror into the optical assembly of the DST instrumentation. This light source produces a uniform inten-

sity that is far-removed from the influence of any spectral lines characteristic of the source. An image of the uniform source (called a *flat*) is then read-out from the CCD, displaying the pixel-to-pixel response to this near-perfectly uniform illumination. The flat is then dark-subtracted and used to generate the gain table of the CCD, defined as the mean of the flat divided by the flat (on a per-pixel basis), then subtracted from the observations of the day. Flat-fielding is typically performed once per day of observation.

A broadband filter, placed immediately after the open DST exit window, is used to reject any light of wavelength shorter than 410 nm, or longer than 720 nm. This spectrally reduced beam then passes through a polarization calibration setup. A linear polarizer and linear retarder in series are used for polarization response calibration, and a *rotating* retarder is used to modulate the polarization, altering the polarization of the light *after* exiting the telescope, but *before* entering any polarimetric instrumentation. The rotating retarder is constructed from a sheet polymer retarder sandwiched between BK7 substrates, used for its high transparency and scratch-resistance. This setup produces a retardance of 0.363 waves at 517 nm, and 0.298 waves at 630 nm, the primary operating wavelengths of the Advanced Stokes Polarimeter, detailed below.

A *field-stop* is placed at the primary focus, effectively reducing the field-of-view to $150'' \times 150''$. Just beyond the field-stop, a pair of mirrors is used to reflect the vertically-oriented beam into a horizontal plane roughly 5 feet above ground level, to facilitate the feeding of the beam into the Advanced Stokes Polarimeter.

3.4 The Advanced Stokes Polarimeter (ASP)

The Advanced Stokes Polarimeter (ASP) was developed jointly by the NSO and the High Altitude Observatory (HAO), located in Boulder, CO, for the express purpose of solar spectropolarimetry with high spatial, spectral, and temporal resolution suitable for examination of the photospheric and chromospheric magnetic structure on a variety of scales (sunspots, pores, faculae, plage, quiet-sun). It consists of a stand-alone polarimeter which, when coupled to the Echelle Spectrograph (ES) and (optionally) the Universal Birefringent

TABLE 3.1 Performance characteristics of the Advanced Stokes Polarimeter. "Polarization Systematic Errors" refers to the minimum amount of detectable peak polarization, in units of the continuum intensity. "Spectral Resolution" is defined as $\lambda/\Delta\lambda$, where λ is the central wavelength of the observational bandwidth, and $\Delta\lambda$ is the spectral pixel scale ($\sim$ 11 mÅ). Taken from Elmore et al. (1992).

Angular Resolution	$\leq$ 1"
Polarization Systematic Errors	$\leq 10^{-3}$
Signal-to-Noise Ratio	$\geq$ 1,000
Spectral Resolution	$\geq$ 200,000
Simultaneous Spectral Coverage	Chromosphere (Mg I 517.3, 518.3 nm)
	Photosphere (Fe I 630.1, 630.2 nm
	or Fe I 524.7, 525.0 nm)
Time Resolution	Active Region in 0.25 hours

Filter (UBF) instrumentation of the DST, allows the measurement of a variety of polarization states as a function of wavelength, thereby allowing the study of polarized radiative transfer phenomena, such as polarization spectral lines. Specifically, the ASP makes simultaneous measurements of multiplet spectral lines of neutral magnesium, which form in the chromosphere, and neutral iron, which form in the photosphere. Table 3.1 shows the spectral coverage and performance characteristics of the ASP.

The optical setup of the ASP is shown in Figure 3.2. The beam is impinged upon a beamsplitter (right triangle prism with coated hypotenuse), which partially reflects and partially transmits it. The reflected beam is then collimated by a lens and fed into a video-rate white-light camera system for real-time monitoring of the solar image. The transmitted beam is then focused onto the spectrographic slit of the ES. The slit is much narrower than it is tall, allowing the solar image to be scanned across the slit to produce spectra of the full image, one slit-position at a time. The ES slit-jaw is polished such that the incident light not directly falling on the slit is reflected back through the collimating lens and fed into the UBF. This allows the simultaneous viewing of the solar scene which surrounds

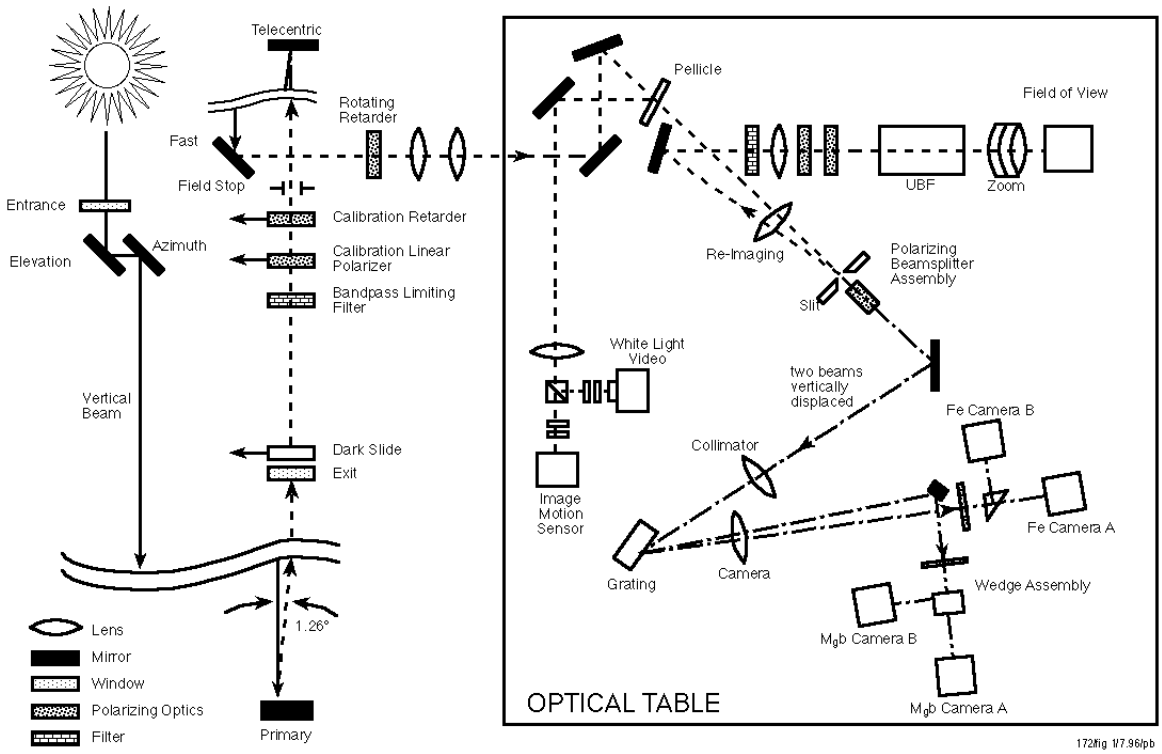

FIG. 3.2 A schematic of the Advanced Stokes Polarimeter (ASP). The instrumentation (not to scale) surrounded by the bold square belongs to the ASP assembly, while that on the left side of the figure are the feed optics (telescope beam path and associated instrumentation).

the current slit position in a 25 pm spectral bandpass. The light transmitted by the slit then passes through a polarizing beamsplitter assembly (see Figure 3.3) which produces two vertically-displaced beams corresponding to different input polarization states. Each of these beams is then incident upon the ES grating, which has dimensions of 250 mm $\times$ 125 mm. The grating contains 316 lines/mm with a blaze angle of 63.43°, and disperses the beams into wavelength-dependent spectra. The dispersed spectra are then filtered to isolate the magnesium lines from the iron lines (and vice versa). The magnesium- and iron-line spectra are simultaneously recorded by separate CCDs, which are coaligned with an accuracy of 0.1 pixel. The ASP control computers accumulate these recordings into eight images corresponding to eight modulation states of the incoming polarization. These images have a vertical spatial extent of 230 pixels and a horizontal spectral extent of 256 pixels,

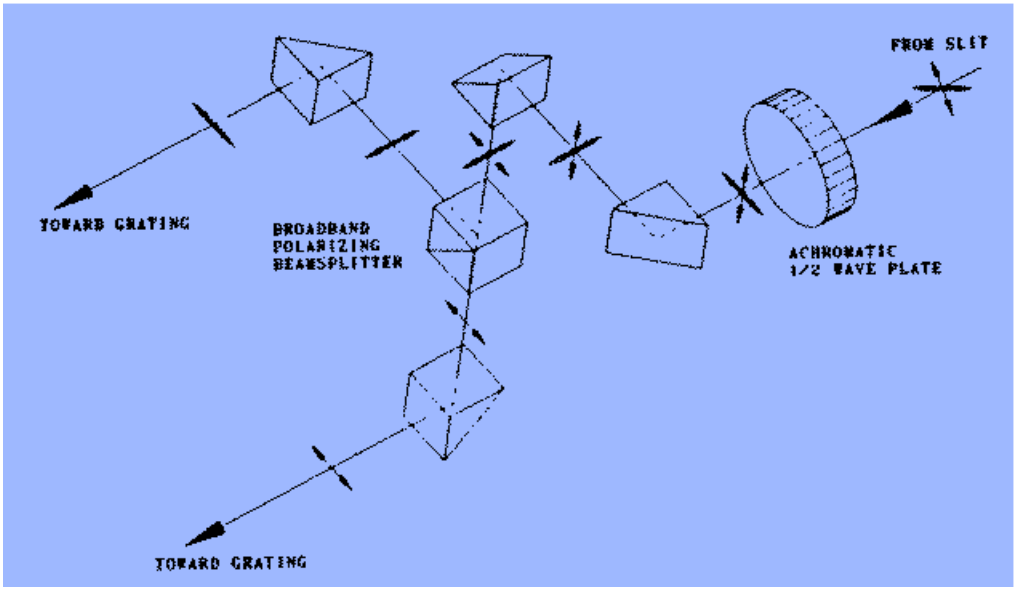

FIG. 3.3 The broadband polarizing beamsplitter assembly. This optical element produces two (2) vertically-displaced beams corresponding to different polarization states to be imaged by two (2) separate CCDs. The separation into two (2) beams is key for spatio-temporal modulation of the polarization signals.

and the Stokes parameters of the previous chapter are then particular *linear combinations* of these imaged modulation states.

The CCDs used in the ASP operate at 60 Hz, twice video-rate, in order to help minimize the effects of seeing-induced crosstalk between the polarization modulation states (Skumanich et al. (1997)). They are identical CCD chips with dimension 755×242 pixels, and are read-out by three serial output channels (one for every third column of pixels). The photoelectron read noise of the CCDs is typically 40–$56e^-$, which under typical conditions is much smaller than the Poisson statistics uncertainty ($\sqrt{N}$).

3.4.1 Calibration of the ASP

The measured Stokes vector is *not* the true Stokes vector. This effect is unavoidable, since the required presence of optical analysis elements will attenuate the intensity and polarization properties of a beam traversing the optical train of the telescope and polarimeter, albeit in known ways. Therefore, the true Stokes vector and measured Stokes vector are related to one another by the effects of the relevant optical elements, which can be modeled

by the equation:

$$\vec{S}_m = g\left(\mathbf{XT}\vec{S}_i\right) + b\hat{\mathbf{e}}, \tag{3.1}$$

where $\hat{\mathbf{e}} = (1, 0, 0, 0)^T$. The quantity g is the gain of the CCD, defined as the ratio of the number of photoelectrons per pixel to the pixel intensity value, and can be adjusted to control the saturation of the CCD pixels. The pixel bias, b, is a result of the fixed offset voltage for each pixel, due to manufacturing imperfections. The 4×4 matrices $\mathbf{X}$ and $\mathbf{T}$ represent the modification of the intensity and polarization due to the optical trains of the polarimeter and telescope, respectively. The polarimeter response matrix, $\mathbf{X}$, typically must be redetermined every 6 months or so, to account for the effects of aging of the polarimeter optical train elements. As mentioned in the last section, the broadband polarizing beamsplitter assembly converts a single input beam into two similarly-polarized beams to be imaged by two separate camera systems. The response matrix for each beam is dependent of the pixel-to-pixel gain variations, transmission coefficients, exposure time, the r.m.s. magnitude of the seeing displacement (which is related to the Fried parameter, r_0), and the mean fractional change (per unit angular displacement on the CCD) of the Stokes vector. The (quite complicated) details of the polarimeter calibration can be found in Skumanich et al. (1997), which provide exact expressions for the elements of the polarimeter response matrix, $\mathbf{X}$.

Just as the polarimeter inherently alters the state of the polarization it is trying to measure, so too does the collection of the light by the actual telescope assembly. However, it may be argued that the determination of the telescope response matrix is much simpler than that for the polarimeter, since only reflections and coordinate system rotations must be taken into account. The telescope response matrix may be determined from the matrix product of its individual optical elements, and the required translations into the different coordinate systems defined by those elements. Physically, the oblique reflections at the DST alt-az mirrors may introduce spurious linear polarization which mixes the Stokes I and Q parameters, while differential phase changes effect Stokes U and V. Furthermore, the

TABLE 3.2 Definitions of the telescope element matrices in the optical train of the DST.

$\mathbf{R}_{pol}$	Defines rotation from the coordinate frame of the DST exit window to the entrance of the polarimeter system
$\mathbf{D}_{exit}$	Mueller matrix of the DST exit window
$\mathbf{R}_{main}$	Defines rotation from the plane of incidence of the primary mirror to the coordinate frame of the optical tables
$\mathbf{M}_{main}$	Mueller matrix of the DST primary mirror
$\mathbf{R}_{az}$	Defines rotation from the plane of incidence of the turret azimuth mirror to the plane of incidence of the DST primary mirror
$\mathbf{M}_{az}$	Mueller matrix of the turret azimuth mirror
$\mathbf{R}_{el}$	Defines rotation from the plane of incidence of the turret elevation mirror to the plane of incidence of the turret azimuth mirror
$\mathbf{M}_{el}$	Mueller matrix of the turret elevation mirror
$\mathbf{D}_{ent}$	Mueller matrix of the DST entrance window
$\mathbf{R}_{sky}$	Defines rotation from the plane of the sky to the plane of incidence of the turret elevation mirror

high vacuum in the DST optical path introduces slight stress-induced birefringence in the entrance and exit windows, further altering the polarization states. Explicitly, these effects can be modeled and accounted for by the composite telescope matrix, given as:

$$\mathbf{T} = \mathbf{R}_{pol}\mathbf{D}_{exit}\mathbf{R}_{main}\mathbf{M}_{main}\mathbf{R}_{az}\mathbf{M}_{az}\mathbf{R}_{el}\mathbf{M}_{el}\mathbf{D}_{ent}\mathbf{R}_{sky}. \tag{3.2}$$

The physical definitions of each of the composite matrices are given in Table 3.2. Note that the telescope matrix acts upon a 4-element column vector, to the left. Therefore, it is built up from right to left, in the same order in which the DST beam traverses the optical train. That is, from the sky to the entrance window to the elevation mirror to the azimuth mirror to the main mirror (primary) to the exit window to the polarimeter.

The entrance and exit windows are taken to be pure retarders with a well-defined optical axis orientation and retardance, and the mirror components have well-defined Mueller matrices describing their reflective effects on the Stokes parameters. Pending non-singularity of the DST Mueller and polarimeter response matrices, this equation can be inverted to give the input (real) Stokes vector originating in the solar photospheric/chromospheric plasma:

$$\vec{S}_i = \mathbf{T}^{-1}\mathbf{X}^{-1}\frac{(\vec{S}_m - b\hat{\mathbf{e}})}{g}. \tag{3.3}$$

Thus, every measurement made by the ASP is converted into the real Stokes vector entering the DST entrance window by utilizing this transformation.

Unavoidably, instrumentation for measuring polarization introduces cross-talk between the Stokes I, Q, U, and V parameters, entangling the polarization measurements and introducing asymmetries where they ought not exist. Sánchez Almeida and Lites (1992) presented a rather nice scheme to estimate the degree of crosstalk for the Stokes II spectropolarimeter at HAO. However, they concluded that their accounting of their instrumental polarization was not able to describe the asymmetries seen in several specially-selected datasets, and so they must be of solar origin.

3.4.2 Data Obtained by ASP

The first dataset utilized in this work originates from spectropolarimetric observations of NOAA active region AR9240, displayed in Figure 3.4, which shows the dark central umbra, surrounded by the filamentary structure of the penumbra, as well as some of the surrounding solar granulation. As can be seen in the figure, the sunspot is very symmetric and comprises the vast majority of the active region, with the other component of this active region consisting of a diffuse group of pore-like structures flanking the main spot, which may be considered the opposite pole of this bipolar active region. At the time of the observations, AR9240 was within 10 degrees of disk-center. A set of subsections of the full ASP intensity and polarization maps, centered on the umbra of AR9240, is shown in Figure 3.5.

The spectral region observed and recorded by the ASP surrounds a spectral line multi-

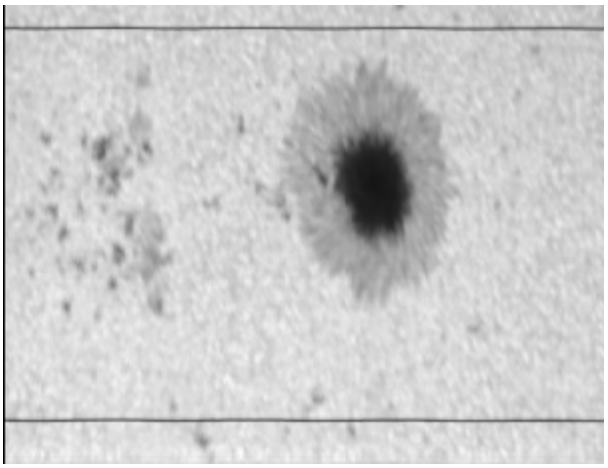

FIG. 3.4 NOAA Active Region 9240. The sunspot is actually very circularly symmetric, but due to the non-square pixels used by the ASP (0.5"×0.3"), the spot appears elliptical.

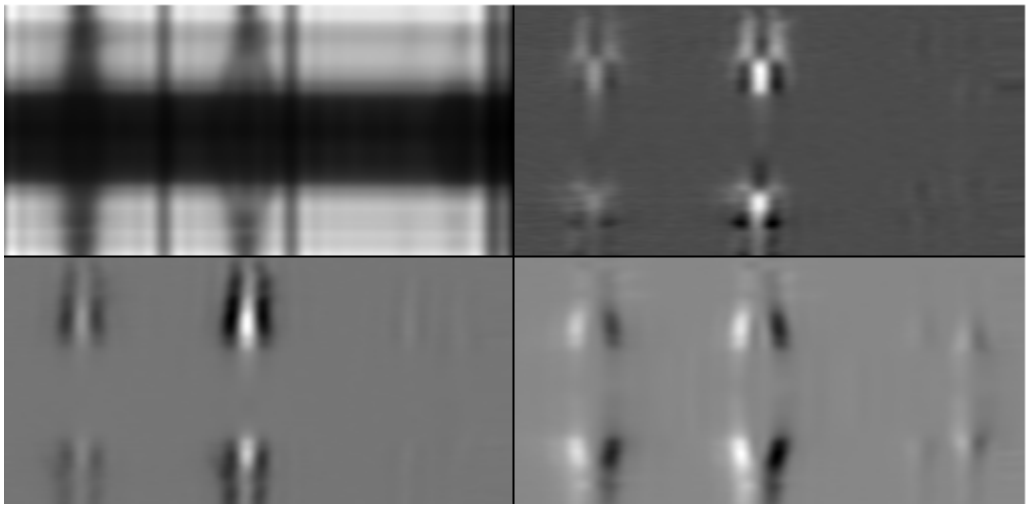

FIG. 3.5 Sample Stokes spectra from AR9240, imaged by the ASP. *(Clockwise from top left)*: Stokes I, Q, U, and V polarization parameters along the spectrograph slit, which runs through the umbra of AR9240.

plet of neutral iron, with the individual spectral lines lying at 630.15 nm and 630.25 nm, respectively, as detailed earlier in this chapter. The 630.15 nm line is a result of a $^5P_2$–$^5D_2$ electronic transition, where the traditional spectroscopic notation is used to denote the quantum state of the electron(s) in the upper and lower levels. Explicitly, the terms have the meaning

$$^{2S+1}L_J, \tag{3.4}$$

where S is the total spin angular momentum quantum number, L is the total orbital angular momentum quantum number, and J is the total angular momentum quantum number ($J = L \oplus S$). The Landé g-factors for the upper and lower levels of this transition take the values:

$$g_{upper} = 1.833 \qquad\qquad g_{lower} = 1.500. \tag{3.5}$$

These values can be used to define an *effective* g-factor that is useful when the magnetic splitting is small, given by Moore (1945) as:

$$g_{eff} = \overline{g} + \frac{\Delta J \Delta g}{4}(2\overline{J} + 1) = 1.667, \tag{3.6}$$

where $\overline{g}$ and $\overline{J}$ are the average of the upper- and lower-level g-factors and the average of the total angular momentum quantum numbers for the upper- and lower-level, respectively. ΔJ and Δg are the differences between the relevant upper- and lower-level quantities. The effective g-factor is useful for characterizing the magnetic sensitivity of spectral lines; lines with a higher effective g-factor typically have larger magnetic splittings for the same magnetic field strength. However, this comparison is usually only valid for lines in the same general region of the electromagnetic spectrum, since the splitting itself is quadratically dependent on the line-center wavelength of the spectral line. Therefore, an infrared line may have a larger magnetic splitting than a visible line with a larger effective g-factor.

The 630.25 nm line is produced by a $^5P_1$–$^5D_0$ electronic transition, with upper and lower Landé g-factors of 2.500 and 0.000, respectively. This leads to an effective g-factor of 2.500, indicating that this line is more sensitive to magnetic fields than the 630.15 nm line. This can be observationally verified, since this line typically shows greater splitting between

TABLE 3.3 Properties of the electronic transitions which produce the Fe I multiplet #816. The excitation energy is denoted by χ, g_{eff} is the effective Landé g-factor, ϖf is the degeneracy-weighted sum of the individual oscillator strengths of each Zeeman component that contributes to the transition. The oscillator strength is a quantum-mechanically derived transition probability dependent on the relative populations of the electronic energy levels.

λ_0	χ (eV)	g_{eff}	$log(\varpi f)$
6301.5091 Å	3.64	1.667	-0.58
6302.5017 Å	3.67	2.500	-1.15

the σ_b, π, and σ_r components than the simultaneously-observed 630.15 nm line. However, this line is contaminated in its long-wavelength wing by a terrestrial absorption line, and so for the present we neglect the 630.25 nm line and focus our efforts solely on the 630.15 nm line. Removal of the contaminating terrestrial line will be a subject of future research and will allow simultaneous analysis of both lines of the multiplet. For completeness, Table 3.3 shows some of the atomic properties of the transitions that give rise to the 630.15 nm and 630.25 nm spectral lines.

3.5 The Diffraction-Limited Spectro-Polarimeter (DLSP)

The Diffraction-Limited Spectro-Polarimeter was designed to be the high-resolution successor to the ASP by permanently observing the Fe I photospheric absorption lines at 6301.5Å and 6302.5Å. Its design is such that it can operate closer to the natural diffraction limit of the DST ($\sim 0.2''$), something that the ASP cannot do under typical seeing conditions. Furthermore, by virtue of its higher spatial resolution, the DLSP can take advantage of the DST adaptive optics subsystem, which gives the best image quality for pixel sizes of the order of $0.1''$ (Lites et al. (2003)). However, to affect this increase in spatial resolution, the DLSP sacrifices a slight amount of spectral resolution, sampling in $\sim$ 38mÅ wavelength bins, compared to the $\sim$ 12mÅ bins of the ASP. A schematic of the DLSP instrumentation is shown in Figure 3.6.

The DLSP has two modes of operation: high-resolution and low-resolution. The high-

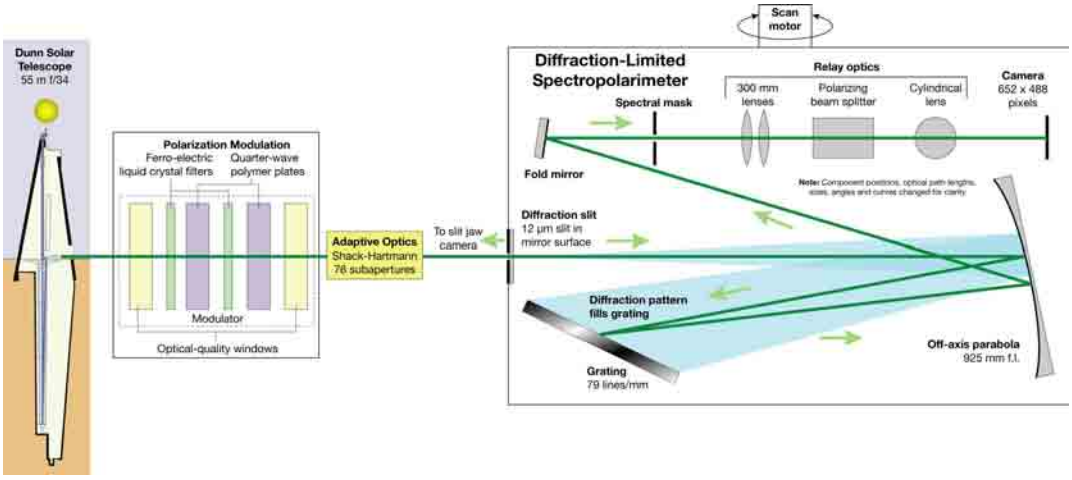

FIG. 3.6 The Diffraction-Limited Spectro-Polarimeter (DLSP). The compact design of the DLSP is attributed to its use of diffraction at the slitjaw to fill the dispersion grating. It utilizes an off-axis parabolic mirror both for redirection of the diffraction pattern to the grating and the subsequent collection and recording of the spectra so-produced. The polarization modulation package switches between four distinct states which are synchronized to the camera, recording images in four distinct polarization states integrated over time to enhance the S/N ratio. The Stokes I, Q, U, and V parameters are then the appropriate additions and subtractions of these polarization intensity images (Gullixson 2007). Courtesy of HAO/NSO.

resolution mode is typically used for investigations of active regions, and has a field-of-view of approximately $60''$ along the spectrograph slit, while the low-resolution mode is used for observations spanning about $180''$ along the slit. A comparison of the two modes, along with the field-of-view of the ASP, is shown in Figure 3.7.

Perhaps the most important piece of post-DST equipment is the adaptive optics subsystem, which can substantially improve the spatial resolution of the observations by partially eliminating the effects of terrestrial atmospheric turbulence.

3.5.1 The Adaptive Optics (AO) Subsystem

The image-blurring effects of seeing are a result of the turbulent motions of the Earth's atmosphere. This ultimately causes the plane wavefronts from the distant source to be distorted into very complicated wavefronts. When these complex wavefronts arrive at the telescope and are imaged by the CCD, the net image appears blurry, because of the super-

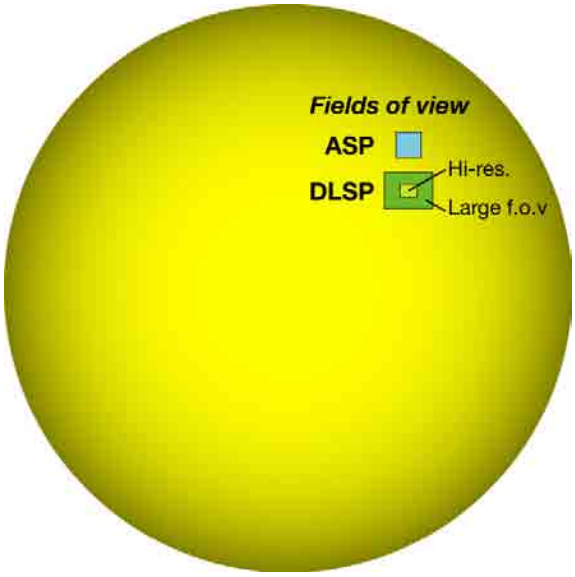

FIG. 3.7 Field-of-view modes of operation for the DLSP. The boxed fields-of-view of the DLSP show the relative differences between the DLSP high- and low-resolution modes of operation, as well as a comparison of both modes to the original field-of-view of the Advanced Stokes Polarimeter. Courtesy of HAO/NSO.

position of many distorted wavefronts. The characteristic size of the atmospheric pockets of gas responsible for producing such seeing effects is given by the Fried parameter, r_0, and was determined to have a median value of 8.7 cm at the NSO site (Brandt et al. (1987)).

The observational impact of the AO subsystem is depicted in Figure 3.8. Adaptive optics subsystems have been utilized in night-time astronomy for many years; however, the implementation of adaptive optics in solar astronomy is not quite as simple, since generally no *point* source is available on the solar surface and no point-spread functions are available. Instead, a solar implementation of AO must sense the wavefronts using the solar granulation, which is a low-contrast, *extended* source.

The DST AO subsystem utilizes a small mirror, 70 mm in diameter, which is divided up into 76 subapertures. 97 piezoelectric actuators are attached to the backside of the mirror, which push and pull on the mirror to subtly alter the shape of its reflective surface (typically on the order of a few μm). Piezoelectrics are used for their high linearity, and the actuators/mirror assembly are kept at an optimal operating temperature of 20° C, in

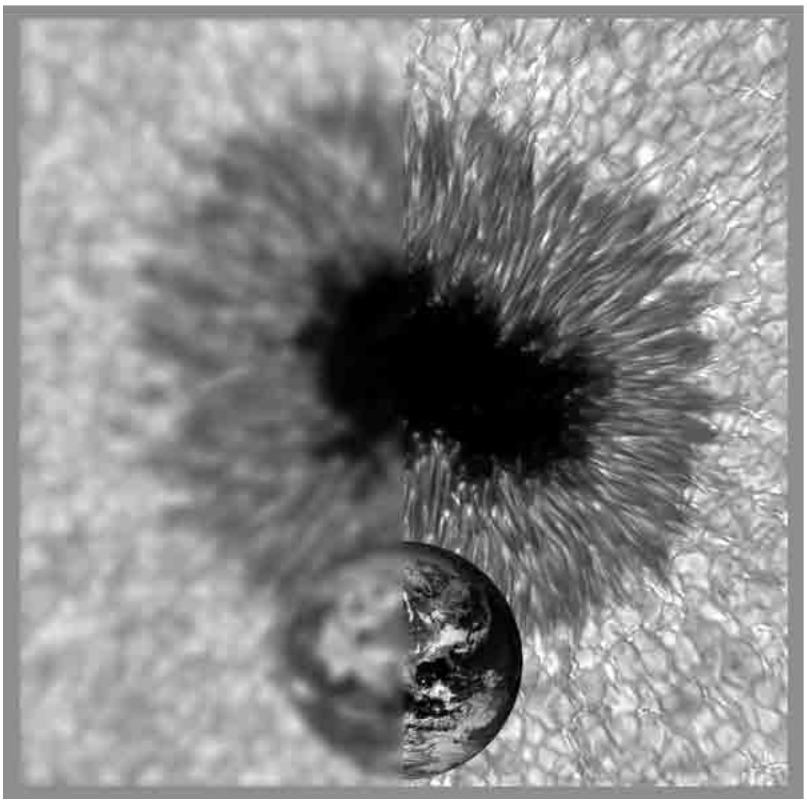

FIG. 3.8 The benefits of adaptive optics (AO). The figure shows the dramatic resolution improvement due to the AO subsystem. The left side of the image (showing AR10810) was artificially smeared out, to simulate uncorrected resolution, since no image was made with the AO subsystem turned off. The Earth is shown for size comparison. Courtesy of NSO/AURA/NSF.

order to minimize hysteresis. The light from the DST is split between the image motion stabilization system (correlation tracker), and the wavefront sensing equipment, which uses a Shack-Hartmann wavefront sensor (SH-WFS). The SH-WFS is sensitive to the gradient of the wavefront phase across the subapertures, and so is able to detect changes in the phase which can be used to reconstruct the image. The SH-WFS is optically fed by a lenslet array of 76 tiny lenses, one for each subaperture, that focuses and forms a separate solar image on each sensor subaperture. Each image is identical, with the exception of the phase information, which is dependent on the position within the subaperture array. These phase differences are determined and analyzed in near real-time, 130 times a second, and fed back to the deformable mirror, which changes shape accordingly to alter the incoming wavefronts

in such a way as to minimize the phase differences, effectively canceling out the effects of turbulent atmospheric gas pockets.

3.5.2 Data Obtained by DLSP

Figure 3.9 shows a size comparison of the field-of-view for the ASP observation of AR9240 and the subfield of an observation made by the DLSP in high-resolution mode and used in this work. The field-of-view for this DLSP observation was 43" by 34". The inset shows NOAA Active Region 10956 (AR10956), observed on May 20, 2007 between 14:59 and 15:25 UTC. During the observation period, the active region was located at 1.6° latitude and 14.7° longitude on the solar disc. As can be seen from the figures, the amount of discernable detail is very much improved over the ASP; individual convective cells are now visible, whereas they are blurred into an almost uniformly gray, slightly mottled background in the ASP image. This increase in spatial resolution should help the stability of the inversion, since the smaller the linear size of the pixel, the more confident one can be that it is occupied by a single thin fluxtube (or, at least, there is less room for other distinct fluxtubes). Figure 3.10 displays polarization intensity images of this active region, near the line-core of the Fe I 6301.5Å absorption line.

3.6 The Hinode Satellite

The Hinode (*sunrise*) satellite, of the Institute of Space and Astronautical Science, Japan Aerospace Exporation Agency, was launched on September 22, 2006, with the express science goal of understanding how photospheric variability affects the activity in the chromosphere and corona, leading to energetic events (Kosugi et al. (2007)). To this end, the major science instrument aboard Hinode, the Solar Optical Telescope (SOT), can feed into a spectropolarimeter to record Stokes polarization profiles. The spectropolarimeter which works in conjunction with the SOT actually served as a model prototype for the DLSP, so many of the operational details are identical to the DLSP (off-axis spectrograph recording the Fe I 6301.5Å and 6302.5Å absorption lines). The main, and perhaps most important, differences are the fact that the Hinode satellite is outside of the Earth's atmo-

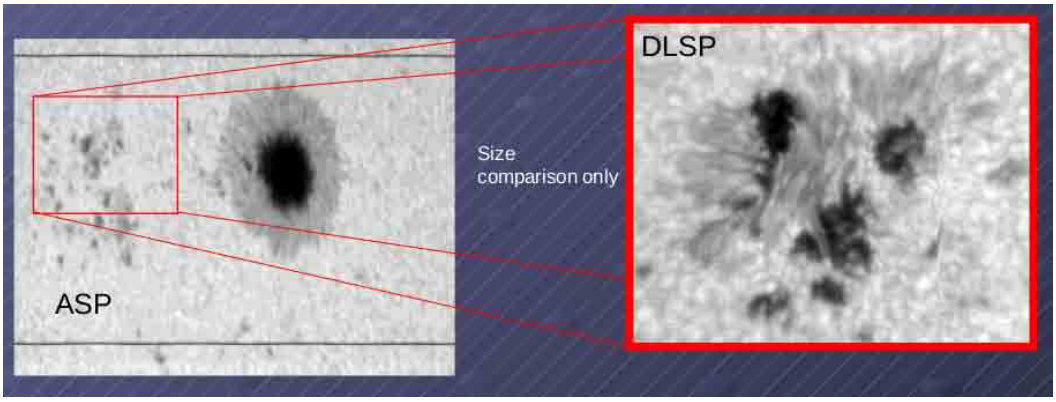

FIG. 3.9 A comparison between NOAA AR9240 and AR10956. *(Left Frame)*: continuum image of AR9240, imaged by the ASP. The image spans 90" in the horizontal direction and 115" in the vertical direction. *(Right Frame)*: for comparison, the inset shows a subfield of AR10956, imaged by the DLSP, spanning 43" (34") in the horizontal (vertical) direction. These active regions were imaged approximately 12 years apart; the image is for size comparison only.

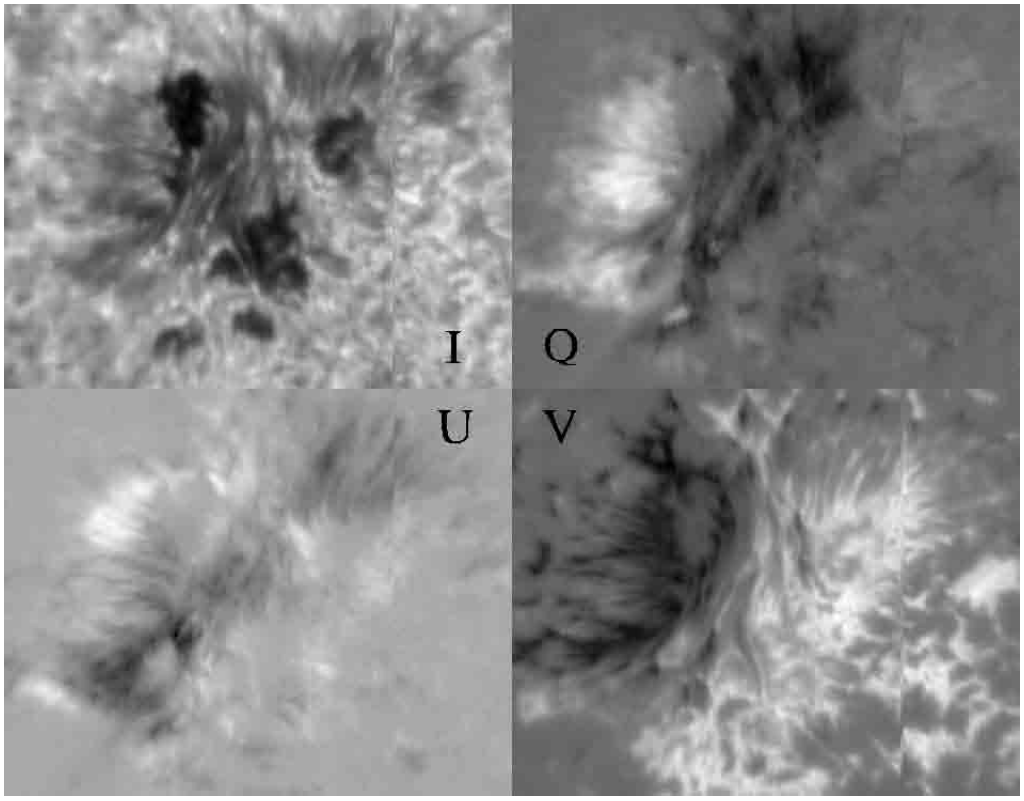

FIG. 3.10 Sample Stokes polarization images from AR10956, taken by the DLSP. These intensity images are formed from light in the line-core of the Fe I 6302.5Å absorption line, as obtained by the DLSP.

sphere, and therefore requires no adaptive optics correction to obtain high spatial resolution ($\sim 0.3''$ diffraction-limited at 630 nm), and has an improved spectral resolution of ~ 21mÅ per pixel. An example of the high-spatial-resolution data obtained by the spectropolarimeter aboard Hinode is shown in Figure 3.11.

Figure 3.11 shows the Stokes parameters in a $62'' \times 81''$ subfield of the full Hinode field-of-view, showing the main sunspot of NOAA Active Region 10923 (AR10923). This active region was observed on November 14, 2006 at 07:51 UTC, during which time the spot was very close to disc-center ($< 1^{\circ}$). Figure 3.12 shows example polarization spectra obtained for this sunspot.

The next chapter details the working procedures used in the genetic algorithm inversion code, the calibration of the data used in this work and presented in this chapter, and the heuristic preprocessing algorithms used to ensure a stable inversion.

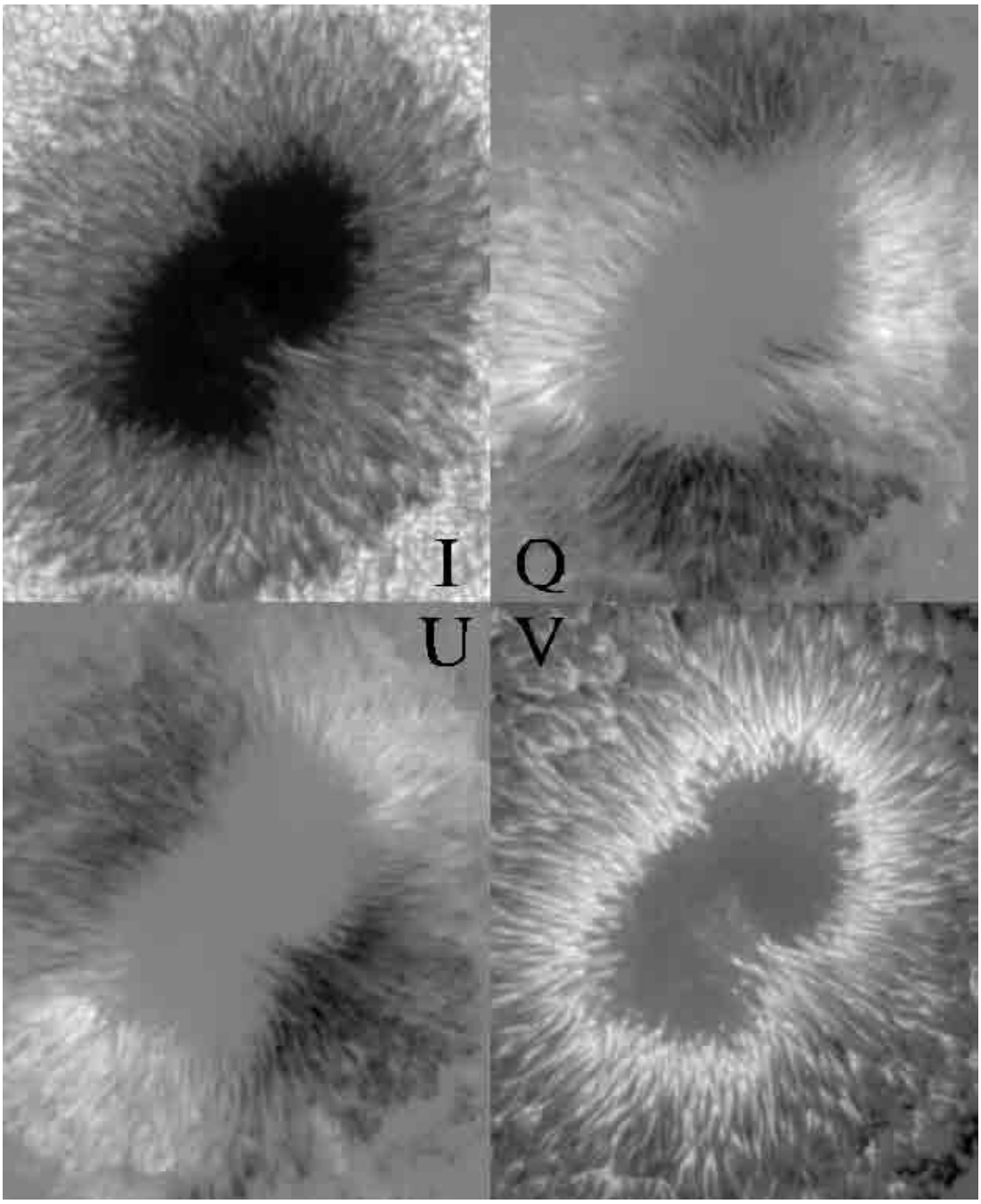

FIG. 3.11 Sample Stokes polarization images from AR10923, taken by the *Hinode* satellite. These Stokes I, Q, U, and V (clockwise from top-left) intensity maps are from a sunspot observed by the Hinode satellite in November 2006. It is clear from this image how the linear polarization signals (Stokes Q and U) are dependent upon the orientation angle of the magnetic field in the plane of the image.

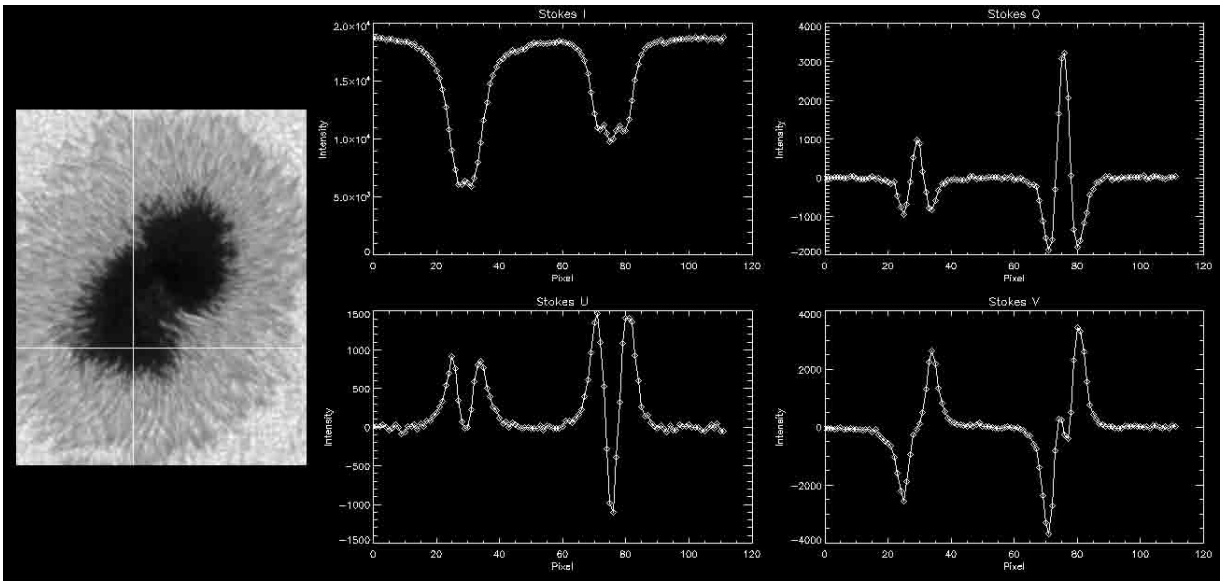

FIG. 3.12 Sample Stokes spectra from AR10923, imaged by the *Hinode* satellite. The wavelength-dispersed Stokes I, Q, U, and V polarization profiles for the pixel at the intersection of the crosshairs in the image on the left. The left-most absorption feature is the Fe I 6301.5Å line, while the other is the 6302.5Å line. Note the clear splitting of the 6302.5Å intensity profile into a multi-lobed profile, and the corresponding reversal at line-center in Stokes V. These are signatures of high field strengths. The 6301.5Å line is just barely showing signs of complete Zeeman splitting.

CHAPTER 4
SPECTROPOLARIMETRIC INVERSIONS WITH A GENETIC ALGORITHM

4.1 Introduction

Recovering solar photospheric magnetic and thermodynamic parameters from information about the wavelength-dependence of spectral line intensity and polarization states is a type of problem known as an *inverse* problem. That is, we cannot make direct measurements of the magnetic field strength at a point on the sun, but given that the presence of the magnetic field modifies (but does not generate) the spectral lines of magnetically-sensitive atomic transitions, we can *infer* the properties that may have given rise to the observed spectrum, from a physics-based model of radiative transfer in the solar interior. To solve this inverse problem, I have employed a version of a class of algorithms collectively known as *genetic algorithms*, to effectively solve the *forward* problem. In this case, we can then compare the spectra generated by the genetic algorithm to the actual observations to determine if the input parameters give a close match. This is done repeatedly, quickly, and efficiently by the genetic algorithm. A detailed description of the solution procedure in the genetic algorithm used for this work is the subject of this chapter.

4.2 Genetic Algorithms

Genetic algorithms (also known as evolutionary algorithms) are a robust set of optimization algorithms designed to computationally exploit the principles of evolutionary biology set forth by Charles Darwin in the 19^{th} century. The driving force of the genetic algorithm does not lie in the algorithm itself, but in a user-defined *fitness function* which the genetic algorithm constantly references in an attempt to find its optimal values. This fitness function can reflect *any* combination of desirable "traits" in a candidate solution, as long as better, more desirable solutions correspond to larger values of the fitness function. The fitness function is a function of N_p parameters, any arbitrary set of which represents a single *individual* in a single *generation*.

The current incarnation of our algorithm utilizes a *decimal encoding*, whereby the parameters are normalized to the range $[0, 1]$ by suitably defining a priori bounds on the parameter space. The set of parameters are then concatenated into a single string of digits in the range $[0, 9]$ by removing the decimal point and sequentially placing these *genes* next to each other. Binary encodings are also possible, and extensively used in the literature, but for our initial purposes a decimal encoding is satisfactory.

The mechanism by which the genetic algorithm maximizes the fitness function is analagous to sexual reproduction in higher organisms at the chromosomal level. A population of N_{pop} individuals is initialized over the entire range of the parameter space by using a random number generator to generate the initial normalized parameters. Each set of parameters (individual) is then rescaled to the proper value and evaluated by the fitness function. Those individuals which correspond to higher fitnesses then must be more desirable than those with lower fitnesses. We wish to keep as much of the "good" genetic information as we can, while dismissing the "bad" genetic information from the lower fitness individuals. This is accomplished by *ranking* each member of the generation according to fitness, and preferentially *selecting* the high-fitness individuals for *breeding*. A so-called roulette-wheel algorithm is employed to do just this. In rank-based selection, parent solutions are chosen to breed by numerically "spinning" a roulette wheel. Higher-ranked individuals are granted larger portions of the roulette-wheel than lower-ranked members, and so when the wheel is spun, the higher-ranked members have more of a chance of breeding, and therefore passing on some of their good genetic material to the daughter solutions.

When two parents are selected to breed, they must exchange some of the their genetic material. This is accomplished by the *crossover* technique, in which a locus on the digit string is selected at random, and only if a probability test returns true. All digits following the selected locus are swapped between the parents, generating two daughter solutions which may contain some of the genetic material from each parent.

Single-Point Crossover

$$...XXXXX|XXXXX... \quad parent1$$

$$...YYYYY|YYYYY... \quad parent2$$

$$\Downarrow$$

$$...XXXXXYYYYY... \quad daughter1$$

$$...YYYYYXXXXX... \quad daughter2$$

A single-point crossover like the one described above is effective, but it lacks one important ability. Imagine an individual whose first and last genes are near-optimal *when simultaneously present*. A successfully executed single-point crossover is clearly unable to propagate that favorable gene combination into subsequent generations, resulting in a net decrease in the fitness of the daughter solutions. This can be avoided by using a two-point crossover, where two loci are selected, and the parents exchange the segment bounded by these loci. In our genetic algorithm, we have chosen to implement single-point crossover approximately half the time, with two-point crossover performed the rest of the time. Which type of crossover to be used is determined by a probability test at runtime.

Two-Point Crossover

$$...XXXX|XXXXXXXX|XXXX... \quad parent1$$

$$...YYYY|YYYYYYYY|YYYY... \quad parent2$$

$$\Downarrow$$

$$...XXXXYYYYYYYYXXXX... \quad daughter1$$

$$...YYYYXXXXXXXXYYYY... \quad daughter2$$

One might be tempted to make the following observation: if there is a finite initial population, then the action of crossover will eventually stall the evolutionary process, since no

new genetic information is being introduced into the gene pool. This is avoided by introducing random mutations into the daughter solutions, in a fashion analagous to biological mutation. A genetic mutation operator is applied to each daughter solution, which progressively scans down the digit string chromosome. If the random number generator returns a value less than p_{mut} (mutation probability) for any gene locus, then that locus is replaced with a random digit in the range $[0, 9]$. Therefore, we have a way of maintaining genetic diversity, even at late times in the genetic algorithm run.

It is well-known that traditional decimal encoding schemes may encounter the so-called *Hamming Wall*, whereby the population gets stuck when a single-point mutation cannot produce an evolutionarily favorable change to the daughter digit string. This is because a single-point mutation causes rather large jumps in parameter space, when the purpose of mutation in the first place is to produce only *slightly* different sets of parameters than one would have had in the absence of mutation. In effect, this means an individual containing a gene that is relatively close to its optimum value might be "rattled" out of the vicinity of the optimum value. This is compensated for by introducing the concept of *creep mutation*. Creep mutation acts as sort of a short-range hill-climber, in that it produces only small changes to the daughter digit string. If a particular locus is targeted for mutation, the creep mutation operator does not replace that locus with an arbitrary single-digit integer, but rather increments the locus by 1, or decrements it by 1, with equal probabilities. If the locus targeted for incrementation by the creep mutation operator is a 9, it is replaced with a 0, and the locus immediately to the left is also incremented by 1. This process is repeated if the locus to the left is also a 9, ad nauseum. The reverse is true for decrementation of a 0. As is obvious, creep mutation also goes by the name "carry-the-one" mutation. It would be unsuitable to adopt creep mutation as the standard, as it produces only small changes in the model parameters. The efficiency of the genetic algorithm is dependent on large-scale changes that lead to greater exploration of the parameter space, such as those produced by single-point mutation. Therefore, *both* types of mutation are coded into our genetic algorithm, with equal probability of occurrence when the *mutate* subroutine

is called. Shown below are the mechanisms of single-point and creep mutation, with the targeted locus shown between the vertical slashes.

Single-Point Mutation	**Creep Mutation**
...678|9|634...	..3499|9|6376...
$\Downarrow$	$\Downarrow$
...678|1|634...	...3500|0|6376...

The mutation rate is dynamically updated to reflect the distribution of similar individuals in the population, so that when the population consists of large numbers of similar individuals, the mutation rate is increased to help spur the introduction of more fresh genetic material. In fitness-based adjustment, the controlling quantity is the normalized difference between the fitnesses of the best and median individual of the previous generation. If this difference is small, the population contains many similar individuals, and hence the mutation rate is increased accordingly. If the difference is large, then the population contains many *distinct* individuals, and the mutation rate is decreased to prevent too many spurious random mutations that might destroy valuable genetic material.

There is another widely used method of adjusting the mutation rate which is based on the normalized metric distance between the best and median solutions. The fitness-based method works well on landscapes with large contrast between areas of high and low fitness, but what if we're searching over a largely "flat" surface with small and diffuse peaks and troughs? Since the individuals on this roughly flat surface would have nearly equal fitnesses, a fitness-based adjustment of the mutation rate is unlikely to introduce sufficient genetic diversity to prevent premature convergence. Distance-based adjustment utilizes the actual difference in decoded parameter space between the best and median individual to remedy this situation. This metric distance is defined as

$$d = \frac{1}{N_p}\sqrt{\sum_{i=0}^{N_p}\left(x_i^{best} - x_i^{median}\right)^2}. \tag{4.1}$$

If this value is small, many individuals are clumped around the best solution of that generation, and the mutation rate is increased in an attempt to promote increased exploration *away* from the cluster, and thereby avoid premature convergence. If the metric is large, then the mutation rate is decreased, for the same reason as in the fitness-based method.

Once the crossover and mutation stages have been completed, the daughter digit strings are decoded into the normalized set of parameters by a process that is essentially the inverse of the encoding mechanism. These decoded parameters are then rescaled to the proper value by utilizing the parameter space bounds, and input to the fitness function to assess their fitness. The selection and breeding process occurs $N_{pop}/2$ times, and when N_{pop} daughter solutions have been generated, this new *generation* of potential solutions replaces the old (parent) generation *en masse*. The new generation is subsequently ranked in terms of fitness, and the breeding process begins anew. This process of information exchange and replacement results in arbitrarily large jumps across parameter space, relative to the coordinates of the parents, and so we see the power of genetic algorithms; as each new generation replaces the old one, we can explore N_{pop} more combinations of parameters in hopes of locating a particular combination that maximizes the fitness function. Shown below is a high-level pseudocode outlining a typical genetic algorithm solution procedure.

$t = 0$

initialize $P(0) = \{\vec{a}_1(0), \vec{a}_2(0), \ldots, \vec{a}_\mu(0)\} \in \{0,1\}^\mu$

evaluate $P(0) : \{\Phi(\vec{a}_1(0)), \Phi(\vec{a}_2(0)), \ldots, \Phi(\vec{a}_\mu(0))\}$ **where** $\Phi(\vec{a}_k(0)) = \Delta(f(\Gamma(\vec{a}_k(0))), P(0))$

while$(\epsilon(P(t)) \neq$ **true**$)$ **do**

$\quad$ *select* : $\vec{a}'_k = s_{\{p_s\}}(P(t))$ **where** $p_s(\vec{a}'_k(t)) = \Phi(\vec{a}'_k(t)) / \sum_{j=1}^{\mu} \Phi(\vec{a}'_j(t))$

$\quad$ *cross* : $\vec{a}'_k = r_{\{p_c, n\}}(P(t)) \; \forall k \in \{1, \ldots, \mu\}$

$\quad$ *mutate* : $\vec{a}''_k(t) = m_{\{p_m\}}(\vec{a}'_k(t)) \; \forall k \in \{1, \ldots, \mu\}$

$\quad$ *evaluate* : $P'(t) = \{\vec{a}''_1(t), \ldots, \vec{a}''_\mu(t))\}$ **where** $\Phi(\vec{a}''_k(t)) = \Delta(f(\Gamma(\vec{a}''_k(t))), P(t-\theta))$

$\quad$ *replace* : $P(t+1) \leftarrow \eta(P'(t), P(t))$

$\quad t = t + 1$

end do

The notation is as follows: there are μ individuals per generation, collectively denoted as the population $P(t)$. Γ is a decoding function which acts on the genotypes $\vec{a}_k(t)$. f and Δ are the fitness and scaling functions, respectively, that are applied to the decoded genotypes. A selection operator, $s_{\{p_s\}}$, has an application probability of p_s. An n-point crossover operator, $r_{\{p_c,n\}}$, has an application probability of p_c. A mutation operator, $m_{\{p_m\}}$, has an application probability of p_m. The generational replacement operator, η, depends on both the offspring and parent populations, and finally, ϵ is some user-defined termination criterion.

4.2.1 The Schema Theorem: Why Genetic Algorithms Work

The first genetic algorithms ever developed used a *binary* encoding of the phenotype, as opposed to the real or decimal encoding used in the present work. This is partly because the base-2 representation of floating-point numbers are easy to alter (have only two possibilities) and so the first genetic operators would only have to operate in a 2-member alphabet (0,1), and hence would be simpler than the 10-member alphabet used in the real-coded representation for this work. A binary *bitstring* $\vec{a} = (a_1, a_2, ..., a_l)$ has elements $a_i = 0, 1 \ \ \forall i$ and resides in l-dimensional binary space. The bitstring is partitioned into segments of equal length l_x, each segment encoding the corresponding real variable.

For *real* function optimization, the so-called *injective decoding functions* (Bäck (1996)) are used to process a binary string representation into a real number representation by decoding the substring into the corresponding integer in the range $[0, 2^{l_x} - 1]$, and mapping the integer into the real interval $[u_i, v_i]$. Formally, the i^{th} gene is decoded as:

$$\Gamma^i(a_{i1}, a_{i2}, ..., a_{il_x}) = u_i + \frac{v_i - u_i}{2^{l_x} - 1} \left[\sum_{j=0}^{l_x - 1} a_{i(l_x - j)} \times 2^j \right]. \tag{4.2}$$

This binary representation therefore represents a kind of regularly-gridded scale which is mapped into the continuum of real numbers where the fitness function is well-defined. The resolution of this scale is then defined by the number, l_x, of binary bits chosen to represent

a single gene in the full genotype, and is given as:

$$\Delta x_i = \frac{v_i - u_i}{2^{l_x} - 1}, \tag{4.3}$$

which confirms that the accuracy of the result may be increased by choosing a large number of bits to represent each gene. A binary encoded genotype then has the form

$$\overbrace{\ldots \underbrace{1010100101}_{l_x \text{ bits}} \underbrace{0001010100}_{l_x \text{ bits}} \underbrace{0001011101}_{l_x \text{ bits}} \underbrace{1010001010}_{l_x \text{ bits}} \ldots}^{l \text{ bits}}. \tag{4.4}$$

The binary selection mechanisms are still fitness-based, and hence are independent of the actual encoding used. The binary crossover operator is functionally identical to the real-coded crossover operator; both select a locus or loci and swap finite segments of their respective encoded genotypes, the only difference being the actual content being swapped. Finally, the binary mutation operator is dissimilar from the real-coded operated outlined above only in that if a binary locus is targeted for mutation, it has only one other option (0 or 1), instead of a range of mutation options (0 through 9). The binary representation is typically slower than real-coded representation, due to the extra computational overhead required by the injective decoding functions. However, due to their simplicity, the binary genotypes allow a somewhat rigorous treatment of how and why this artificial evolution produces quality results to difficult problems. This theoretical framework of a genetic algorithm has come to be known as *Holland's Schema Theorem* (Holland (1975)). The schema theorem states how "building blocks" consisting of (loosely) connected groups of 1's and 0's are preferentially propagated into later and later stages of the artificial evolution. These building blocks can have a maximum length equal to the size of the entire encoded genotype, and a minimum size of 1 (a single binary bit), and act as a similarity template representing multiple distinct but piecewise-similar genotypes. For our purposes, we define a *schema*, S, to be any binary bit substring of a form such as:

$$S \equiv \{*10*1\}, \tag{4.5}$$

where the "*" is a wildcard for either 0 or 1, and the other loci are referred to as *fixed* positions. Any genotype containing a substring matching the schema S is said to encode

for, or belong to, S. That is, the binary substrings in the set {11001,11011,01011,01001} all match the example schema S shown above. These schema are the actual "building blocks" mentioned previously. The action of the binary crossover and mutation operators serves to disseminate schema of various length and order throughout the population, rewarding genotypes that encode for beneficial schema by allowing them to breed more frequently via an appropriate selection mechanism, fueling a positive feedback cycle that ultimately leads one (or a few) specific genotypes to completely dominate the population. Since the exchange and transmission of good genetic material is carried out by the crossover and mutation operators (and indirectly by the selection operator), an analysis of the schema theorem should start with them.

The goal of a genetic algorithm is to continually propagate good genetic material into later stages of the algorithm, in the form of coherent schema that may be created or destroyed by the crossover and mutation operators. The schema theorem attempts to assign a lower bound to the rate(s) at which relatively good schema are transmitted to future generations. To do this, we must ascertain the probabilities that a binary string containing some arbitrary schema S is selected for breeding, and still contains that schema after the crossover and mutation operators have been applied. The particular selection mechanism also contributes to the expected number of instances of a particular schema in subsequent generations.

The contribution from a selection operator utilizing fitness-proportionate ranking is given by:

$$m(S, t+1) = m(S, t) P_{s,select}, \tag{4.6}$$

where $P_{s,select}$ is the probability that a given individual encoding for schema S is selected to breed, and is given by:

$$P_{s,select} = \frac{f(S,t)}{\overline{f(t)}} \tag{4.7}$$

for fitness-proportionate selection. This gives the expected number of instances of schema S in the next generation, $m(S, t+1)$, based solely on the differential reproductive success

of the schema instances in the current generation. $f(S,t)$ is the average fitness of all strings containing the schema S, and $\overline{f(t)}$ is the average fitness of the whole population. Since the selection mechanism is fitness-proportionate, higher-fitness schema tend to be selected more frequently, resulting in their preferential transmission into future generations.

The probability that the schema survives the action of the binary crossover operator is determined by examining the *disruption* probability. A schema will be disrupted *only* if the crossover operator selects a crossover locus that corresponds to one of the *fixed* positions within the schema. The fraction of the total bitstring length taken up by the fixed positions of schema S is given by $\delta(S)/(l-1)$, where $\delta(S)$ is the *defining length* of schema S. The defining length is the number of bits between the first and last fixed positions in the schema. Therefore, the probability that any of the fixed positions are destroyed by the crossover operator is

$$p_c \frac{\delta(S)}{l-1}, \tag{4.8}$$

where p_c is the crossover probability. This gives a complimentary survival probability of

$$P_{s,cross} = 1 - p_c \frac{\delta(S)}{l-1}. \tag{4.9}$$

The probability that the schema is disrupted by the mutation operator is the joint probability that any one single *fixed* position within the schema is mutated. Therefore, the probability that a single fixed position survives the mutation operator is $1-p_m$, where p_m is the mutation probability. Besides its defining length, a schema is also characterized by its *order*, $o(S)$, which is the number of fixed positions within the schema. For the entire schema to survive the action of the mutation operator, all $o(S)$ fixed positions must survive, giving a total survival probability of

$$P_{s,mutate} = (1-p_m)^{o(S)}, \tag{4.10}$$

since all loci have equal probability of being mutated.

The expected number of instances of schema S in the next generation is then dependent on the current number of instances as well as the probability that a string containing S is

selected to breed, *and* S survives both crossover *and* mutation:

$$m(S, t+1) = m(S, t) P_{s,select} P_{s,cross} P_{s,mutate}, \tag{4.11}$$

which ultimately yields

$$m(S, t+1) \geq m(S, t) \frac{f(S,t)}{\overline{f(t)}} \left[1 - p_c \frac{\delta(S)}{l-1}\right] [1 - p_m]^{o(S)}. \tag{4.12}$$

Traditionally, this is the mathematical statement of Holland's Schema Theorem. The inequality refers to the fact that a schema may not only be propagated into future generations intact, but also may be *constructed* from bits and pieces of other lower-order schema. It should be noted, however, that since the genetic algorithm is a stochastic mechanism, the schema theorem does not serve as a rigorous prediction of the distribution of schema from one generation to the next, but rather is more of a description of the *average* propagation of schema to the next generation.

Another apparent consequence of the schema theorem is that instances of schema with below-average fitness then monotonically decreases, to allow the high-fitness schema to represent a greater fraction of the total population. Schema with low defining number and low order are also preferentially propagated into future generations which helps illustrate the principle of "building blocks" in genetic algorithms. Relatively small similarity templates act as the basis for myriad combinations and rearrangements of the multiple genotypes they represent, ultimately leading to the generation of relatively large, high-fitness schema.

For schema with low defining number, $\delta(S)/(l-1) \ll 1$, which leads to the crossover survival probability of

$$P_{s,cross} = 1 - p_c \frac{\delta(S)}{l-1} \approx 1, \tag{4.13}$$

which implies that schema with short distances between their first and last fixed positions easily survive the schema-disruptive action of the crossover operator. This is intuitive, since shorter defining lengths logically yield a higher probability that the locus selected for single-point crossover does not correspond with a fixed position of a schema.

The propagation of low-order schema can be analyzed by performing a binomial expansion on the mutation term in the schema theorem

$$[1 - p_m]^{o(S)} \approx 1 - o(S)p_m \approx 1, \tag{4.14}$$

since the mutation rate is typically required to be very small (in this work, $p_m = 0.005$). This implies that low-order schema are easily propagated by the mutation operator, which again is intuitive; schema with low numbers of fixed positions (e.g., {***1*0}) have a smaller chance of being disrupted by any mutation than a schema of the form {1101*0}.

From the formal statement of the schema theorem, it can be conclusively argued that everything on the RHS of equation 4.12 that is multiplied by $m(S,t)$ must be greater than 1 in order for the above-average schema H to receive ever-increasing representation within the gene pool

$$\gamma = \frac{f(S,t)}{\overline{f(t)}} \left[1 - p_c \frac{\delta(S)}{l-1}\right] [1 - p_m]^{o(S)} \geq 1. \tag{4.15}$$

To see this, assume that some low-order, low-defining-length schema S remains above-average by an amount proportional to the average fitness of the population, i.e.,

$$f(S,t) = \overline{f(t)} + C\overline{f(t)} = (1 + C)\overline{f(t)}. \tag{4.16}$$

Then the number of "trials" allocated to the schema S in the first generation is given by the schema theorem as:

$$m(S,1) = m(S,0)(1 + C), \tag{4.17}$$

which by recursion yields for some arbitrary generation:

$$m(S,t) = m(S,0)(1 + C)^t, \tag{4.18}$$

which shows that above-average schema receive exponentially-increasing attention in the global gene pool. The converse is also true; the transformation $C \longrightarrow -C$ shows that below-average schema are represented in an exponentially-decreasing fashion.

Therefore, we finally see that it is a probabilistic survival of members of the set of individuals containing higher-than-average schema that allows the binary-encoded SGA to

effectively and efficiently search over the entire parameter space. A proper schema theorem for real-coded genetic algorithms (used in this work) is slightly more complicated (see e.g. Zhi-Hua and Jian-Chao (2002)), due to the extended 10-alphabet used in the real encoding, but contains the same basic principles of probabilistic destruction and construction of similarity templates. In fact, the precise definition of a schema ($\delta(S)$ and $o(S)$) is independent of the number of symbols used to represent it, since it is really just a similarity template. Therefore, the only complication to be added to the schema theorem is how the maximum number of schema that can be represented changes with the 10-alphabet. The probability that a particular schema of order $o(S)$ matches *at least* one genotype in a random population of N genotypes in a k-ary alphabet is

$$P(S) = 1 - \left[1 - \left(\frac{1}{k}\right)^{o(S)}\right]^{N}. \tag{4.19}$$

The total number of all schema of order $o(S)$ on a string of length l in a k-ary alphabet is

$$N[o(S)] = \binom{l}{o(S)} k^{o(S)}, \tag{4.20}$$

which implies that the total expected number of instances of schema S in the random population is

$$< m(S, t=0) >= N[o(S)]P(S) = \binom{l}{o(S)} k^{o(S)} \left(1 - \left[1 - \left(\frac{1}{k}\right)^{o(S)}\right]^{N}\right) \tag{4.21}$$

The above-average genotypes containing this schema will again tend to propagate more copies of their schema into subsequent generations, with the below-average individuals still propagating their genetic material, but in a monotonically-decreasing fashion. Goldberg (1989) showed that the number of representable schema on a string of length l is optimal when $k = 2$, i.e., when a binary alphabet is used for the encoding. It is generally accepted that binary encodings can process many more schema than higher representations, although the choice of a real-coded genetic algorithm for this work was motivated by the computational ease of implementing real-coded operators on a domain of real numbers, as is frequently the case in practical problems.

This section has summed up the mechanism(s) by which the SGA evolves a population of strings toward an optimal result. The next section focuses on some of the additions/improvements I have made to the basic SGA procedure.

4.2.2 More Sophisticated GA Strategies

In the course of my work, I have explored many different incarnations of genetic algorithms in hopes of finding/designing a more effective, more efficient evolutionary solution to the Stokes inversion problem. This section gives a list of various genetic paradigms that I have incorporated into the SGA, along with a brief synopsis of their functions.

Adaptive Creep Mutation

The normal creep mutation operator (Charbonneau (2002)) works in concert with the traditional single-point mutation operator to help produce both large mutations (which produce large jumps across the parameter space) and small mutations (which act as a hillclimber in local regions). The original algorithm used a 50/50 mix of the two operators. I have introduced an adaptive scheme, based on fitness and distance in parameter space, which increases/decreases the creep mutation probability based on a few measures of the degree of "convergence" in the population:

$$\Delta f = f_{best}(t) - f_{med}(t) \tag{4.22}$$

$$\Delta d = \|\vec{\nu}_{best} - \vec{\nu}_{med}\|. \tag{4.23}$$

Both of these measures are incorporated into a single probability modification expression that determines when an increase/decrease in the creep mutation rate would be advantageous. The following four situations may be envisioned.

- Δf, Δd small:

  In this situation, the population is tightly clustered both in fitness value and

distance/separation in parameter space. The mutation rate should therefore be increased, to promote more local exploration around the population.

- Δf small, Δd large:

  In this situation, the fitness contrast is small over a large region of the parameter space. We should decrease the creep mutation rate, so that normal mutations will have a higher probability of "jumping" around this "flat" region.

- Δf large, Δd small:

  In this situation, the presence of many individuals has revealed strong gradients in the fitness function. We should increase the creep mutation rate in hopes of quickly climbing up these gradients.

- Δf, Δd large:

  In this situation, nothing justifiable can be done.

This mode of operation showed a lot of promise when used in conjunction with niching techniques (below) so that subpopulations could have different mutation rates. What determines whether a particular value of Δf or Δd is small or large? A good candidate would be these same values evaluated on the initial (random) population. It should also be noted that the evolution of these values can help to estimate the "convergence velocity" of the algorithm.

Arithmetic Crossover

The arithmetic crossover operator is suitable only for genetic algorithms using a real-coded representation. Instead of swapping segments of the parents' genotypes, offspring genes are calculated as a weighted combination of the corresponding genes in the parents

$$G_{i,off} = \epsilon G_{i,p1} + (1 - \epsilon) G_{i,p2}, \tag{4.24}$$

where G_i represents the i^{th} real-valued gene (parameter) on the chromosome, and ϵ is a constant. This constant can be arbitrarily set by the user, but I also offer the following

adaptive scheme, whereby the mixing constant ϵ can be adjusted to take advantage of the higher-fitness parent's good genes. Let

$$\epsilon = \frac{f_1}{f_1 + f_2}, \tag{4.25}$$

where f_1 and f_2 are the fitnesses of the first and second parent, respectively. Then the offspring genotype is closer to that of the parent with the highest fitness, and the degree of difference is set independently by *each crossover operation*. This ensures that good evolutionary information is still propagated to subsequent generations, but also helps to introduce innovation into the population. By innovation I refer to information that was not present in either parent, but is present in one or more offspring.

Binary Encoding(s)

The archetypal genetic algorithm uses a binary (bitstring) encoding of the real-valued parameters to be optimized. The use of a binary encoding is not typically required, although it is easiest to analyze the average convergence properties of a binary-coded genetic algorithm (see previous section on the Schema Theorem). There is widespread debate among the genetic algorithm community regarding the effectiveness of real-coding versus binary-coding, but it seems that which encoding is "better" is highly problem-dependent. Therefore, I have included an option for the user to specify which particular encoding to use. The details of the encoding/decoding can be found in the previous section on the Schema Theorem.

The standard binary encoding does have one drawback worth mentioning here. The role of mutation is to introduce new genetic information into the population by making *small* random changes directly to a genotype. However, the change of a single binary bit (in the right place) can cause huge changes in the decoded parameter. This is ultimately counterproductive and is likely to destroy good genetic information. The cure to this ailment is to use a *gray* binary code. An equivalent gray binary string can be easily constructed from the standard binary string with an addition step in the encoding/decoding processes, and it has the following extremely convenient property:

- Two gray-coded binary strings that differ by only a single bit (i.e., have a Hamming distance of 1) represent two *consecutive* positive integers.

Therefore, binary mutations on a gray-coded string will produce only small changes to the decoded parameters, and this is exactly what we want in a binary mutation operator. For completeness, the injective encoding and decoding functions for a gray binary code are given here as:

$$\Xi^i(g_{i1}, g_{i2}, ..., g_{il_x}) = u_i + \frac{v_i - u_i}{2^{l_x} - 1} \left(\sum_{j=0}^{l_x - 1} \left(\bigoplus_{k=1}^{l_x - j} g_{ik} \right) \times 2^j \right). \tag{4.26}$$

Fitness Scaling

Fitness scaling works by increasing the contrast between low-fitness and high-fitness regions of the parameter space, without changing the locations of these regions, thereby altering the selection pressure and forcing more individuals toward the nearest peak in the parameter space. This strategy is implemented by exponentially scaling (Deb and Goldberg (1989)) the raw fitness of an individual by some factor, β:

$$f_{scaled} = (f_{raw})^{\beta}, \tag{4.27}$$

where $\beta > 1$. The effect of fitness scaling on a sample fitness-function landscape can be seen in Figure 4.1. In my algorithm, β can be constant throughout the entire run, or may be dynamically adjusted away from an initial value set by the user at a rate also set by the user. The best value of β to use varies wildly from application to application, although it is not uncommon for $\beta > 10$ to be used (e.g., Darwen and Yao (1995)). Scaling can also be initialized at any generation; that is, one can start the fitness scaling at any generation. In fact, I have found that this particular strategy can be very effective at quickly driving individuals toward their respective nearest optimum. However, the caveat here is that it can drive them to the optima *too* quickly, causing loss of diversity in the population which then fails to find other optima. This is one of the major drawbacks to fitness scaling.

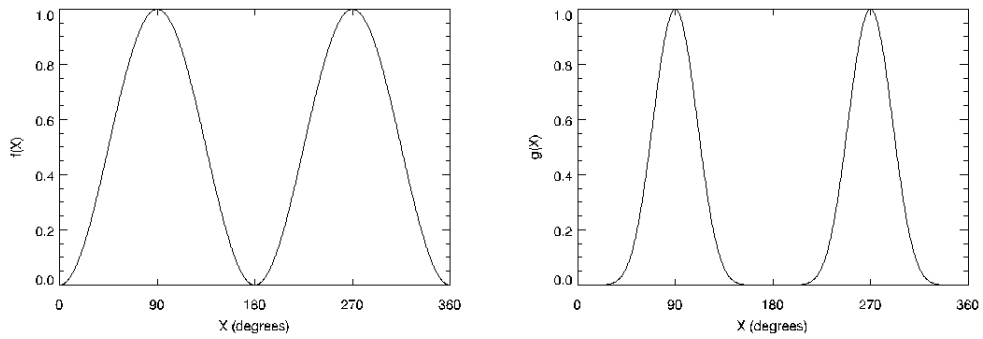

FIG. 4.1 The effect of scaling on a sample fitness function. *(Left):* A sample fitness function, $f(x) = sin^2(x)$, which has maxima at $x^* = \pi/2,\ 3\pi/2$. *(Right):* A scaled fitness function, $g(x) = f^4(x)$. The positions of the maxima have not been altered, but the contrast between high-fitness and low-fitness regions has been dramatically increased.

Hillclimber

The hillclimber operator is a modified version of the creep mutation operator, used only for a real-coded genetic algorithm. This operator is typically called only once, at the end of an evolutionary run, to operate on the best individual found by the evolutionary search. This is, in part, due to the computational expense of the operator. The hillclimber operator systematically increments/decrements the real-valued genotype, testing for fitness improvement along the way. The operator terminates when no further increment or decrement can produce a measurable change in the fitness of the individual. At its core, this operator acts like a gradient search algorithm, but without the need for evaluating those nasty derivatives.

This operator is the exploitative counterpart to the exploratory search performed by the genetic algorithm. That is, during the course of a GA run, large parts of the parameter space are explored. In contrast, during the course of the hillclimber's operation, the information processed by the genetic algorithm is exploited to fine-tune the final solution. It should be noted that this method is very efficient at producing the *exact* optimal solution from a slightly suboptimal individual for relatively low parameter space dimensions. Practical, real-world problems typically have high dimensions and, depending on the complexity of the fitness function, it may be impractical to utilize this operator, since it requires a fairly

high number of function evaluations.

Segmented Crossover

The original SGA on which this work is based utilized a mixture of single- and two-point crossover. I have since generalized this to n-point crossover, whereby n crossover loci are selected and the parents exchange alternating segments. This is expected to aid the distribution of good genetic information, especially for high dimensional problems with large numbers of genes. It is expected to be more useful in a binary-coded genetic algorithm where a single parameter is represented by a binary string, typically of length $l > 15$, such that the full binary genotype is a very long string of bits. In a real-coded algorithm, the full genotype is much shorter, so that the presence of many crossover loci may be more destructive to good genetic information than in the binary case. In fact, through various simple test cases, I have found that (on average) the largest "safe" number of crossover loci that can be used on a real fitness function of N parameters (i.e., an N-dimensional function) is $N-1$. With any more loci, the segmented crossover operator shuffles too many *short* blocks of information between the two parents, effectively randomizing the daughter solutions.

Multi-Point (Segmented) Crossover

$$XX|XXX|XXXXX|XXX|XX \qquad \textit{parent } 1$$

$$YY|YYY|YYYYY|YYY|YY \qquad \textit{parent } 2$$

$$\Downarrow$$

$$XXYYYXXXXXYYYXX \qquad \textit{daughter } 1$$

$$YYXXXYYYYYXXXYY \qquad \textit{daughter } 2$$

Penalty Functions for Constraint Handling

Many parameter optimization problems (particularly in the field of engineering) must be solved subject to constraints which must be satisfied by the solution. I first considered

this approach when attempting to fit a three-component model of solar surface magnetic fields to observational data. The particular constraint I was dealing with was that the sum total of the three components must occupy the entire pixel field-of-view, and I could not envision a scheme that would let me independently determine all three components while still retaining and orthogonal parameter space. Mathematically, constraints are formulated as follows:

$$g(\vec{p}) = \lambda, \tag{4.28}$$

or equivalently

$$g(\vec{p}) - \lambda = 0, \tag{4.29}$$

where g is some function of the parameters. For an individual, $\vec{p}$, the degree of violation of this constraint is given by the LHS of the above equation. That is, an individual that completely satisfies the constraint returns 0 for the RHS of the above equation. There are two main paradigms for dealing with individuals that violate their constraints:

- destroying the violating individual, and replacing it with another, or
- penalizing the individual, but not entirely eliminating it from the population.

The first method drastically effects the runtime of the algorithm, since randomly producing new individuals is still unlikely to produce one that satisfies the constraint (Carlson and Shonkwiler (1998)). Therefore, the second method (penalization) is typically used, since it incurs no extra computational overhead. Instead, the violating individual has its fitness reduced by an amount that is a function of its degree of constraint violation. This way, the genetic information possessed by the individual is not destroyed, but instead is *probabilistically* removed from the population by the aforementioned selection techniques. Again, the form and amount of penalization is very problem-dependent, and a brief synopsis of a few major strategies can be found in Michalewicz (1995), and the references therein.

Tournament Selection Techniques

Tournament selection is generally the other front-runner when it comes to selection methods, the other being proportionate (also known as rank-based or roulette-wheel) selection. In tournament selection, to determine which individuals will have crossover applied to them, a random sample of q individuals is selected from the population and compared. The individual with the highest fitness is selected to be a parent. This process is repeated to select another parent. Common values of q are 2 (binary tournament) and 3 (tertiary tournament), but in practice the choice of q depends on the fitness function landscape as well as the population size. This type of selection is easily implemented in the algorithm.

A somewhat more sophisticated strategy has been dubbed *correlative tournament selection*, and involves a measure of the fitness-distance correlation between two selected parents. It is based on the idea that the genetic search will be more effective if offspring are only *slightly* different from their parent genotypes, therefore aiding in more detailed local exploration. Matsui (1999) implemented an altered version of the "seduce" operator found in Ronald (1995) that used the Hamming distance between two binary strings to measure the correlation of two parents, and found enhanced performance on several types of GA-hard problems, with very little extra computation that is not already present in more traditional selection operators. This type of reproduction strategy could perform very well on multimodal optimization problems, where many optima exist, and where the recombination of individuals that are *too* disparate severely degrades the genetic search.

Uniform Crossover

Uniform crossover has its roots in the original binary-coded genetic algorithm. In the binary representation, an offspring allele (a single bit in the offspring genotype) is produced by using the corresponding allele from either one parent or the other. This is determined by a probability test, usually with equal probabilities for each parent. That is, an offspring allele has a 50% chance of originating from the first parent. A fitness-proportionate scheme here would be very similar to the arithmetic crossover operator for real-coded genetic algorithms.

That is, the i^{th} bit in the offspring genotype has a $p_i = f_1/(f_1 + f_2)$ probability of coming from parent 1, and a $1 - p_i$ probability of originating from parent 2.

This can also be applied to a real-coded genetic algorithm. The method is exactly the same, only with a different alphabet: the single-digit integers [0,9]. However, it is as yet untested on a real-coded genetic algorithm, although I am of the opinion that it would be more destructive than helpful for a k-ary alphabet with $k > 2$, for the same reason that highly-segmented crossovers are very destructive to the parent genotypes.

Judgment Day

A major problem in genetic algorithms is the prevention of premature convergence. Because of the fitness-proportionate selection mechanism, a very high-fitness solution (relative to the rest of the population) will eventually dominate the population, and diversity will be lost, stalling the evolutionary pressure by which a genetic algorithm is able to search the parameter space. While this is desirable if the aforementioned high-fitness solution is near the global optimum, in practice there is no way to determine this. Therefore, a suboptimal solution may eventually dominate the population, and the global optimum may not be located. To prevent this premature convergence and to help the genetic algorithm continually access other regions of the parameter space, I have included a genetic operator which I refer to, somewhat menacingly, as the judgment day operator, which is a variant on the "selectively destructive restart" operator or Maresky et al. (1995). If the population has converged to a large degree, measured by comparison of the average fitness of the entire population to the best current solution, a fraction of the population is destroyed and replaced by a new subpopulation, initialized randomly, as in the initial generation of the genetic algorithm. This injection of new genetic material into the mating pool stimulates exploration away from the region of convergence, *if it is desirable to do so*. That is, if the reinitialization finds a point in the parameter space with fitness comparable to, or greater than, that of the currently converged population, the fitness-proportionate selection operator will then start the exploration of this new region.

4.2.3 Subpopulations and Niches

This final section on improvements to the basic SGA structure covers a vast area that focuses on the ability of a genetic algorithm to maintain and evolve *separate* subpopulations of individuals, hereafter referred to as *niches*. The motivation for this approach stems from the fact that many optimization problems have slightly sub-optimal solutions that are "good enough." A good example is circuit design. A circuit required to have a certain property (resistance, capacitance, impedance) can typically be constructed any number of ways, but the best configuration will generally use the least number of circuit elements, thus cutting down production cost. For a complex circuit design problem, it may be impractical or expensive to construct the fully optimal circuit, but there may be a slightly sub-optimal design that is equivalent (to some specified tolerance), and easier to construct.

The action of genetic drift in an SGA causes an initially diverse population to converge to only one or a few high-fitness genotypes. Unmodified, the SGA is therefore unsuitable for finding a number of "good enough" solutions to optimization problems. To effect such a modification, we hark back to evolutionary biology for the concept of *fitness sharing*.

In nature, many different species survive in the same environment. Because of this co-existence, there exists a competition for any present natural resources (food, water, shelter, mates), such that all these resources must be *shared* among the different species in order for a stable ecosystem to emerge and thrive. This concept of sharing is thus extended into a computational paradigm whereby an individual's true fitness is dependent not only on its raw fitness, but on its closest neighbors and their fitnesses as well. In this way, a region of parameter space that has been explored by the evolution of a number of individuals can be maintained by a small number of individuals while other, perhaps more promising, regions can be explored. It will be shown below by a detailed analysis how this method of *fitness sharing* allows multiple distinct subpopulations to evolve in parallel, allowing the identification of multiple promising, high-fitness regions. The analysis will concentrate on two strategies: standard fitness sharing and a considerably more sophisticated strategy named

Dynamic Niche Clustering (Gan and Warwick (2000)).

Fitness Sharing

The workhorse of many subpopulation-based GA strategies is the concept of *fitness sharing* (Sareni and Krähenbühl [1998], Beasley et al. [1993], Dick [2005], Mahfoud [1994a,b and 1995], Shir and Bäck [2006], Miller and Shaw [1995]), whereby a candidate's raw fitness is reduced by an amount that depends on the presence and distribution of other candidates within some specified (static or dynamic) distance in parameter space. Let f_i be the raw fitness of the i^{th} candidate, then the shared fitness may be expressed as:

$$f_i^* = \frac{f_i}{n_i}, \tag{4.30}$$

where n_i is the *niche-count* of the i^{th} individual. The niche-count is actually a sum of *sharing functions* for each candidate within a distance σ_{sh} of the i^{th} individual, and is given by:

$$n_i = \sum_{j=1}^{N} Sh(d_{i,j}) = \sum_{j=1}^{N} \left[1 - \left(\frac{d_{i,j}}{\sigma_{sh}}\right)^{\alpha}\right] \qquad d_{i,j} \leq \sigma_{sh}, \tag{4.31}$$

where the sum runs over all candidates within a metric distance σ_{sh} of the i^{th} individual, $d_{i,j}$ is the metric distance between the i^{th} and j^{th} candidate

$$d_{i,j} = \sqrt{\sum_{k=1}^{N_{dim}} (x_{k,i} - x_{k,j})^2}, \tag{4.32}$$

and α controls the shape of the sharing function. Note that other distance metrics may be used (L-norms), but the effect on algorithm performance of using these alternate metrics is not investigated here. Typically, α is set to 1, giving the so-called "triangular sharing function" (see Figure 4.2). The geometry of the sharing function as α is varied is also shown in Figure 4.2 for a niche centered at $X = 0.5$, with a radius of 0.2.

The purpose of the sharing function is to reduce the contribution to the shared fitness from candidates that lie further away from the clustered group to which it belongs. This can be seen in the $d_{i,j}$ term above; if $d_{i,j}$ is small, the sharing function is approximately 1, while for $d_{i,j} = \sigma_{sh}$, the sharing function is 0. Intermediate values of the distance metric

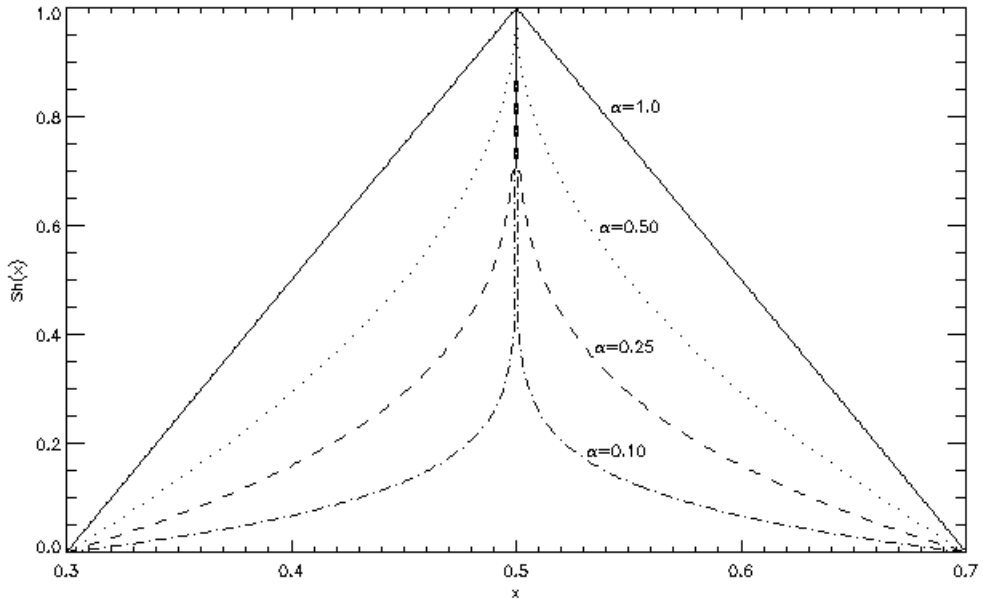

FIG. 4.2 The effect of the sharing parameter, α, on the geometry of the sharing function. The sharing function is evaluated for each individual within a niche-radius of the center of the niche, and the individual contributions are summed and divided into the raw fitnesses. The more individuals that lie close to the center of the niche, the larger the summed contributions, and hence the probability that that niche will be further explored by selection, crossover, and mutation decreases.

lead to sharing functions that lie between these two extremes. When calculating the niche-count for the i^{th} individual, the sum includes the i^{th} individual itself, guaranteeing that the niche-count function is always greater than 1, leading to a reduced shared fitness when compared to the raw fitness. Therefore, an SGA that incorporates basic fitness sharing will start to populate optima in the parameter space through the genetic drift mechanism, but when enough of the population has clustered around some area, the sharing mechanism starts to reduce the chances that the neighborhood around that populated optimum will be further explored by reducing the fitnesses, and hence the selection probabilities of the candidates will subsequently go down.

It should be noted that the fitness sharing *does not* automatically retain the best candidate within a niche. To accomplish this, the NGA turns to the concept of *elitism* used in the SGA. After enough candidates have been bred to produce the next generation, the best

member of each niche is compared to the newly-generated pool of candidates. If the best member of each niche has a higher fitness than the worst member of the newly-generated population, that member is copied unaltered into the next generation, replacing the worst of the new population. This way, once an optimum has been found, it will never be lost due to the stochastic nature of the genetic operators. This also has the effect of weeding out very poor (rogue) candidates in the new population that were generated by some detrimental evolutionary step backwards, for example a random mutation that places a candidate in a very low-fitness region of the parameter space.

Once every candidate is assigned a shared fitness value, the population is ranked from lowest to highest *shared* fitness. Selection, crossover, and mutation can then proceed as in the SGA. Therefore, by ranking candidates based on a *dynamic* function of their fitnesses, multiple optima can be identified and maintained against the pressure of genetic drift exerted by the stochastic operation of the selection and crossover operators.

Estimation of the optimal value for the cutoff metric distance (or niche radius) σ_{sh} has been the subject of much work since the development of niching techniques. A simple but restrictive geometrical argument for estimating the proper niche radius can be found in Deb and Goldberg (1989) and is as follows:

- Assume each niche is represented by an N-dimensional hypersphere of radius σ_{sh}.
- If there are q optima in the parameter space, assume each niche hypersphere occupies a fraction $1/q$ of the parameter space (i.e., no overlapping).
- Calculate the radius of the hypersphere that encloses the *entire* population:

$$r = \frac{1}{2}\sqrt{\sum_{k=1}^{N_p}(x_{k,max} - x_{k,min})^2}. \tag{4.33}$$

- If the niches are not allowed to overlap, then the hypervolume of q niche hyperspheres should be equal to the volume of the whole parameter space

$$q\left(C\sigma_{sh}^N\right) = Cr^N, \tag{4.34}$$

where C is some constant that does not depend on the size of the hypersphere. For example, for $N = 3$, $C = 4\pi/3$. This then yields

$$\sigma_{sh} = \frac{r}{\sqrt[N]{q}}. \tag{4.35}$$

A glaringly obvious downside to this method is that the (approximate) number of optima in the parameter space must be known beforehand, and this information is rarely available in practical real-world problems. Furthermore, this procedure assumes the optima are roughly evenly distributed through the parameter space (no niche overlap), and are roughly the same size. Again, this is a dangerous assumption on many problems with unknown fitness function topologies. Therefore, a niching technique is needed in which the number, size, and distribution of the niches can be dynamically adjusted as the genetic algorithm explores new regions of the parameter space, thereby eliminating the need for detailed *a priori* knowledge of the fitness function landscape. Another inherent downfall of the NGA is that a niche is not strictly well-defined, except within the part of the algorithm which calculates the shared fitnesses of the population. This means that the evolution of clusters of individuals cannot be tracked dynamically in the rest of the algorithm, since there is no information available about the size, shape, or content of a niche. This is a tremendous loss of information, which can be remedied by utilizing a niche representation that is external to the fitness function.

The Dynamic Niche Clustering Algorithm

The Dynamic Niche Clustering (Gan and Warwick [2000, 2001, 2002]) technique maintains a persistent set of niches, aptly named the *nicheset*, which evolves alongside the main population. The niches are persistent in that, unless a niche is targeted for deletion or scheduled for creation (see below), the niches in generation t are retained in generation $t + 1$. As with the NGA, the DNC algorithm is built on the foundation of the SGA, with added operators and instructions. Each niche in the nicheset consists of a unique set of the following variables:

- $\vec{m}_j$ is the vector midpoint, in parameter space, of the j^{th} niche;

- σ_j is the radius of the j^{th} niche; and
- $[\Lambda_j]$ is a list of individuals that currently belong to the j^{th} niche.

A candidate (or individual) $\vec{\nu}_i$ is considered to be a member of the j^{th} niche if it lies inside the hypersphere defined by the niche midpoint and radius

$$\vec{\nu}_i \in [\Lambda_j] \quad \text{iff } \|\vec{\nu}_i - \vec{m}_j\| \leq \sigma_j. \tag{4.36}$$

Whereas the NGA was limited to a static value for the niche radius, DNC incorporates a variable-radius scheme whereby the niche size is allowed to vary during the execution of the algorithm, to more accurately map out the overall extent of an optimum. Nevertheless, an initialization is still required. An initial niche radius is defined based on the size of the population (P) and the dimension of the parameter space (D)

$$\sigma_{init} = \lambda \frac{\sqrt{D}}{\sqrt[D]{P}}, \tag{4.37}$$

where λ is a constant parameter used to control the amount of overlap between niches.

To initialize the DNC algorithm, a random population is generated, as in the NGA and SGA. A niche is added to the nicheset for every individual in the population, centered on that individual ($\vec{X}_j = \vec{\nu}_j$) and with radius σ_{init}. Since there will be a fair amount of overlap in the initial randomly generated nicheset, it is preprocessed to eliminate excessive overlap. If the distance between any two individuals (and hence niche midpoints) is less than half the initial niche radius, those individuals can be represented by a single niche. The redundant niche with the lowest fitness at its midpoint is deleted from the nicheset. Initially, there are P niches, and DNC requires $\mathrm{O}(N_n^2)$ niche comparisons each generation, where N_n is the number of niches, so cutting down on the initial computational expense by deleting redundant niches in the initial generation yields a tremendous savings in required computations. Every individual is then compared to every niche in the current nicheset, using Eq. 7, to map out the initial members of each niche. An important aspect of DNC is that individuals are not constrained to be members of only a single niche. An individual can

belong to more than one niche, and niches may overlap to a certain degree without being *merged* (see below). In each generation, the following procedures are executed after any fitness scaling (if applicable) and/or genetic operators are used, and before the application of any fitness-sharing technique(s):

(1) The current members of each niche are redetermined using Eq. 36. If an individual does not belong to any of the niches in the current nicheset, a new niche centered on that individual is spawned and added to the nicheset. If a niche has no members, it is deleted from the nicheset.

(2) The midpoint (centroid) of each niche is moved to the fitness-weighted average position of all the members of the niche

$$\vec{m}_j = \vec{m}_j + \frac{\sum_{\vec{\nu}_i \in [\Lambda_j]} (\vec{\nu}_i - \vec{m}_j) f_i}{\sum_{\vec{\nu}_i \in [\Lambda_j]} f_i}, \tag{4.38}$$

where f_i is the fitness of individual $\vec{\nu}_i$.

(3) The niche members are recalculated using Eq. 36, since the newly moved niche may have encompassed new individuals.

(4) Every niche in the current nicheset is compared to every other niche, and the Euclidean distance between their midpoints is calculated and stored in the array D.

(5) If $D_{ij} \equiv ||\vec{m}_i - \vec{m}_j|| \leq (\sigma_i \; or \; \sigma_j)/2$, the i^{th} and j^{th} niches are *merged* together to form a single niche that replaces the i^{th} niche, and all references to the j^{th} niche are deleted from the nicheset. The merger can occur in two different ways:

(a) If one niche completely envelops the other, the members of the smaller niche are absorbed into the larger niche, replacing the i^{th} niche, and the smaller niche is deleted from the nicheset. The niche radius of the new niche is simply the niche radius of the larger niche. By examining the intersection of a line connecting the midpoints of

the two niches with the niche boundaries themselves, it is straightforward to determine if a niche is completely enveloped by another niche, and hence should be absorbed.

(b) If neither niche can absorb the other, they are simply coalesced into a single niche that replaces the i^{th} niche. The midpoint of the new niche is recalculated using the Fitness Distribution From Midpoint method, which moves the midpoint to the average fitness-weighted distance of all members of both niches from the *average midpoint*, defined as

$$\vec{m}_{avg} = \vec{m}_i + \frac{\vec{m}_j - \vec{m}_i}{2}. \tag{4.39}$$

The new midpoint of the merged niche is then calculated as

$$\vec{m}_{new} = \vec{m}_{avg} + \frac{\vec{w}_i + \vec{w}_j}{\sum\limits_{\vec{\nu}_p \in [\Lambda_i]} f_p + \sum\limits_{\vec{\nu}_q \in [\Lambda_j]} f_q}, \tag{4.40}$$

where $\vec{w}_{i,j}$ are the weighting vectors for the i^{th} and j^{th} niche

$$\vec{w}_{i,j} = \sum_{\vec{\nu}_k \in [\Lambda_{i,j}]} (\vec{\nu}_k - \vec{m}_{avg}) \ f_k. \tag{4.41}$$

The radius of the resulting niche is then recalculated by observing where $\vec{m}_{new}$ is placed, relative to the old midpoints of the i^{th} and j^{th} niche. If the new midpoint is closer to the previous midpoint of the i^{th} niche, the niche radius is recalculated as the distance from $\vec{m}_{new}$ to the furthest extent of the i^{th} niche

$$\sigma_{i,new} = \|\vec{x}_i\| + \sigma_i, \tag{4.42}$$

where $\vec{x}_i = \vec{m}_{new} - \vec{m}_i$. This procedure is conversely applied (with the transformation $i \rightarrow j$) if the new midpoint is closer to the j^{th} niche. To prevent unlimited growth or contraction of a niche boundary, upper ($2\sigma_{init}$) and lower ($0.5\sigma_{init}$) limits to the niche radius must be observed. Since this process moves the midpoint of the i^{th} niche, the i^{th} row and column of the midpoint distance array, D, must be recalculated before proceeding with any other comparisons.

(6) Redundant niches are once again deleted from the nicheset, the current niche members are recalculated, and the sharing function (one of Eq. 37, 47, or 48) is applied to the entire population to obtain the shared fitnesses of all individuals. Ranking, selection, breeding, and mutation now proceed as in the NGA and SGA, generating a new population.

(7) Steps (1)–(6) are repeated with the new population until convergence criterion is satisfied, or until some predefined maximum number of generations have evolved.

There are many sharing functions that can be used with any niching method, but here I present only two. The first sharing function is only slightly different from that used with the NGA

$$m_i = \sum_{\vec{\nu}_j \in [\Lambda_i]} n_i \left[1 - \left(\frac{d_{i,j}^m}{2\sigma_i} \right)^\alpha \right], \tag{4.43}$$

where now $d_{i,j}^m$ is the metric distance of the j^{th} individual from the centroid (midpoint) of the i^{th} niche. This sharing function serves to penalize individuals that are closely-spaced within the niche, so that the action of this function is to spread the members of the niche around the niche midpoint, since individuals that are further away from the midpoint contribute more to the shared fitness of any one niche member. This can be illustrated with a simple example; imagine the j^{th} niche has N members. In one case, all N individuals lie exactly at the midpoint of the niche, so that $d_{i,j}^m = 0 \ \forall i$. This gives $m_i = N$, leading to a shared fitness of $f_{sh,i} = f_{raw,i}/N$. Now consider the case when the N niche members are spread out, giving $d_{i,j}^m \geq 0 \ \ \forall i$. This leads to a net niche count $m_i \ \leq \ N$, which, when divided into the raw fitness, yields a higher shared fitness than the previous tightly clustered example. Therefore, this particular sharing function exerts an evolutionary pressure on the members of each niche, favoring situations where many niche members are spread over the niche. Whereas this is useful for exhaustive exploration *around* the actual optimum, in some cases it may cause the solution inferred by the algorithm to be slightly suboptimal.

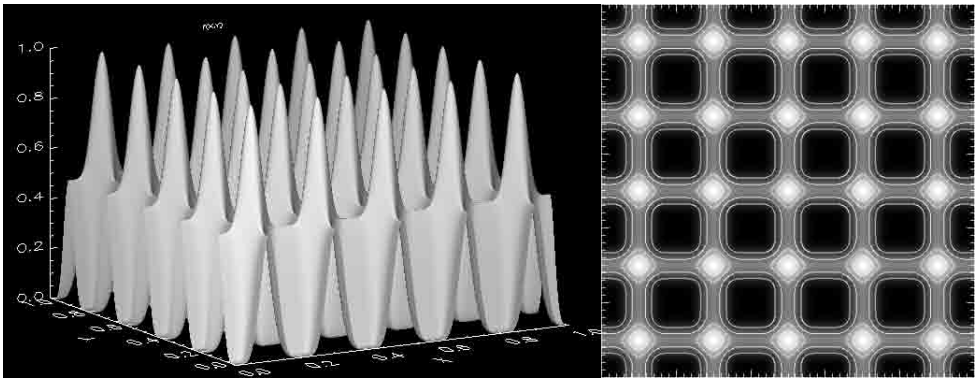

FIG. 4.3 A sample multimodal function. The figure shows a 3D representation of Equation 4.45 (left frame) and the corresponding contour plot (right frame). The 25 maxima are located at 0.1, 0.3, 0.5, 0.7, 0.9 $\forall\ x_i$.

The second sharing function is simply the number of niche members, i.e.,

$$m_i = n_j, \quad \text{iff } \vec{\nu}_i \in [\Lambda_j]. \tag{4.44}$$

This has the effect of reducing the fitness of every niche member by the same amount, retaining the *relative* fitness distribution within the niche. Since there is no differential fitness reduction based on position within the niche, this sharing function should show excellent convergence to an arbitrarily small neighborhood around the optimum, resulting in much tighter clusters at the optimal positions, as well as slightly lower execution times, due to the lack of Euclidean distance computations. The computational savings to be gained from the use of this function should become increasingly noticeable as the dimension of the parameter space increases.

As a testament to the power of this approach, Figure 4.3 shows a simple multimodal function for which all the maxima (of value 1) are to be located. Explicitly, it is described by the equation

$$f(\vec{x}) = 1 - \frac{1}{2}\sum \left[1 - sin^6\left(5\pi x_i\right)\right], \quad x_i \in [0, 1]. \tag{4.45}$$

An SGA attempting to find the maxima of this function would locate only a single peak, due to the premature convergence of genetic drift. Furthermore, as can be seen from the figure, the function contains many saddle-like areas that would also fool a local gradient

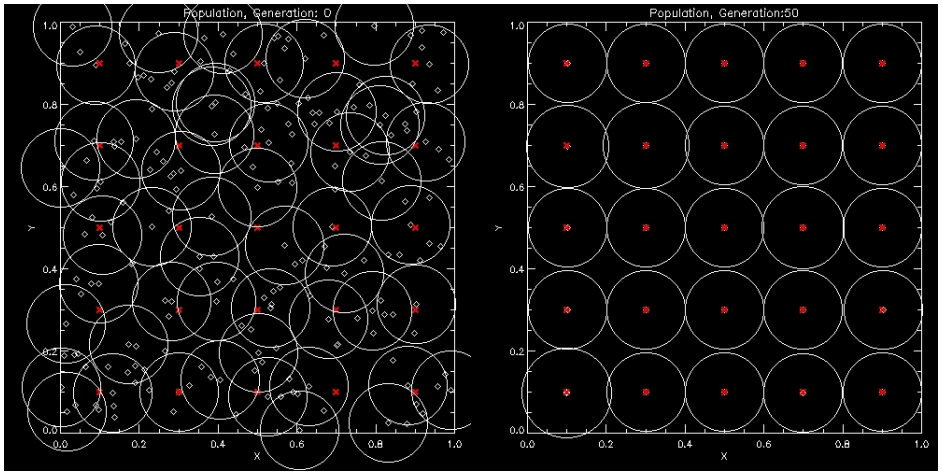

FIG. 4.4 The power of the Dynamic Niche Clustering algorithm. *(Left Frame):* The randomly-initialized population of 100 individuals (diamonds). Each individual belongs to at least one niche. The circles represent each individual niche, with an independent midpoint and radius. *(Right Frame):* The final configuration of the population after 50 generations. The red crosses mark the positions of the optima. Over the course of the evolution, the niches have adjusted their sizes to be representative of the peaks which they occupy. All 25 optima have been located very precisely.

ascent algorithm. Figure 4.4 shows an initial population of 100 individuals, as well as the configuration of the final population, 50 generations later, in which *all* optima have been identified.

The previous subsections have described the inner workings of a genetic algorithm, as well as some improvements to the basic mechanisms contained therein. The next section details how I have applied the genetic algorithm to the problem of inferring the magnetic field geometry in solar active regions from observations of their polarization profiles, along with some computational considerations on calibrating the genetic algorithm performance.

4.3 Stokes Polarization Profile Calculations

Before the genetic inversion can be applied to real data, several pre-processing steps must be taken to calibrate the data, and to massage it into a form appropriate for the genetic algorithm. This section details these steps, namely wavelength calibration, continuum determination, normalization, and the establishment of appropriate boundaries on the

Milne-Eddington parameter space. Finally, to bring this chapter to an end, I present several tests of the genetic inversion on synthetic, noisy data as well as a test case using data from the ASP.

4.3.1 Wavelength Calibration

The first step in applying the genetic algorithm to Stokes inversion is to ensure several ancillary quantities needed in the inversion routine(s) are properly calculated. Firstly, we calibrate the wavelength of the observations. This is done differently for the datasets obtained by different instrumentation. For the ASP and DLSP datasets, we have the convenient advantage of working in a spectral region containing two absorption lines of molecular Oxygen that are produced in the Earth's atmosphere. The advantage of this approach is that these terrestrial absorption features are not subject to the Fe I wavelength shifts produced by the motion of the solar photosphere (convection, field-guided plasma flows), and therefore are sharp, *stationary* spectral features. The two telluric absorption lines have a well-defined spectral separation of 0.7624 mÅ in their line cores (Pierce and Breckinridge (1973)). Figure 4.5 shows a sample quiet-sun intensity profile obtained by the DLSP, highlighting these terrestrial absorption features.

The line cores of the Telluric absorption lines typically falls between two ASP/DLSP sampled wavelength bins. Therefore, a 2^{nd} order polynomial is fit to the range of data lying strictly in the Telluric features (i.e., no continuum is included). From the fits, an accurate line-center (in fractional pixels) is then determined for both Telluric lines. The known spectral separation of the lines is then divided by the difference in fractional pixel line-center just determined. This gives a scale of approximately 11 mÅ/pixel for ASP data and approximately 38 mÅ/pixel for DLSP data. These scales (dispersions) are then used, along with the known location of one of the Telluric lines, to set the wavelength value for each pixel in the spectra.

The situation is very different for the data obtained with the Hinode satellite. Being in low Earth orbit, the satellite is outside the spectral influence of the Earth's atmosphere,

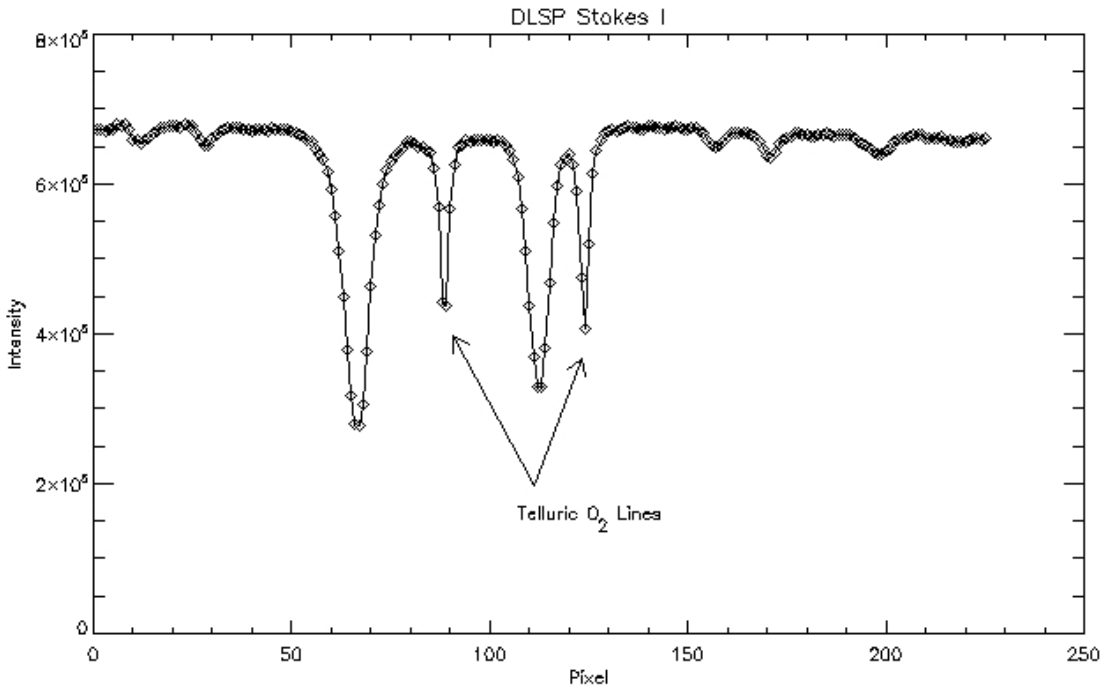

FIG. 4.5 Telluric O_2 terrestrial absorption lines. The figure shows two spectrally-stationary O_2 absorption lines inherent in ground-based spectra of the region surrounding the Fe I multiplet.

and hence the Telluric lines are not present in Hinode data. In this case, wavelength scale is determined by averaging the quiet-sun intensity profile over a large portion of the quiet-sun pixels in the field-of-view. This averaged profile is then compared to the Liegi Atlas of the solar spectrum, which contains standardized measure of solar intensity as a function of wavelength between 3601Å and 9300Å. The averaged quiet-sun profile is interpolated between the wavelengths of the Liegi Atlas such that the line-cores of the Fe I multiplet exactly agree with the atlas.

4.3.2 Continuum Determination

The continuum intensity of every spectral line profile obtained for a pixel in the field-of-view is the "background" intensity far-removed from the spectral line core. In deeper, hotter atmospheric layers, the source function increases, and therefore so does the photon flux. From the Eddington-Barbier relation, the flux of photons of wavelength λ comes

TABLE 4.1 Average and corresponding 1σ uncertainties in the continuum intensities for the ASP, DLSP, and *Hinode* spectropolarimeters.

Instrument	$< I_c >$ (detector counts)	σ_I (%)
ASP	12218.4	0.29
DLSP	624794.0	0.60
Hinode	27279.1	0.55

from layers where the optical depth at that wavelength, τ_{λ_0}, is approximately 2/3. The continuum intensity is an important physical quantity for the inversion, and in this work we calculate it by averaging the Stokes I intensities over a range of wavelengths that are devoid of other absorption features, and far-removed from the center of the Fe I 6301.5Å line. Figure 4.6 shows the appropriate quiet ranges for the ASP, DLSP, and Hinode data used in this thesis.

A benefit of this determination is the calculation of a corresponding instrumental uncertainty in the intensity values, effected by calculating the standard deviation of the intensities in the quiet range. The average continuum intensity and corresponding instrumental uncertainties so-calculated for the ASP, DLSP, and Hinode datasets is shown in Table 4.1.

4.3.3 Normalization

In the standard Milne-Eddington atmosphere, the assumptions are as follows:

- The opacity ratio, η_0, is constant with optical depth.
- The magnetic field, $\vec{B}$, is constant with optical depth.
- The source function, $S(\tau)$, is linear in optical depth: $S(\tau) = S_0 + \mu S_0 \tau$, where $\mu = 1/cos(\theta)$ and θ is the observer's viewing angle.

These assumptions, although restrictive, are generally applicable in the very thin atmospheric layer we call the photosphere; the thickness of this layer is much less than the pressure or temperature scale height, so quantities like the opacity ratio and magnetic field

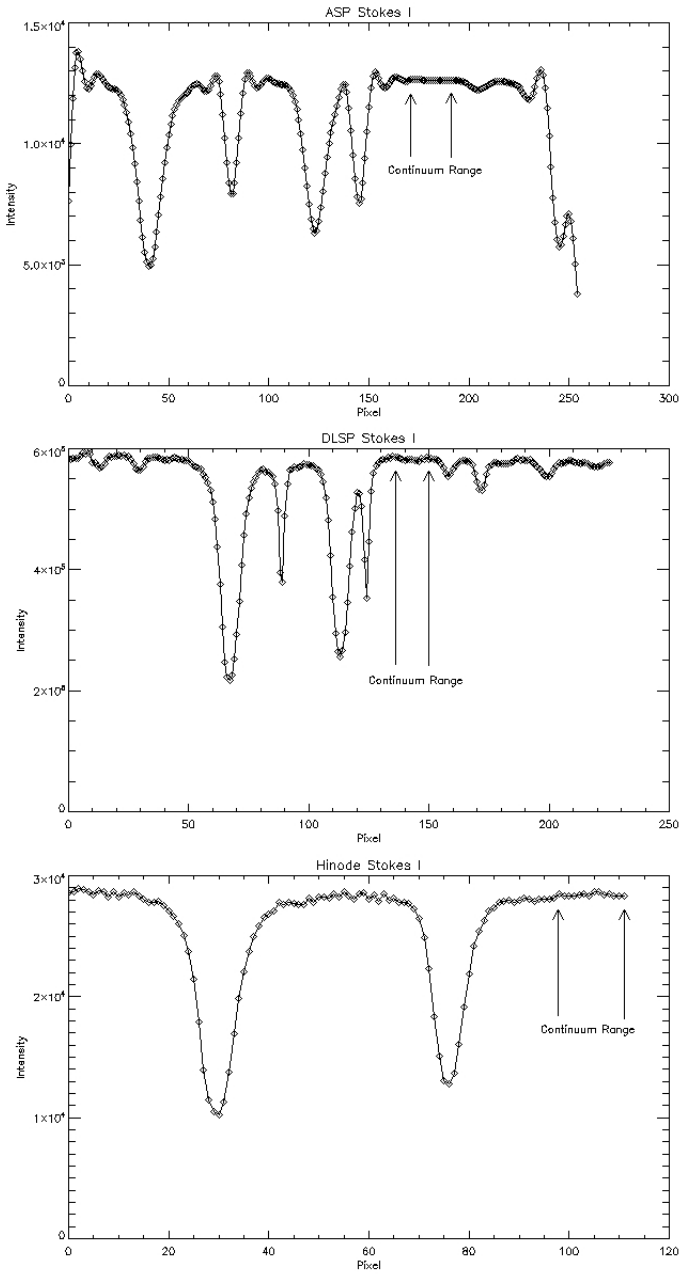

FIG. 4.6 Continuum ranges for ASP, DLSP, and *Hinode* data. The figure shows the range of spectrally-quiet wavelengths used to determine the continuum intensity and a corresponding estimate of the uncertainty in intensity for the ASP (top), DLSP (center), and Hinode (bottom) spectropolarimeters.

must not change appreciably from the bottom to the top of this layer. The source function coefficients, S_0 and S_1, typically enter the model as free parameters to be determined. However, by a judicious choice of normalization, I have completely eliminated the dependence of the fit on these parameters, as well as the μ-value. Instead of working directly with the Stokes profiles, I choose to work with the *line depression*, which is essentially the line profile turned upside-down. Recall from Eqns. 2.131 through 2.134 in Chapter 2 that the Milne-Eddington synthetic Stokes profiles (as a function of wavelength at the surface $\tau = 0$) are given by:

$$I(\lambda,0) = S_0 + \frac{\mu S_1}{\Delta(\lambda)}(1+\eta_I(\lambda))\left[(1+\eta_I(\lambda))^2 + \rho_Q(\lambda)^2 + \rho_U(\lambda)^2 + \rho_V(\lambda)^2\right] \tag{4.46}$$

$$Q(\lambda,0) = -\frac{\mu S_1}{\Delta(\lambda)}\left[(1+\eta_I(\lambda))^2\eta_Q(\lambda) + (1+\eta_I(\lambda))(\eta_V(\lambda)\rho_U(\lambda) - \eta_U(\lambda)\rho_V(\lambda)) + \rho_Q(\lambda)R(\lambda)\right] \tag{4.47}$$

$$U(\lambda,0) = -\frac{\mu S_1}{\Delta(\lambda)}\left[(1+\eta_I(\lambda))^2\eta_U(\lambda) + (1+\eta_I(\lambda))(\eta_Q(\lambda)\rho_V(\lambda) - \eta_V(\lambda)\rho_Q(\lambda)) + \rho_U(\lambda)R(\lambda)\right] \tag{4.48}$$

$$V(\lambda,0) = -\frac{\mu S_1}{\Delta(\lambda)}\left[(1+\eta_I(\lambda))^2\eta_V(\lambda) + (1+\eta_I(\lambda))(\eta_U(\lambda)\rho_Q(\lambda) - \eta_Q(\lambda)\rho_U(\lambda)) + \rho_V(\lambda)R(\lambda)\right] \tag{4.49}$$

with the continuum intensity given by $I_c = S_0 + \mu S_1$. The first step in the normalization is to subtract the line from the continuum:

$$I_c - I(\lambda,0) = S_0 + \mu S_1 - \left(S_0 + \frac{\mu S_1}{\Delta(\lambda)}(1+\eta_I(\lambda))\left[(1+\eta_I(\lambda))^2 + \rho_Q(\lambda)^2 + \rho_U(\lambda)^2 + \rho_V(\lambda)^2\right]\right) \tag{4.50}$$

$$= \mu S_1 - \frac{\mu S_1}{\Delta(\lambda)}(1+\eta_I(\lambda))\left[(1+\eta_I(\lambda))^2 + \rho_Q(\lambda)^2 + \rho_U(\lambda)^2 + \rho_V(\lambda)^2\right] \tag{4.51}$$

$$= \mu S_1\left(1 - \frac{1}{\Delta(\lambda)}(1+\eta_I(\lambda))\left[(1+\eta_I(\lambda))^2 + \rho_Q(\lambda)^2 + \rho_U(\lambda)^2 + \rho_V(\lambda)^2\right]\right). \tag{4.52}$$

To complete the normalization, note that since all the terms on the RHS of the above equations are actually functions of wavelength, we can divide by the same quantity evaluated at some convenient wavelength, which eliminates the multiplying factor μS_1. The most-

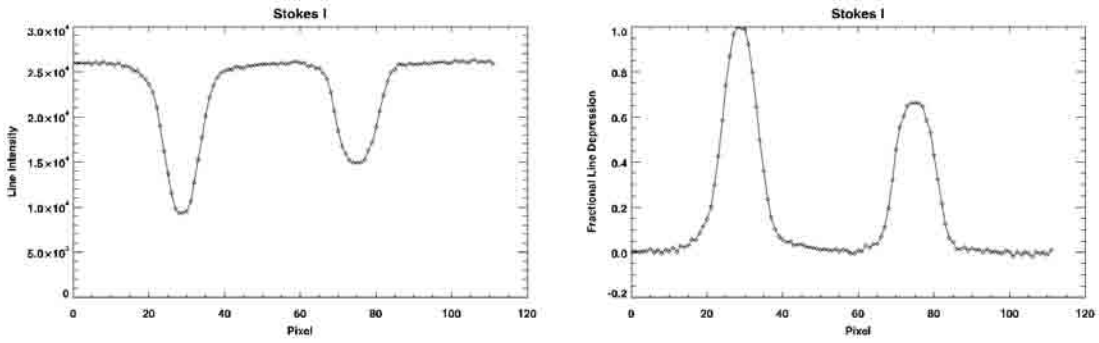

FIG. 4.7 Normalization of raw spectral data. *(Left Frame):* The raw intensity spectral line profiles from the Hinode satellite. The units along the vertical axis are detector counts, a proxy for intensity or brightness. *(Right Frame):* The normalized profiles used in the inversions in this work. Note the units along the vertical axis are no longer in absolute intensity units, but rather in relative depth from the continuum.

obvious choice would be to evaluate it at the line-center, λ_0:

$$\frac{I_c - I(\lambda,0)}{I_c - I(\lambda_0,0)} = \frac{1 - \frac{1}{\Delta(\lambda)}(1+\eta_I(\lambda))\left[(1+\eta_I(\lambda))^2 + \rho_Q(\lambda)^2 + \rho_U(\lambda)^2 + \rho_V(\lambda)^2\right]}{1 - \frac{1}{\Delta(\lambda_0)}(1+\eta_I(\lambda_0))\left[(1+\eta_I(\lambda_0))^2 + \rho_Q(\lambda_0)^2 + \rho_U(\lambda_0)^2 + \rho_V(\lambda_0)^2\right]}. \tag{4.53}$$

Therefore, the inversion algorithm works explicitly with the normalized Stokes profiles:

$$I'(\lambda) = \frac{I_c - I(\lambda,0)}{I_c - I(\lambda_0,0)} \tag{4.54}$$

$$Q'(\lambda) = \frac{Q(\lambda,0)}{I_c - I(\lambda_0,0)} \tag{4.55}$$

$$U'(\lambda) = \frac{U(\lambda,0)}{I_c - I(\lambda_0,0)} \tag{4.56}$$

$$V'(\lambda) = \frac{V(\lambda,0)}{I_c - I(\lambda_0,0)}, \tag{4.57}$$

and in doing so, we reduce the dimension of the searchable parameter space by 2, which will increase the efficiency with which the other dimensions can be explored. Figure 4.7 shows the difference between the raw Stokes profiles and the normalized profiles used in this work.

4.3.4 Boundaries on the Model Parameters

It is important for a genetic algorithm to have the smallest bounds possible, for efficient searching capabilities. To this end, I have implemented an algorithm that determines what

TABLE 4.2 Boundaries on pixel brightness for determining if a pixel is situated in the umbra, penumbra, or quiet-sun. The percentages refer to the line continuum (intensity averaged over quiet wavelengths) relative to the same value calculated over the average quiet-sun profile. Note that since the quiet-sun profile is averaged over a large field-of-view, quiet-sun pixels are allowed to vary by 10%.

Instrument	Umbra ($\%I_c$)	Penumbra ($\%I_c$)	Quiet-Sun ($\%I_c$)
ASP	$< 55\%$	$> 55\%$ and $< 90\%$	$> 90\%$
DLSP	$< 70\%$	$> 70\%$ and $< 90\%$	$> 90\%$
Hinode	$< 35\%$	$> 35\%$ and $< 90\%$	$> 90\%$

TABLE 4.3 Boundaries within the parameter space defined by the M-E model atmosphere. The boundaries on physical parameters of the model are determined by the comparisons in Table 4.2. Refer to the text for an explanation of the bounds on the field inclination, Ψ.

Parameter	Umbra	Penumbra	Quiet-Sun
$\|\vec{B}\|$ (G)	2000 - 4000	1000 - 3000	0 - 2000
Ψ ($^\circ$)	*	*	*
χ ($^\circ$)	0 - 180	0 - 180	0 - 180
α	0.8 - 1.0	0.5 - 1.0	0.0 - 1.0
$\Delta\lambda_D$ (mÅ)	15.0 - 25.0	20.0 - 40.0	20.0 - 60.0
a	0.0 - 1.0	0.0 - 1.0	0.0 - 1.0
η_0	0.0 - 25.0	0.0 - 25.0	0.0 - 25.0

region (umbra, penumbra, quiet-sun) the current pixel belongs to, and adjusts the parameter space bounds accordingly. Which region a pixel belongs to is determined by comparing the line continuum intensity to the quiet-sun continuum intensity. Table 4.2 shows the range of intensity values which determine region-membership. Using some "safe" *a priori* estimates of the magnitudes of the various parameters in each region, the genetic algorithm dynamically changes its boundaries according to Table 4.3.

The inclination of the magnetic field from the observer's line-of-sight is the only physical

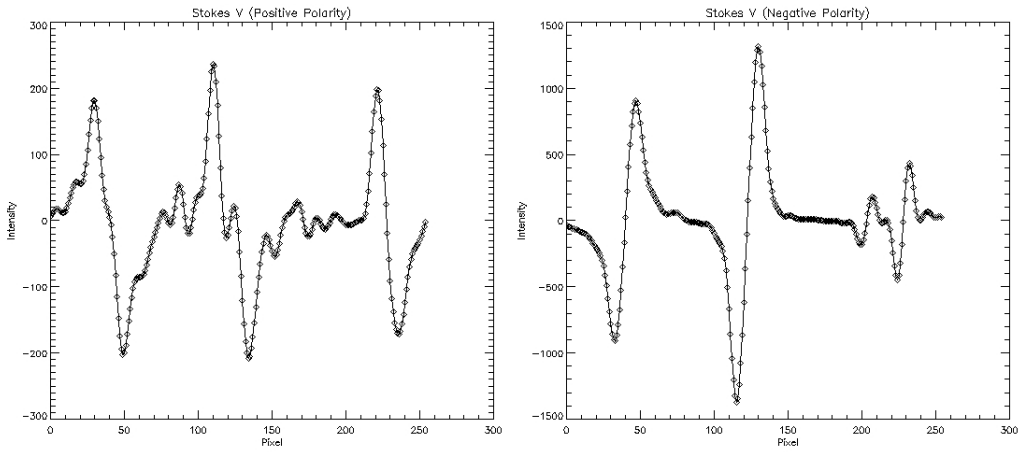

FIG. 4.8 Sample Stokes V spectra for opposite-polarity magnetic fields. The figure shows Stokes V circular polarization profiles from a positive-polarity umbral pixel (left) and a negative-polarity pore pixel (right) in AR9240. The spectral position of the positive and negative lobes of the Fe I absorption line reverse for opposite polarity magnetic fields.

parameter that allows a safe estimation of its appropriate boundaries, *from the data itself*. Figure 4.8 shows the difference between the Stokes V circular polarization profiles for a magnetic field inclined toward, and away from the observer's line-of-sight, respectively. The physical difference in this situation is that the former represents a positive-polarity field (coming out of the surface, toward the observer) and a negative polarity field (going into the surface, away from the observer). The difference is quite obvious, and allows a further reduction of the parameter space. For each pixel, the position of the maximum and minimum Stokes V signal is located. If the maximum position occurs at shorter wavelengths than the minimum position, the genetic algorithm is constrained to locate positive-polarity fields ($\Psi \in [0°, 90°]$). If the converse is true, we limit the inclination range to ($\Psi \in [90°, 180°]$) for negative-polarity fields. This ensures that the genetic algorithm does not "waste any time" handling individuals with the completely wrong field orientation.

One final preprocessing step occurs before the genetic algorithm is allowed to search for the correct model parameters. The total degree of polarization, $p(\lambda)$, is calculated from the

data as

$$p(\lambda) = \frac{\sqrt{Q^2(\lambda) + U^2(\lambda) + V^2(\lambda)}}{I(\lambda)}. \tag{4.58}$$

A perfectly-polarized observation will have $p(\lambda) = 1, \ \forall \ \lambda$, while an unpolarized observation will have $p(\lambda) = 0, \quad \forall \ \lambda$, and a typical observation obtains values between these two extremes. The degree of polarization depends on the magnetic field strength in the pixel area; therefore, if the degree of polarization is low there is weak or no magnetic field in the pixel, and it is pointless to determine a magnetic profile fit to a nonmagnetic spectral line. To this end, if the integrated degree of polarization (between the continuum limit of the 6301.5Å spectral line),

$$P = \int_{\lambda_0 - \Delta\lambda}^{\lambda_0 + \Delta\lambda} p(\lambda) d\lambda, \tag{4.59}$$

has a value of 0.5 or less, and if $MAX[p(\lambda)] < 0.05$, that pixel is excluded from the inversion process, and the algorithm cycles to the next available pixel. This step, depending on the field-of-view, affects a moderate reduction in the total runtime of the inversion algorithm.

4.3.5 Testing the Genetic Inversion

The above considerations, along with calibrating the genetic inversion procedure, have allowed me to reduce the dimension of the searchable parameter space from 10 (see Table 2.1) to 7. These 7 free parameters of the model atmosphere to be determined by the genetic algorithm are then as follows:

$$\vec{x} = (|\vec{B}|, \psi, \phi, \Delta\lambda_D, a, \eta_0, \alpha)^T. \tag{4.60}$$

To ensure the correct operation of our spectral line profile calculation subroutines, we present two special cases to the solution of the PRTE, and check for consistency. For a magnetic field parallel to the observer's line-of-sight ($\psi = 0°$), the Unno-Rachkovsky solutions (Unno [1956], Rachkovsky [1962]) satisfy the following identity:

$$I(\lambda + \Delta\lambda_B) + V(\lambda + \Delta\lambda_B) = I_{nm}(\lambda), \tag{4.61}$$

where $\Delta\lambda_B$ is the Zeeman splitting in the line due to a magnetic field of strength B. Furthermore, for a field perpendicular to the observer's line-of-sight ($\psi = 90°$), the following

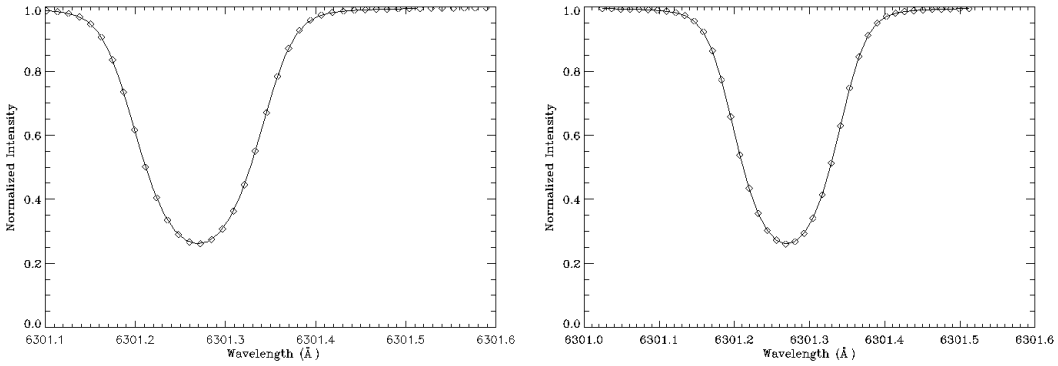

FIG. 4.9 Identities of the Unno-Rachkovsky solutions. The figure shows a confirmation of the correct workings of the Stokes polarization profile calculations in the algorithm. *(Left:)* The diamonds represent the quantity $I(\lambda+\Delta\lambda_B)+V(\lambda+\Delta\lambda_B)$, while the line is the non-magnetic line profile. *(Right:)* The diamonds represent the quantity $I(\lambda)+Q(\lambda)$, and again the solid line is the non-magnetic profile.

identity is true:

$$I(\lambda) + Q(\lambda) = I_{nm}(\lambda). \tag{4.62}$$

Figure 4.9 shows the line profiles given by both sides of the above equation. As can be easily seen in both cases, the identities are extremely well satisfied, and we may be confident that our line profile calculations are functioning as they should.

For this work, the fitness function used to evaluate individuals is the inverse of the χ^2 goodness-of-fit describing the agreement between our model and the observations. Specifically,

$$\chi^2 = \frac{1}{N_{dof}} \left(\sum_{i=I,Q,U,V} \sum_{j=1}^{N_{obs}} w_i \left[\mathbf{I}_{obs,i}(\lambda_j) - \mathbf{I}_{mod,i}(\lambda_j) \right]^2 \right), \tag{4.63}$$

where $\mathbf{I}_{I,Q,U,V}$ is the Stokes vector ($\mathbf{I} \equiv [I,Q,U,V]^T$), $\sigma_{i,j}$ are the (potentially wavelength-dependent) observational uncertainties, and the subscripts "obs" and "mod" refer to the observations and mathematical model, respectively. N_{dof} is the number of degrees of freedom in our model. The coefficients w_i are weights used to increase/decrease the importance of deviations of the model from the observations. Because the Stokes polarization profiles (Q, U, and V) typically take values that are roughly one or two orders of magnitude smaller than the total intensity profile, the measurement of the deviation between the observed

profiles and the profiles generated by the model will be dominated by the intensity profile. To ensure that the genetic algorithm treats deviations between all line profiles on an equal footing, we have employed a weighting scheme that scales the deviations to roughly the same value. The χ^2 merit function has been modified to

$$\chi^2 = \frac{1}{N_{dof}} \sum_{i=I,Q,U,V} \sum_{j=1}^{N_\lambda} w_i \left[\vec{I}_{obs}(\lambda_j) - \vec{I}_{mod}(\lambda_j) \right]^2, \tag{4.64}$$

where

$$w_I = \frac{1}{\sum_{j=1}^{N_\lambda} I_{obs}(\lambda_j)} \tag{4.65}$$

$$w_Q = \sum_{j=1}^{N_\lambda} |Q_{obs}(\lambda_j)| \tag{4.66}$$

$$w_U = \sum_{j=1}^{N_\lambda} |U_{obs}(\lambda_j)| \tag{4.67}$$

$$w_V = \sum_{j=1}^{N_\lambda} |V_{obs}(\lambda_j)|. \tag{4.68}$$

While this scaling will not, in general, allow the deviations to take completely equal values, it will ensure that they are at least of comparable magnitude, such that a small deviation from a low-signal-strength polarization profile will be treated the same as a small deviation from a high-signal-strength polarization profile.

To diagnose any potential problems in the genetic algorithm, we have generated a 1-component synthetic dataset from the model described above. Since these "observations" were generated using the mathematical model we wish to utilize in the genetic algorithm, it should be able to recover the known input parameters (almost) exactly. This diagnostic should shed light on trouble areas in the genetic algorithm. The resulting performance is displayed in Table 4.4, and the average inferred Stokes profiles are shown in Figure 4.10. The genetic inversion was performed 500 times on the synthetic spectral line profiles generated from the parameters in the "Input" column in the table. The average optimal parameter set inferred by the genetic algorithm and the corresponding 1σ standard deviations of the distribution are also shown. As can be seen, the genetic algorithm is very efficient and

TABLE 4.4 Robustness of the parameters recovered by the genetic inversion. Clearly, the thermodynamic parameters are the least-reliably determined of the whole set, although given their spread, their reduced reliability does not seem to effect the accuracy of the magnetic field parameters. The 500 independent iterations were performed with a population size of 100 evolving over 50 generations.

Parameter	Input	Average	Std. Deviation
$\|\vec{B}\|$ (G)	3000.0	3034.4	56.3
Ψ (deg)	45.0	44.2	1.2
ϕ (deg)	45.0	45.0	1.3
a	0.005	0.0049	0.0030
$\Delta\lambda_D$ (mÅ)	25.0	24.5	1.2
α	0.80	0.79	0.02
η_0	20.00	19.66	2.39

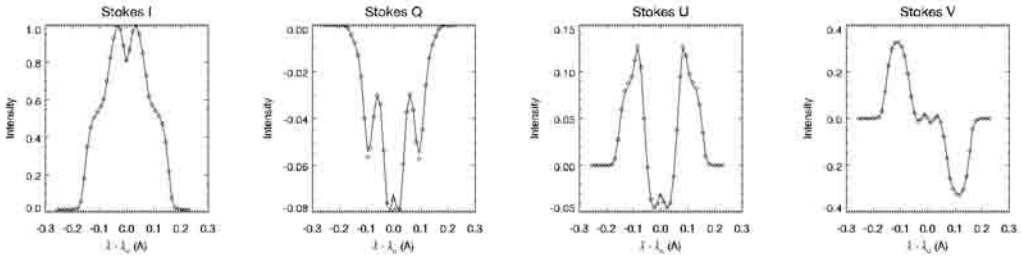

FIG. 4.10 Genetic recovery of the synthetic test dataset. These Stokes profiles were inferred by the genetic algorithm for the average values listed in Table 4.4. The fit (solid line) to the synthetic data (diamonds) is very good, and variations in the model parameters by values of the same magnitude as the standard deviations produce profiles that are indistinguishable from the average fit.

reliable, despite the pseudo-randomness behind many of its internal operations. The input parameter set was chosen to be indicative of a position within a typical sunspot umbra, near to the umbra-penumbra boundary—that is, with relatively high field strength ($|\vec{B}| = 3000\ G$), moderately inclined from the observer's line-of-sight ($\Psi = 45°$), with a high fill-factor ($\alpha = 0.80$) and low temperature ($\Delta\lambda_D = 0.025$ Å).

Taking this analysis one step further, I have run 500 separate genetic inversions on the

synthetic dataset, for various population sizes, with a static number of generations, set to 50. Figure 4.11 shows the average of the optimal parameters inferred by the genetic inversion as a function of increasing generations, and Figure 4.12 shows the corresponding 1σ standard deviations, indicating very high stability and repeatability for my algorithm. One can clearly identify a configuration which gives the best stability for a required runtime (being directly proportional to the population size). Any larger than this critical size simply wastes time and computational resources to obtain results that could be obtained with a small population size.

The figure above clearly demonstrates one of the limiting factors in genetic algorithm performance; because of the genetic drift phenomenon, genetic algorithms are exceedingly good at finding *near* optimal solutions with arbitrary accuracy, but struggle to locate the precisely *exact* optimal solution. This is particularly apparent in real-world, high-dimensionality problems. In this case, the loss of accuracy is mitigated by the fact that spectropolarimeters, which obtain the real data used in this work, typically can only distinguish magnetic fields which differ in strength by about 25 G. The standard deviations in field strength, even for a "small" population of 100 individuals, are not too dissimilar to this theoretical lower limit. The field orientation (inclination and azimuth) are recovered in many cases with tolerances of less than a degree. The remaining parameters ultimately assume similarly low spreads around the true input values, demonstrating that the genetic inversion algorithm is capable of obtaining accurate and repeatable estimates for the Milne-Eddington model parameters.

Figures 4.13 through 4.15 show the convergence properties and rates of the synthetic profile recovery for a population of 500 individuals evolving over 50 generations. From these figures we may be confident that the genetic algorithm is not wasting time around local optima, but is instead truly focusing on the most globally-promising regions of the parameter space, while at the same time promoting diversity to prevent the stalling of the evolutionary search.

As a final measure of the difficulty of the genetic inversion problem, a measure of the

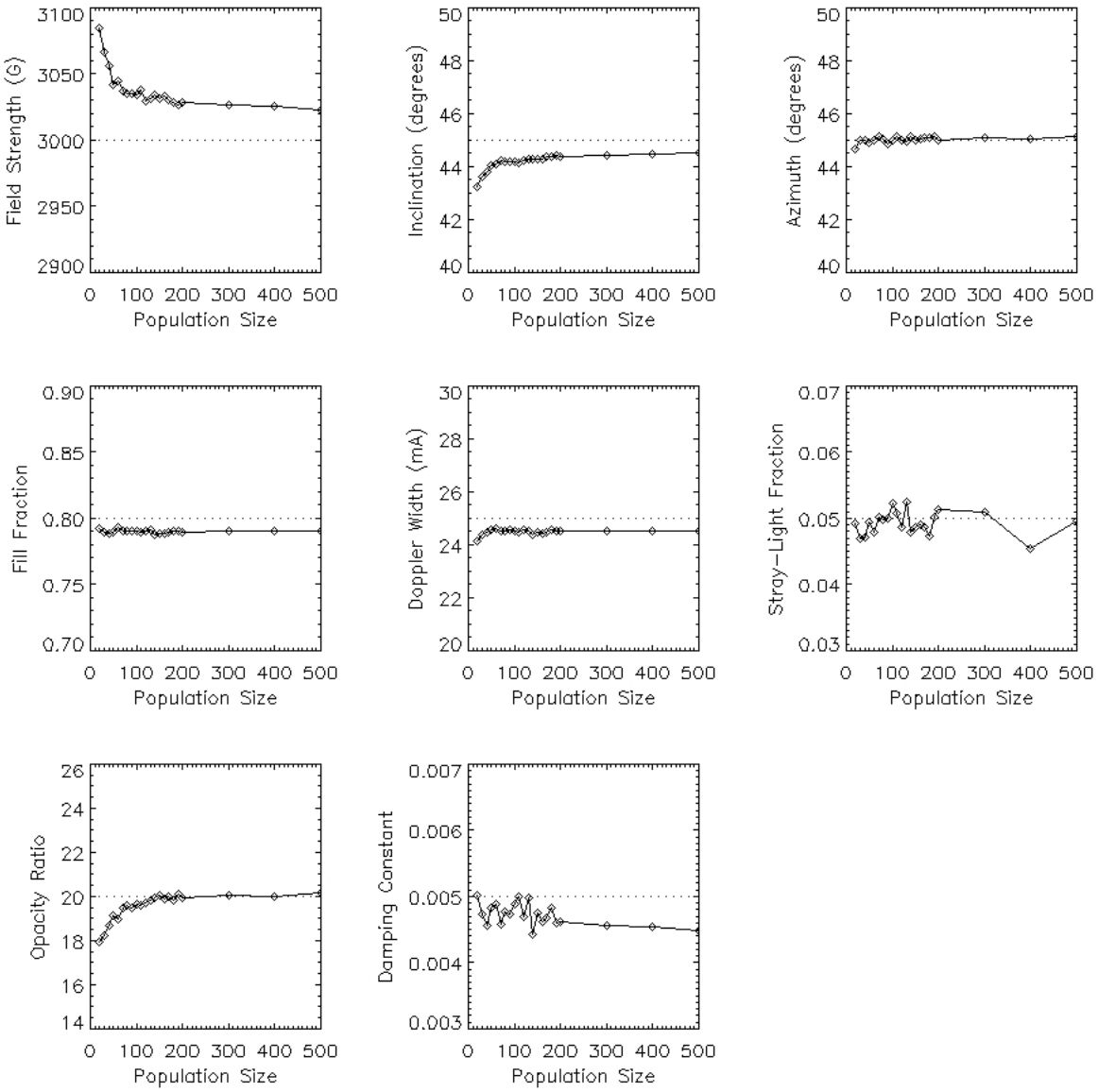

FIG. 4.11 Effect of population size on inferred parameters. The recovery of the known inputs (dotted lines) which generated the synthetic "observations" are shown as a function of population size in the genetic algorithm. It is obvious that small population sizes do not contain enough diversity in the initial stages of the genetic algorithm to accurately recover the input parameters. For larger population sizes, it is more likely that a member of the initial population will contain better initial estimates of the optimal parameters, which would then be exploited by the genetic algorithm.

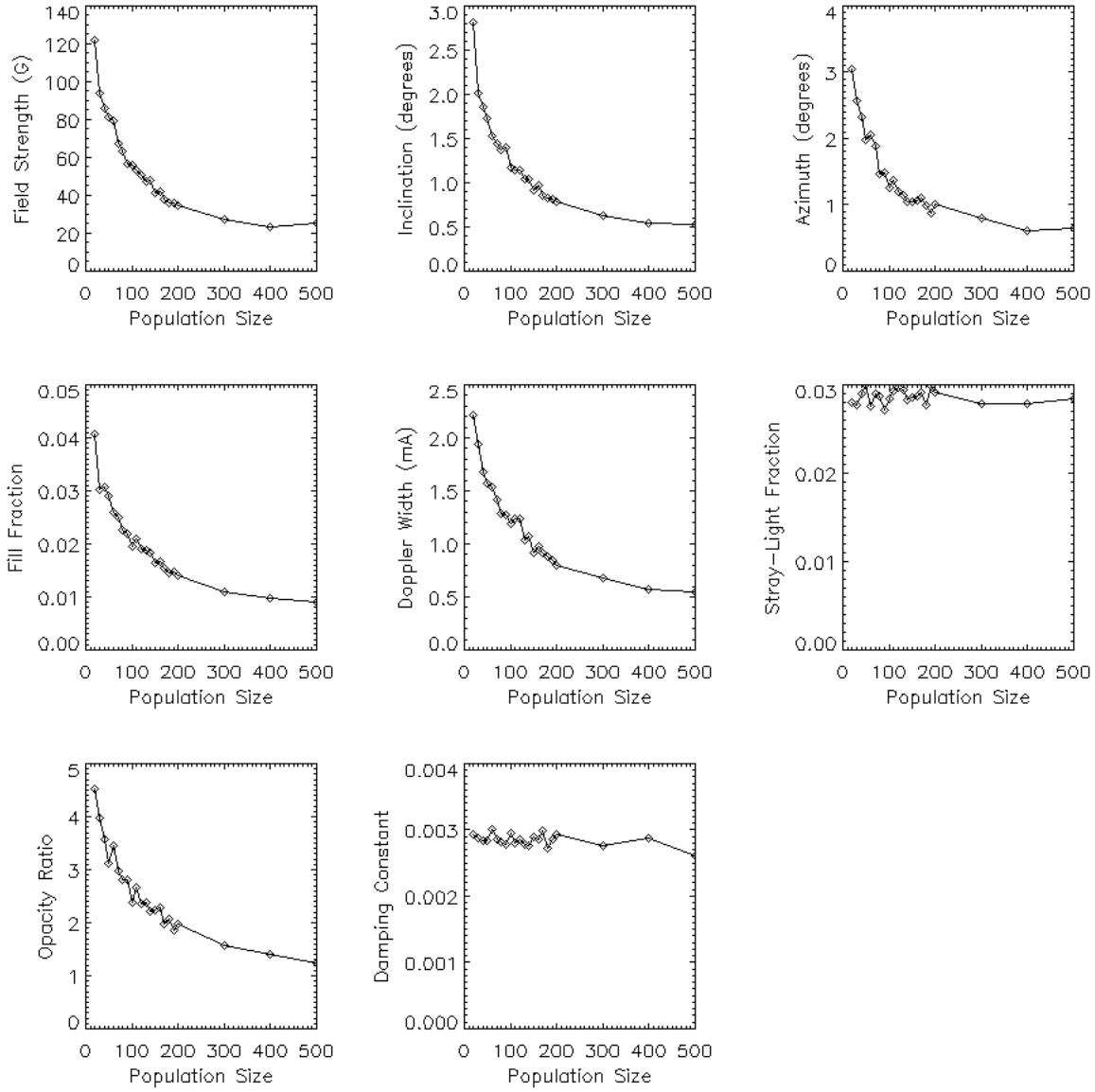

FIG. 4.12 Effect of population size on uncertainty in inferred parameters. As the population size increases, the stability of the algorithm is also increased. The spread around the average recovered values (see Figure 4.11) is well within the tolerances delimited by the observational uncertainties. Even for small population sizes, the uncertainties in the derived model parameters are not prohibitively large.

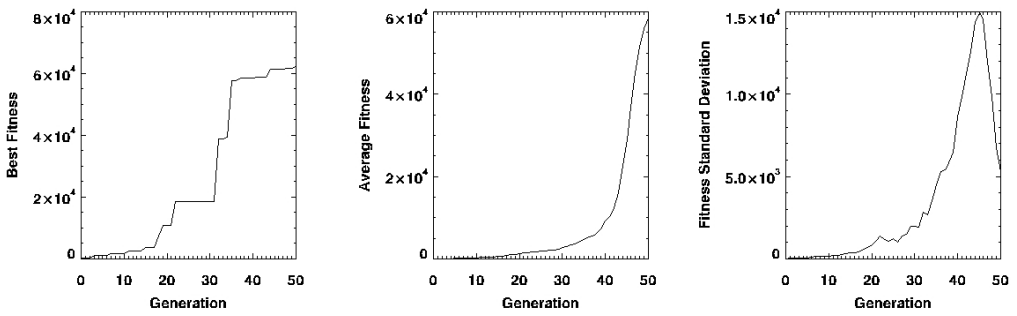

FIG. 4.13 Population convergence measured by fitness value. The figure shows the best fitness found by the genetic inversion at each generation *(left)*, the average fitness of the entire population at each generation *(center)*, and the standard deviation of the fitness around the average value in each generation *(right)*. The punctuated equilibrium of the genetic algorithm is evident in the left frame, while the center and right frames show the convergence of the rest of the population around a high-fitness solution found by the genetic algorithm. As the population begins to converge, the average fitness draws closer to the best fitness, though the high standard deviations indicate the population contains many individuals with very low fitness. At the end of the genetic algorithm run, as genetic drift takes over and converges the population to a small neighborhood, the standard deviation of the population fitness drops drastically.

correlation between fitness and distance from the optimum is shown in Figure 4.16. This correlation coefficient, C_{FD}, is given by

$$C_{FD} = \frac{1}{n\sigma_f\sigma_d}\sum_{i=1}^{N_p}\left(f_i - \bar{f}\right)\left(d_i - \bar{d}\right), \tag{4.69}$$

where n is the dimension of the fitness function, f_i, d_i are the fitness and distance from current optimum, respectively, and $\bar{f}, \bar{d}, \sigma_f, \sigma_d$ are the corresponding averages and standard deviations of these values over the entire population. This coefficient loosely measures how the fitness function behaves as distance from the optimum increases or decreases. Jones and Forrest (1995) and Tomasinni et al. (2005) have made an exhaustive analysis of C_{FD} for various types of optimization problems, and has shown that many types of problems that are typically GA-hard have $C_{FD} > 0.15$, while GA-easy problems have $C_{FD} < -0.15$. The intermediate region is somewhat of a GA-limbo, where the GA performance is hard to predict. A negative value for C_{FD} indicates that, generally, fitness decreases as distance from the optimum is increased. The positive C_{FD} indicates that fitness tends to *increase* as

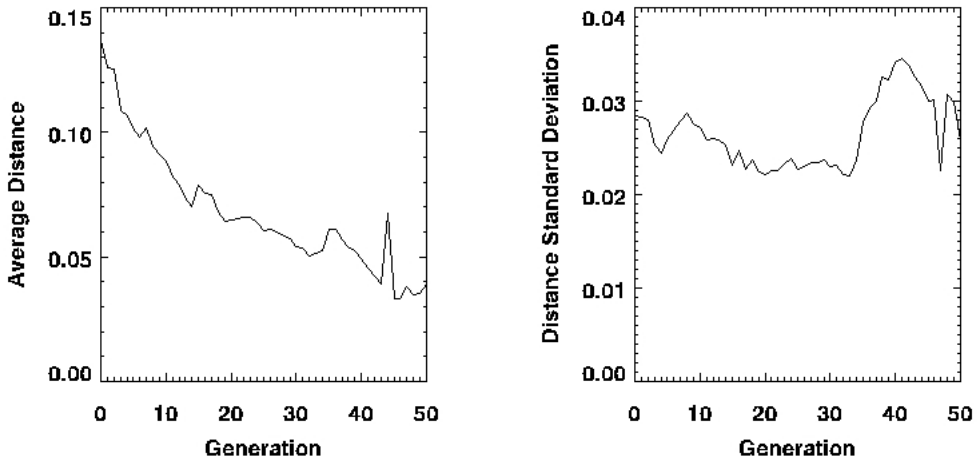

FIG. 4.14 Population convergence measured by Euclidean distance in parameter space. The figure displays the spatial convergence of the population around an optimal solution. The general downward trend of the average distance from the currently-optimal individual *(left)*, signals that the population is drawing closer to the region of parameter space occupied by the optimal individual. However, the spatial spread of the population *(right)* shows no such marked decrease, indicating that the population is retaining a degree of diversity despite its convergence, which is desirable to prevent premature convergence to a local optimum.

distance from the optimum increases. This somewhat misleading property is mitigated by the fact that no single scalar number can fully characterize any fitness function, but rather indicates an overall general behavior of the fitness function. This misleading property is exactly what makes an optimization problem GA-hard, and these types of problems are referred to as *deceptive* problems.

Although the previous considerations have shown that the genetic inversion is capable of inferring the correct model parameters with high accuracy, the real data we use in this work is far from perfect. To this end, the synthetic genetic inversion test is repeated with the same input model parameters, but with 1%, 5%, and 10% Gaussian white noise artificially added to the synthetic observations. The percentages refer to the maximum amount of contamination; 1% noise implies that the most the signal can be contaminated by is 1% of the continuum intensity. These noisy synthetic observations are more true to the nature

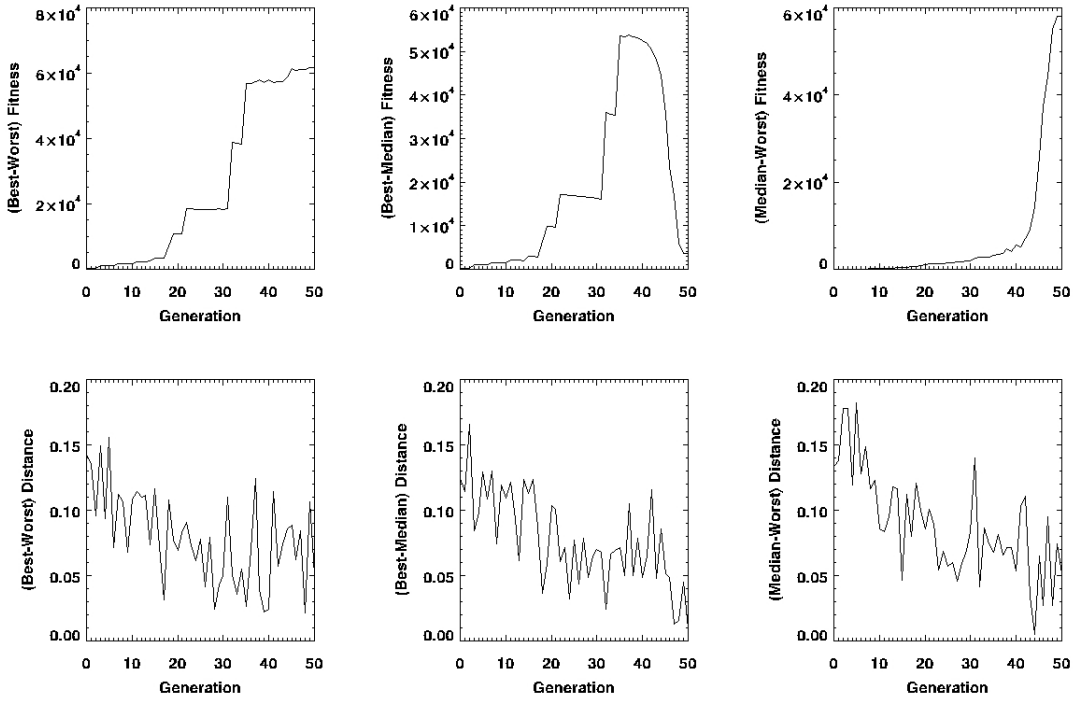

FIG. 4.15 Several differential measures of population diversity. The upper frames display fitness differential measures, while the lower frames present distance differential measures. *(Left):* The difference between the best and worst member of the population. *(Center):* The difference between the best and median member of the population, where median is defined as the individual with fitness ranking of $N_p/2$, where N_p is the population size. *(Right):* The difference between the median and worst individual. The difference between the best and worst fitnesses is very similar to the best fitness curve (see Figure 4.13), indicating that there are still some individuals with remarkably low fitness values. The fitness difference between best and median individual indicates that the genetic algorithm had found a high-fitness region early in the run, then at later times began to populate this region more heavily, as evidenced by the sudden drop at later generations. The overall downward trend of the distance measures again indicates that portions of the population were being drawn to the high-fitness regions, while the superimposed "noisiness" shows that some degree of genetic diversity is being maintained against this evolutionary pressure.

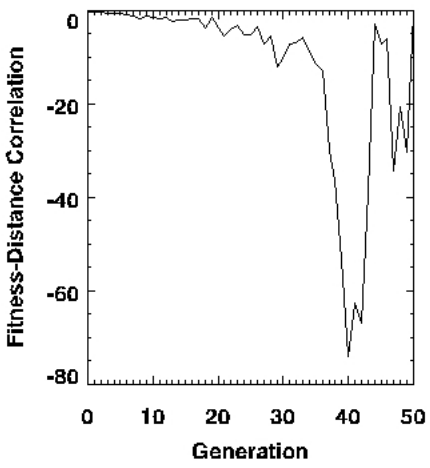

FIG. 4.16 The fitness-distance correlation measure, C_{FD}, as a function of generation. This indicates quite clearly that this type of optimization problem should present the genetic algorithm with no deceptive qualities. The initial value of this correlation measure, evaluated on the randomly-initialized population, was –0.19. This generally indicates that there is only one optimum of interest, in that there may be other local optima in the parameter space, but their peak values are less than that in the neighborhood around the true global optimum, which was populated by at least one member of the initial generation.

of the real data we use, so that the average performance of the genetic inversion will now be more representative of the type of results we can expect when inverting real datasets from (e.g.) the ASP. Table 4.5 shows the recovery statistics for the noisy cases. Figure 4.17 displays the quality of the profile fits to the noisy data. As can be seen from the figure, the Stokes I and V profiles are relatively insensitive to the noise, since they typically have signal strengths that are much greater than the observational uncertainty/noise. However, Stokes Q and U linear polarization profiles have signal strengths that are typically only a few percent of the Stokes I and V profiles, and are therefore much more sensitive to the presence of noise, as can be seen in the degradation of the quality of the fit for the Stokes Q and U profiles in the 5% and 10% noise cases.

Given the success of the synthetic profile recovery, we test the repeatability of the genetic algorithm on actual observational data. ASP data is used for this test, due to its higher

TABLE 4.5 Recovery of the known input parameters for the genetic inversion of noisy synthetic data. The averages and errors/uncertainties in the recovered model parameters are satisfactory, with the exception of the opacity ratio, η_0, in the 10% noise case, and the damping constant, a, in both the 5% and 10% noise cases. However, the magnetic parameters ($\|\vec{B}\|, \Psi, \phi, \alpha$) are all recovered with accuracies that parallel the no-noise results, as well as being within theoretical detection limits.

		1% Noise		5% Noise		10% Noise	
Parameter	Input	Average	Std. Dev.	Average	Std. Dev.	Average	Std. Dev.
$\|\vec{B}\|$ (G)	3000.0	3030.3	24.5	2976.5	21.0	3111.8	34.9
Ψ (deg)	45.0	44.6	0.52	44.9	0.40	41.9	0.71
ϕ (deg)	45.0	44.5	0.42	47.6	0.52	46.6	0.97
a	0.005	0.0056	0.0030	0.0031	0.0023	0.0085	0.0015
$\Delta\lambda_D$ (mÅ)	25.0	25.0	0.6	22.9	0.4	27.4	0.8
α	0.80	0.788	0.009	0.784	0.008	0.803	0.013
η_0	20.00	18.69	1.22	22.39	1.12	9.75	1.03

spectral resolution, and a pixel lying in the penumbral region of AR9240 was selected. As before, a total of 500 independent genetic inversions were performed on the spectral profiles from this pixel, with a population size of 100 individuals evolving over 50 generations. Table 4.6 shows the averages and standard deviations of the distribution of inversions, and Figure 4.18 shows a typical set of inferred line profiles compared to the observations.

Finally, we apply the genetic inversion to the full ASP field-of-view of AR9240. Several trial runs have been performed to assess the sensitivity of the genetic inversion to population size. Specifically, Figure 4.19 shows the magnetic field configuration (field strength, inclination, azimuth) inferred by the genetic inversion, for population sizes of 20, 50, 100, and 200, each evolving over 50 generations. It is evident that an increased population size reduces the amount of "noise" in the inferred parameters.

To assess the validity of the genetic inversion results, we compare them to the standard HAO one-component Milne-Eddington inversion code results. This particular inversion

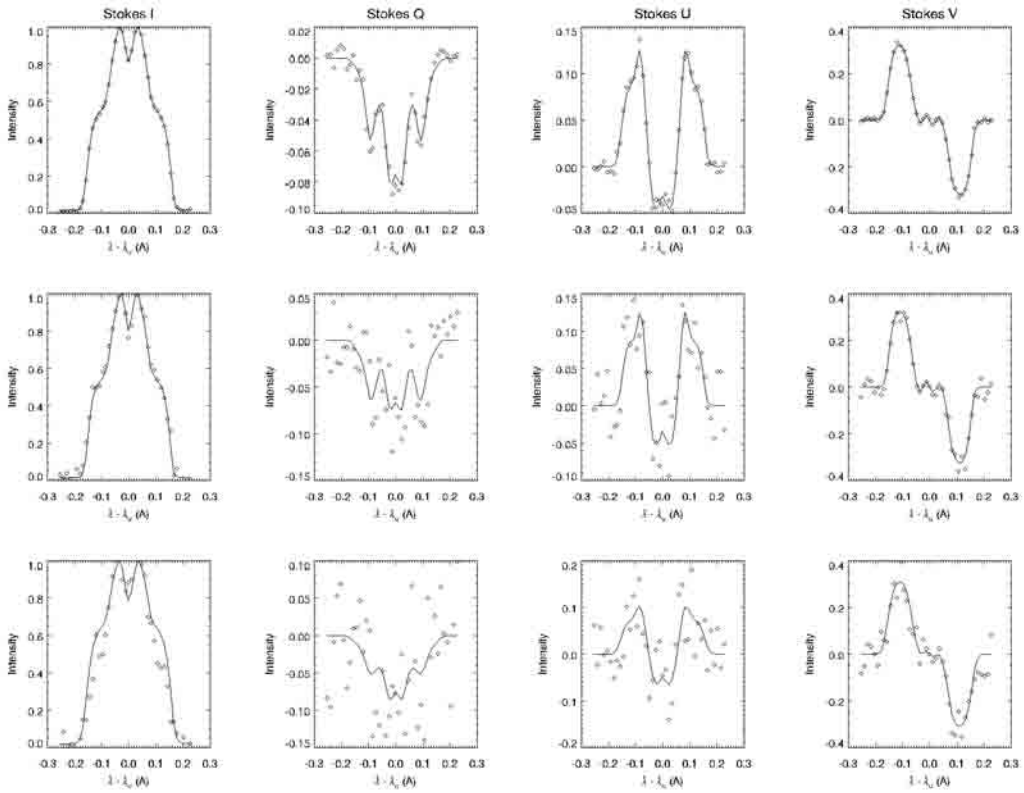

FIG. 4.17 The effect of noisy synthetic data on the genetic inversion. Shown are the Stokes I, Q, U, and V profiles inferred by the genetic algorithm (solid line) from the synthetic data (diamonds), with 1% (top row), 5% (middle row), and 10% (bottom row) Gaussian white noise superimposed on the synthetic profiles. The addition of noise has the most impact on the Stokes Q and U profiles, which already have low signal strength compared to the Stokes I and V profiles of typical real data. However, the recovered profiles are remarkably resilient to the presence of noise in the synthetic data.

code is not freely available, and is distributed only as a Sun executable for use on ASP data obtained at the National Solar Observatory. Figure 4.20 shows side-by-side comparisons of the HAO inversion and the genetic inversion. As can be seen from the figure, the overall structures are comparable, however there are some important differences. The smooth variation high field strengths from the center of the umbra outward present in the HAO inversion is also present in the genetic inversion, but to a lesser degree. This could be in part due to the low signal-to-noise ratio in the dark umbral regions, which introduces spurious variations in the intensity and polarization signals that are thwarting the genetic algorithm.

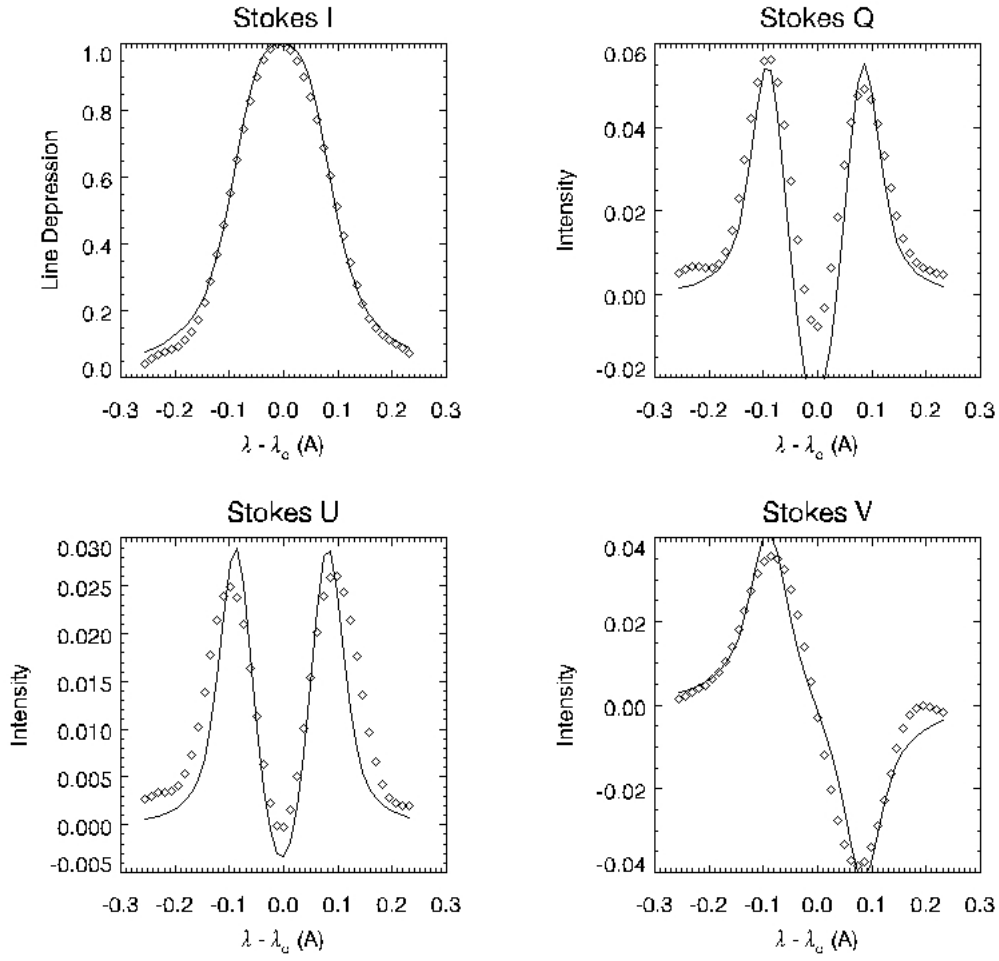

FIG. 4.18 Penumbral Stokes profiles inferred by the genetic inversion. The figure shows the average Stokes profiles inferred by the genetic inversion on a penumbral pixel for AR9240. The slight asymmetries present in real data somewhat degrade the quality of the fit.

TABLE 4.6 Recovery statistics for the genetic inversion on a sample ASP dataset.

Parameter	Average	Std. Deviation
$\lvert\vec{B}\rvert$ (G)	1242.7	104.6
Ψ (deg)	79.7	0.33
ϕ (deg)	9.62	0.27
a	0.416	0.022
$\Delta\lambda_D$ (mÅ)	38.1	0.9
α	0.913	0.048
η_0	11.73	0.65

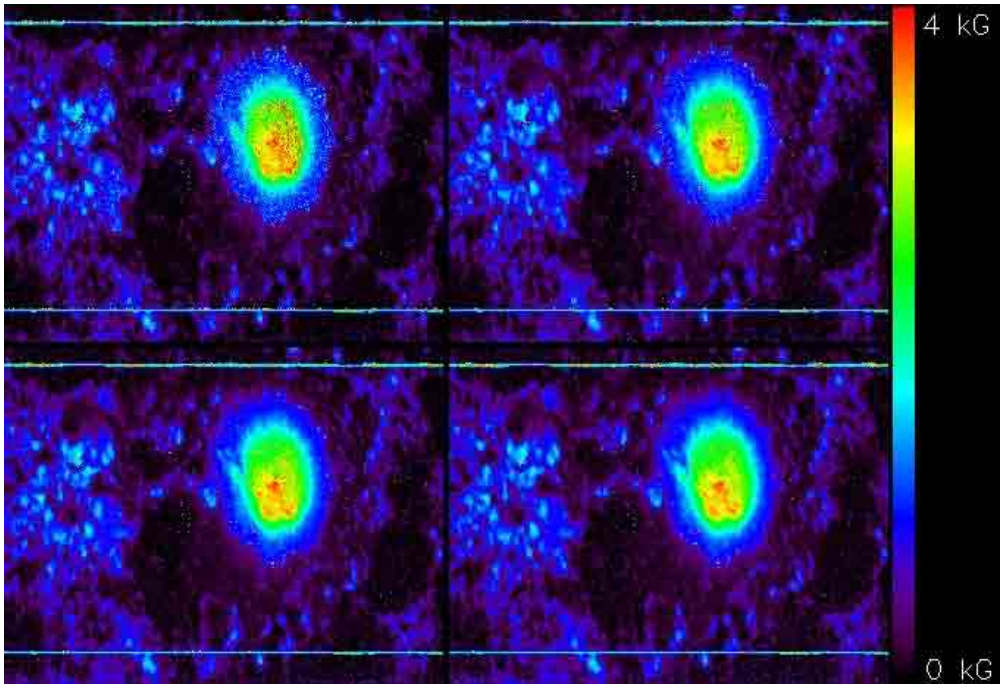

FIG. 4.19 The genetic inversion as a function of population size. This figure illustrates the effect of increasing the size of the population on the inversion results, for which the field strength is shown as a representative example. The population sizes are 20 (top left), 50 (top right), 100 (bottom left), and 200 (bottom right). As expected, increasing the size of the available gene pool promotes better exploration of the parameter space. The small population sizes recover the general structures, and therefore may be useful as "quick-look" techniques, but the required smoothness of the parameters is recovered only with larger populations. The noise in the recovered parameters is also monotonically decreased as the population size increases.

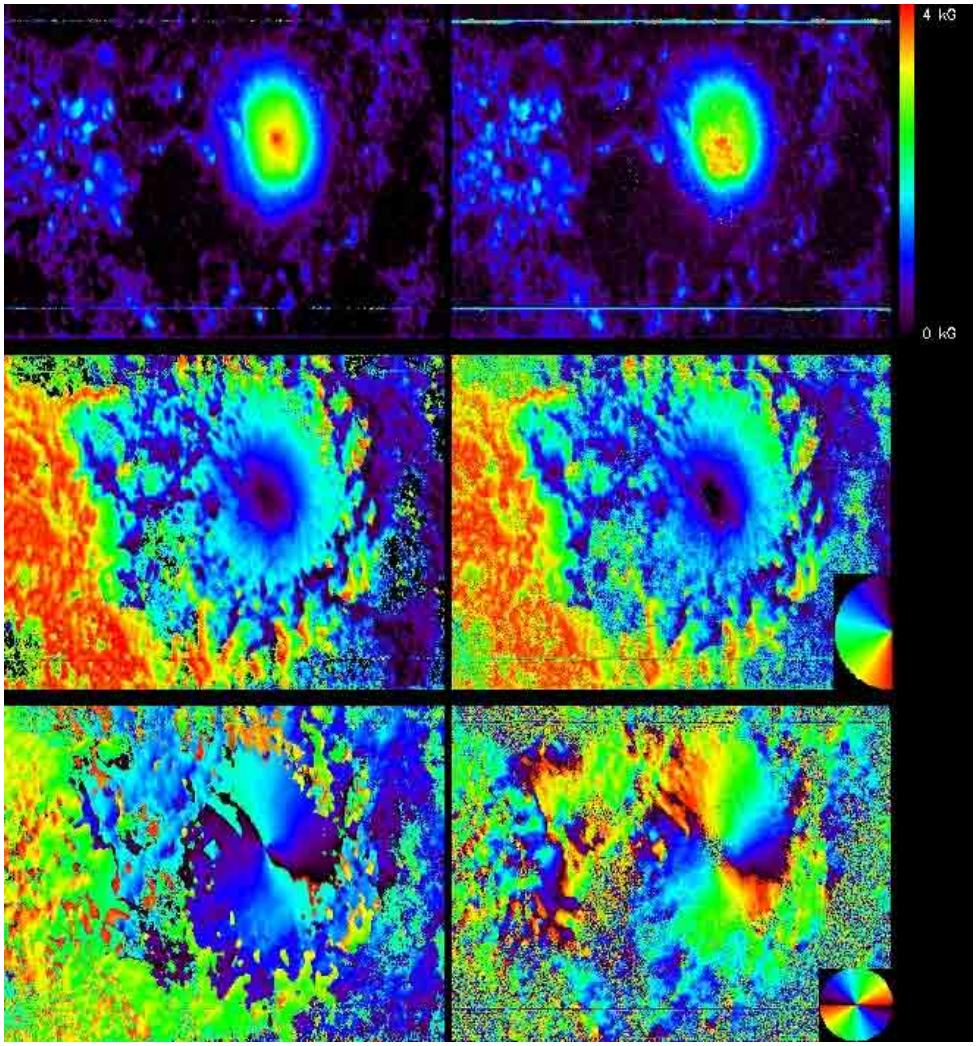

FIG. 4.20 Comparison between the HAO M-E inversion and the genetic M-E inversion. *(Left, top to bottom):* HAO M-E magnetic field strength, inclination, and azimuth. *(Right, top to bottom):* Genetic inversion magnetic field strength, inclination, and azimuth. The color wheel legend for the inclination maps indicate magnetic fields pointing out of the plane of the page by black, those pointing into the page by red, and horizontal fields by the aquamarine/green boundary. The wrap-around in the azimuth color wheel legend is indicative of the 180° ambiguity; the orientation of the field for a given color may either be that of the upper half or lower half of the color wheel. Although, the color table used for the azimuth plots is the same, the differences occur because the HAO inversion calculates the azimuth in the range $[-180°, 180°]$, but does not attempt to resolve the 180°-ambiguity. See Figure 4.24 for a more straightforward comparison of the azimuth values.

It may be possible to reduce this effect by increasing the population size or evolution time of the algorithm, or by introducing a post-processing step to fine-tune the results from the genetic algorithm. Contrarily, the HAO inversion seems to ignore the umbral substructure which is evident in intensity variations within the umbra itself. The inferred field strength seems to be "too smooth" (Balasubramaniam, K.S., personal communication), to the point of being non-physical, so that the differences between the HAO and genetic inversions seem to reinforce the idea that the genetic inversion is capable of resolving the finer structures in the active region. Furthermore, with the relative large size of the ASP pixel, there is ample room for multiple magnetic components, so such smooth variation really is not to be expected if the structures are averaged- or smoothed-over in some sense. The details of the field inclination structures are almost identical between the two inversion methods, as is the transverse orientation (azimuthal angle).

Figure 4.21 show the pixel-to-pixel variation in magnetic field strength for the HAO and genetic inversions, for three representative cross-sections through the penumbra, umbra, and quiet-sun. The HAO inversion is much smoother than the genetic inversion, although the latter itself does not display an inordinate amount of variation between neighboring pixels.

Figures 4.22 through 4.24 show scatterplots of the genetic inversion results plotted against the HAO inversion results. Only the points within the field-of-view with significant polarization degree have been included; as explained in a previous section, only pixels with a significant degree of polarization, p, are inverted to prevent the genetic algorithm from trying to fit a magnetically-sensitive absorption profile to data that clearly shows no signature of magnetism in the Stokes polarization profiles.

The correctness of the previous special cases along with the robustness of the blind iterations (i.e., small standard deviations) in the average inferred parameters leads me to believe that the method is as stable as the pseudo-random nature of the algorithm allows, and therefore can be used to reliably determine the structure of magnetic fields at the solar photospheric level. The next chapter extends the genetic inversion to the entire field-of-view

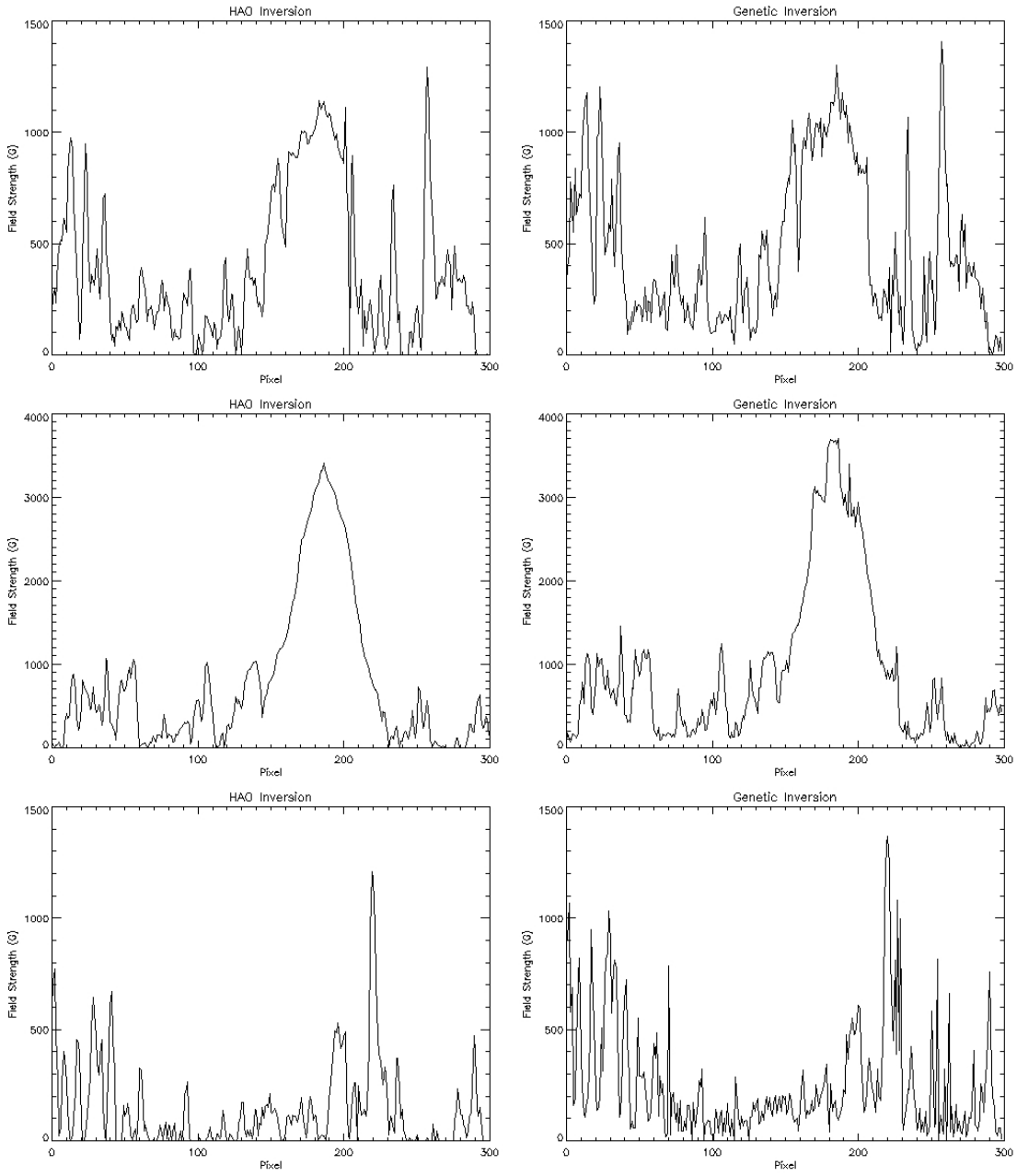

FIG. 4.21 Pixel-to-pixel variation in inferred field strength. The figure compares the results from the HAO inversion (left) and the genetic inversion (right), for the penumbra (top), umbra (middle), and quiet-sun (bottom). In most respects, the two are almost identical. The most notable, and perhaps most important difference is the lack of smoothness in the umbral region, present in the HAO inversion, but somewhat lacking in the genetic inversion. Again, this is most likely due to the low S/N in the umbra, which causes the intensity profile to contain spurious signal that does not correspond with any absorption mechanism. On the other hand, there is some structure to the umbral cores as seen in white-light details, so this may represent a more realistic magnetic configuration. Further analysis will be required.

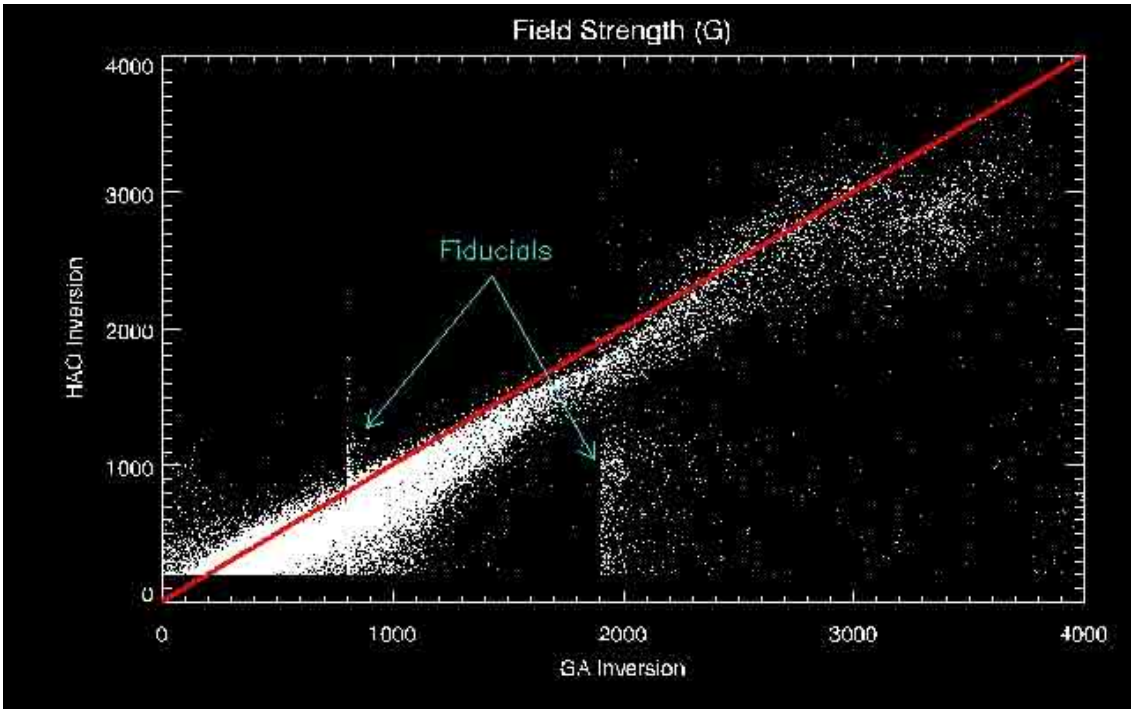

FIG. 4.22 Comparison between HAO and genetic inferred field strength. The red line denotes perfect agreement between the two methods. There is some scatter around this agreement line, but overall the results are consistent. The striated features at $\|\vec{B}\| \approx 800,\ 1900$ Gauss are due to the fiducial (guider) lines from the ASP.

of the Diffraction-Limited Spectropolarimeter (DLSP) as well as spectropolarimetric data from the Solar Optical Telescope aboard the *Hinode* satellite. We show that the magnetic structure of a solar active region, as well as its time-dependent dynamics, can be recovered with the methods used in this work.

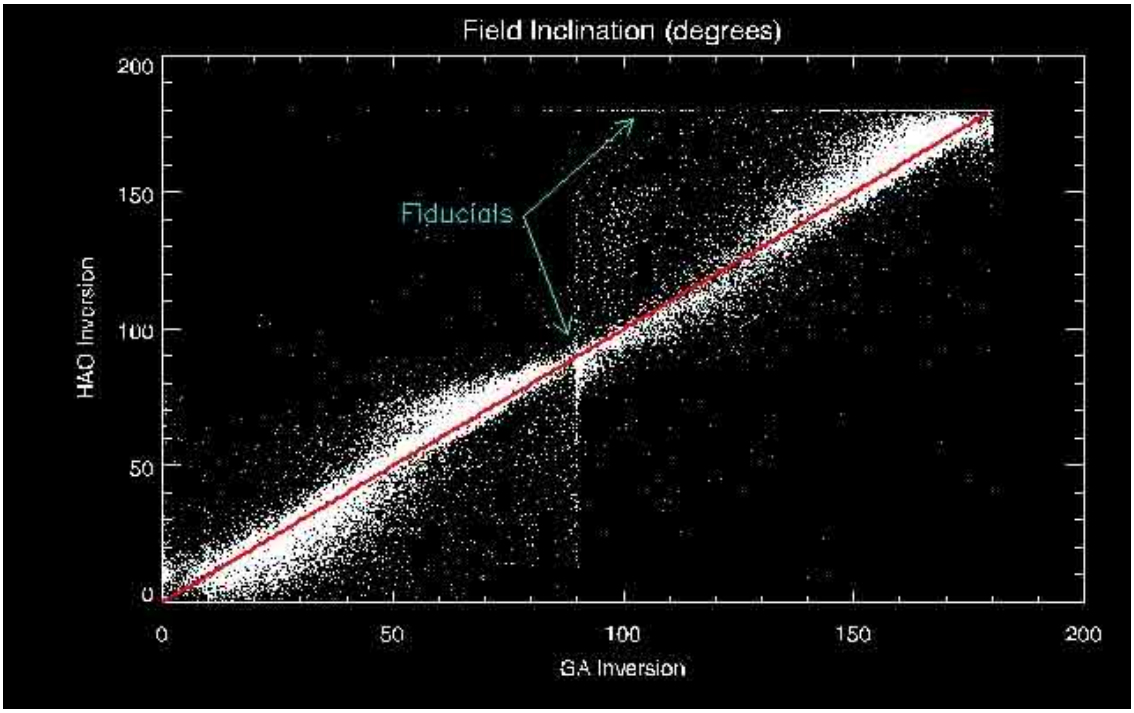

FIG. 4.23 Comparison between HAO and genetic inferred field inclination. Again, the agreement between the two methods is very tight, and the influence of the ASP fiducial lines is again noticeable.

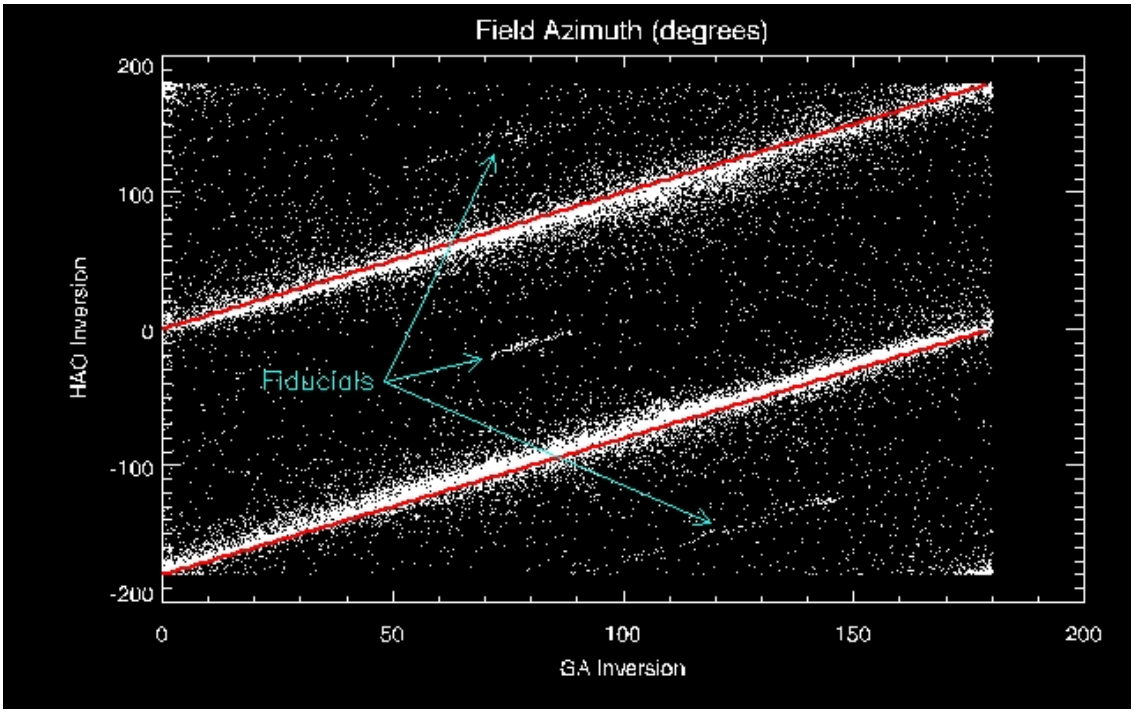

FIG. 4.24 Comparison between HAO and genetic inferred field azimuth. The double-banded nature of this plot is due to the azimuth range used by the HAO inversion ($\chi \in [-180°, 180°]$), while the genetic inversion restricts itself to the range $[0°, 180°]$ because of the 180°-ambiguity inherent in the Stokes inversion problem.

CHAPTER 5

THE VECTOR MAGNETIC FIELDS OF SUNSPOTS

Chapter 4 presented several test cases for which it was shown that the performance of the genetic inversion is indeed satisfactory. By comparing the genetic inversion results to those from an established inversion method, it was shown that the genetic inversion is capable of recovering the correct magnetic field strength and geometry. In this chapter, we extend the analysis to infer the magnetic and thermodynamic configuration of a sunspot observed with the Hinode satellite. The increased spatial resolution of Hinode allows us to discern greater detail, particularly in the penumbra, such that we may be more confident that a single magnetic component model of the Fe I line formation region is appropriate. For comparison, the angular resolution of each ASP pixel is roughly $0.525'' \times 0.37''$, giving a linear resolution of roughly 400 km by 270 km on the photospheric surface. The thin flux tube approximation (Parker [1974], Roberts and Webb [1978], Spruit and Roberts [1983]) states that photospheric magnetic fields are arranged in cylindrical *fluxtube* configurations embedded in a non-magnetic or weakly-magnetic plasma. Under this approximation, the diameter of a typical flux tube (magnetic element/component) is calculated to be < 200 km. Therefore, within the pixel resolution of the ASP, there could very easily be *four* separate magnetic flux tubes, such that the observed Stokes profiles result from the attenuation by four separate magnetic fields. The asymmetries present in such a configuration cannot be fit by a single component model. Compare this to the resolution of the Hinode SOT (about 0.2" diffraction limited at 630 nm), we obtain a linear size of about 150 km by 150 km per pixel. Therefore, a single component model should describe the Stokes profiles very well, given that the pixel size is comparable to, but somewhat less than, the size of individual magnetic elements at the photospheric level. The results from the inversion of Hinode data is now presented.

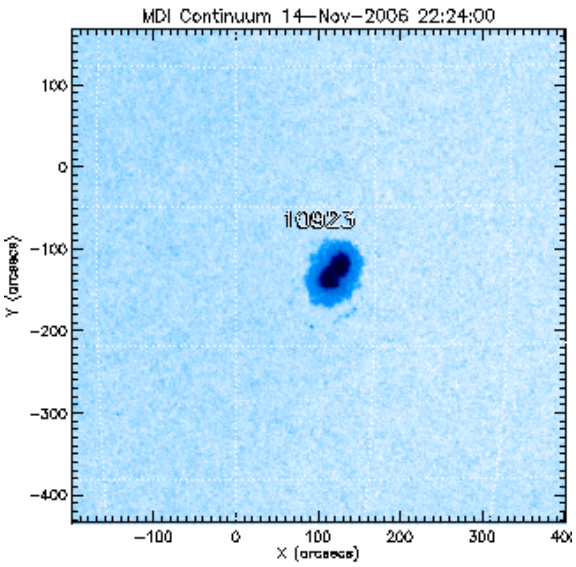

FIG. 5.1 A SOHO MDI continuum image of AR10923. The figure shows AR10923 roughly 12 hours after being observed by the Hinode satellite. Extrapolating the movement of the active region backwards shows that during the Hinode observation window, the sunspot was very close to disc-center, roughly centered on the zero-longitude meridian with a solar latitude of $\sim$ few degrees south.

5.1 Hinode Data—AR10923

As mentioned in Chapter 3, NOAA AR10923 was observed by Hinode between 10 November 2006 and 19 November 2006, during which time the sunspot traversed the solar disc from the solar east limb to the solar west limb. To avoid projection effects, we first focus on the data obtained on 14 November 2006, when the sunspot was very close to disc-center, such that the local solar normal is coincident with the line-of-sight. Figure 5.1 shows a continuum image obtained by SOHO MDI, approximately 15 hours after Hinode observed the sunspot. The formation of a light bridge is apparent in this image.

Figure 5.2 shows the sunspot in white-light, as well as in polarization intensity at the photospheric level, and Figure 5.3 shows several X-ray wavelength images, detailing the magnetic structure above the photosphere. These loop arcade structures are governed by the magnetic configuration at their photospheric footpoints, so the inference of the vector magnetic field at the photospheric level is key to understanding how such structures

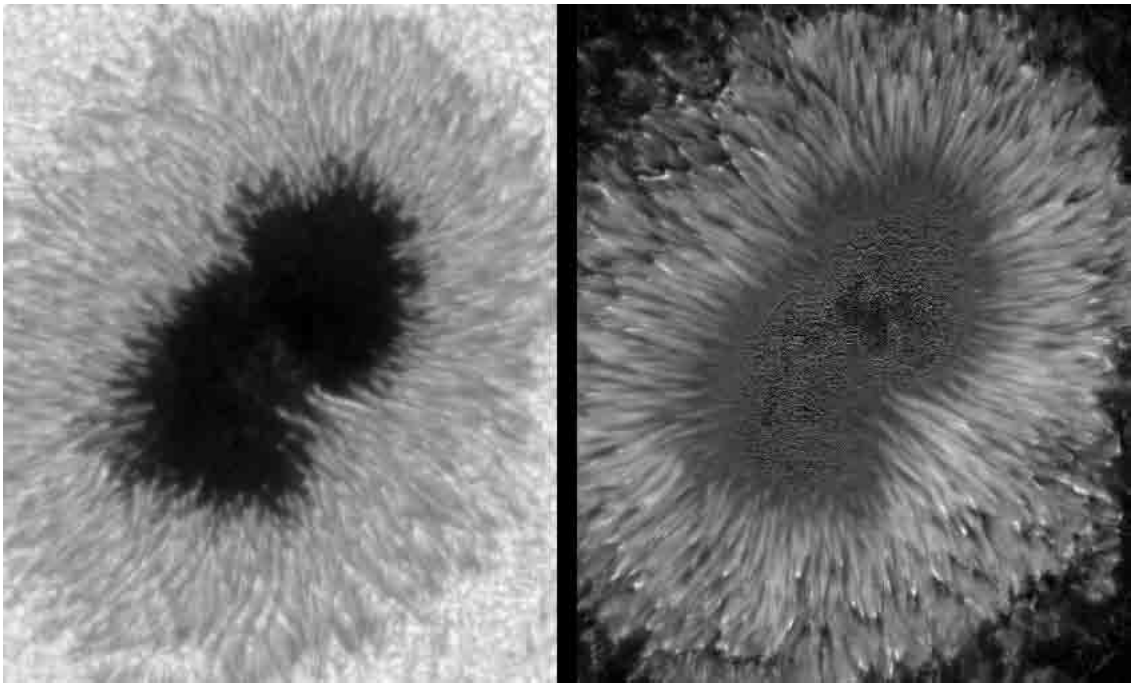

FIG. 5.2 Total polarization in AR10923. The figure displays NOAA AR10923 as observed by Hinode in continuum white-light (left frame), as well as in its fractional degree of polarization, $p \equiv \sqrt{Q^2 + U^2 + V^2}/I$ (right frame). In the right frame, bright areas correspond to regions of high-field strength, horizontal or vertical field inclination, or some combination thereof.

are formed and how they evolve. The field-of-view of this sunspot is 44.4 Mm along the horizontal image direction and 59.0 Mm along the vertical direction. For comparison, the diameter of the Earth is roughly 12.72 Mm.

5.2 Inversion Results

In this section we present the results of the genetic inversion of NOAA AR10923. We limit ourselves to the discussion of the field strength, inclination, azimuth, fill-fraction, and Doppler width returned by the genetic inversion. The remaining parameters, namely the damping constant and opacity ratio, are of minimal interest to the structure of the active region and are therefore ignored in what follows. Figure 5.2 displayed a white-light continuum image of AR10923, and from this image, we can make the following few simple observations outlining the results we expect to see with the high-resolution data:

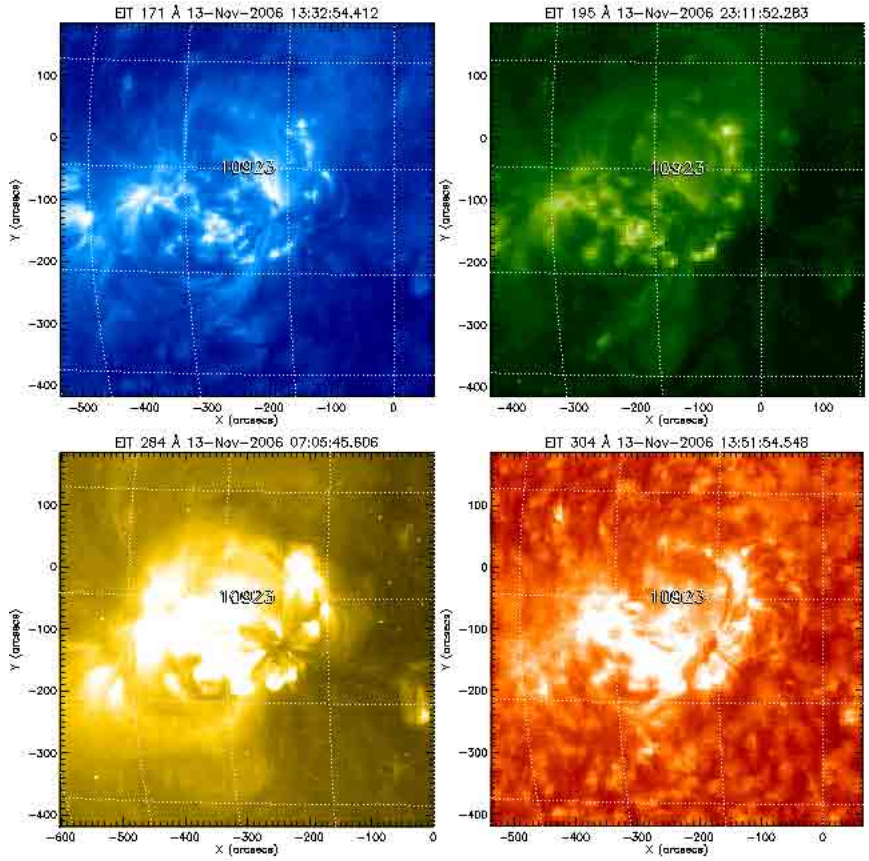

FIG. 5.3 Several X-ray observations of AR10923. NOAA AR10923 was observed by SOHO EIT (reproduced from *solarmonitor.org*), detailing the magnetic arcade structures in the corona that originate from the photospheric sunspot fields below.

- The umbra of a sunspot is dark because the presence of magnetic field inhibits convection. The darkest regions of the umbra should therefore contain the highest field strengths, since strong magnetic fields are more efficient at inhibiting the solar convection from bringing bright, hot subsurface material up to the surface.

- With the same rationale, the magnetic fill-fraction should obtain values close to 1.0 in the entire umbra, where the strongest fields are located.

- Since the sunspot was close to disc-center during the observation, the line-of-sight corresponds to the local solar normal, such that the umbral fields should be relatively uniformly directed away from the observer ($\Psi \approx 180°$), based on the order of the

positive- and negative lobes of the Stokes V profiles.

- The increased spatial resolution allows the penumbral fine structure to be resolved; the more horizontally oriented fields should be resolved separately from the more vertical fields in which they are embedded. This is the so-called uncombed model of the penumbra.

- The more-horizontal fields of the penumbra are thought to guide the hot upflow channels of the Evershed effect, so the thermal Doppler widths near the umbra-penumbra boundary should show a similar stratification to the inclination angles.

- The sunspot has a high-degree of radial symmetry, so correspondingly radial field directions should be recovered.

Figure 5.4 shows the magnetic field strength inferred by the genetic algorithm. The red contour outlines the darkest parts of the umbra, as seen in continuum wavelengths. The figure shows the highest field strengths are indeed found to lie within the darkest regions of the umbra, where the expected high field strengths are most effective at inhibiting convection from bringing hot, bright, subsurface plasma to the surface proper.

Figure 5.5 displays the magnetic field inclination inferred by the genetic algorithm, with the convention that black corresponds to vertical fields emerging from the surface, toward the observer, and white represents vertical fields plunging into the surface, away from the observer. The red contour surrounds the entire umbral region, as determined internally by the genetic inversion, from a comparison to the continuum intensity far from the absorption line. As can be seen, the umbral fields are all nearly perfectly vertically-oriented away from the observer, with little pixel-to-pixel variation. Also evident in the figure are several moving magnetic features (MMFs) of polarity opposite to the that of the sunspot. These MMFs typically occur in the outer penumbral filaments, where a portion of a penumbral fluxtube has plunged back below the surface, leading to the so-called "sea-serpent" model of MMFs.

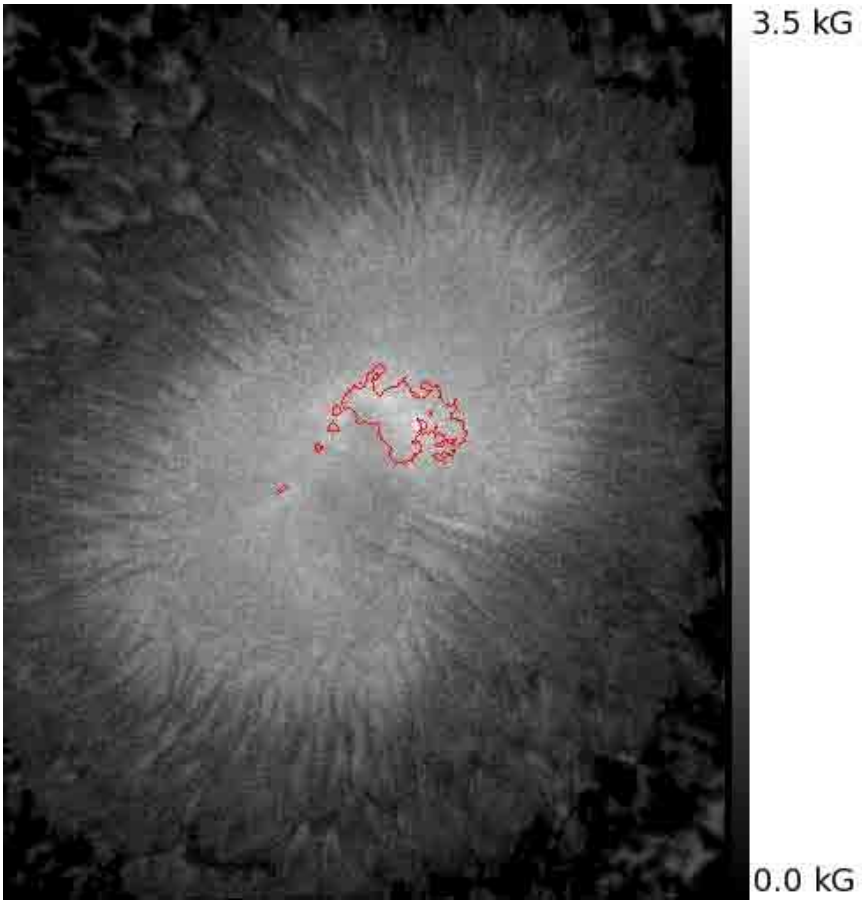

FIG. 5.4 Magnetic field strength in AR10923. The penumbral protrusion into the umbra (see Figure 5.2, center region) contains lower field strengths than in the umbra itself, in agreement with expectations. The general variation from high field strengths in the umbral regions to lower field strengths as one moves outward is nicely recovered.

Figure 5.6 shows the magnetic field azimuth (horizontal field angle) inferred by the genetic algorithm. Because this sunspot has a relatively simple, circularly-symmetric structure, the expected radial orientation of the penumbral fields is recovered. In the umbral region, the nearly perfectly-vertical fields do not have a well-defined azimuthal angle, and as such, the linear polarization signals in this region are lost below the noise level of the observations. Therefore, the field orientation in the umbra is composed of the "static" seen in the figure. Because of the 180° ambiguity in the Stokes inversion problem, the recovered azimuthal angle has a "wrap-around" effect at 0° and 180°.

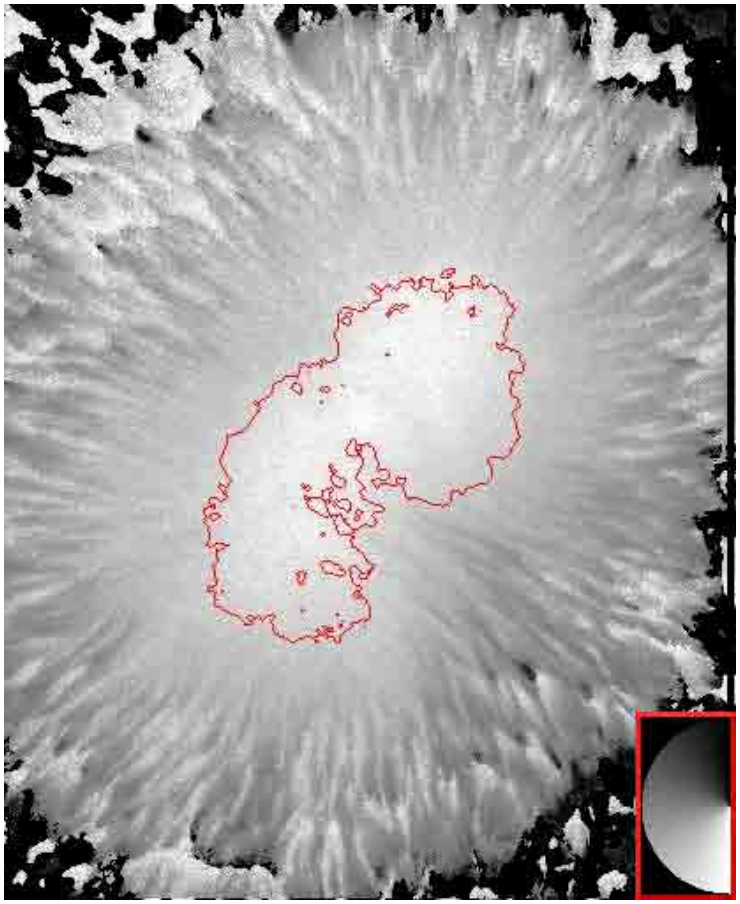

FIG. 5.5 Magnetic field inclination in AR10923. Since the sunspot was very near disc-center at the time of the observations, one expects that the umbral fields will be almost perfectly vertical, in this case they have negative polarity and are directed into the plane of the page.

Figure 5.7 displays the magnetic filling-fraction, defined as the fractional pixel area occupied by the magnetic field. The filling-fraction is used to mix the Stokes I profiles, generating a superposition of Stokes I from magnetic and non-magnetic plasmas. As expected, in the umbral regions where high-field strengths exist, we expect high filling-fractions (close to 1.0). Also, notice the decrease and subsequent increase as one moves radially outward from the umbral center to the umbra/penumbra boundary to the inner penumbra. This is a result of the penumbral fields being "offshoots" from the main core of the vertical umbral fields (see Figure 1.21).

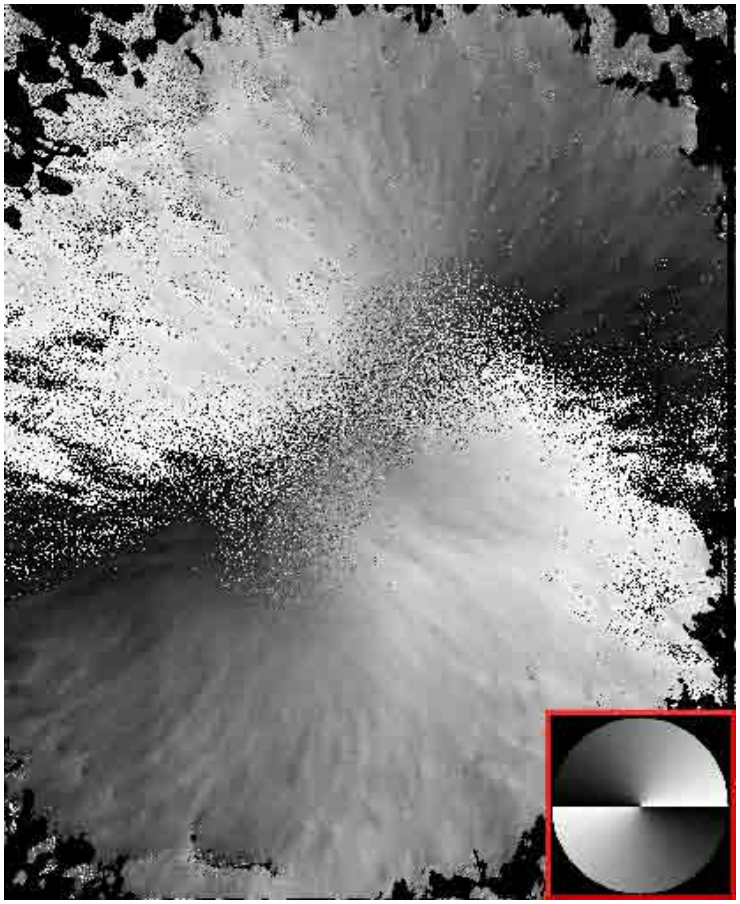

FIG. 5.6 Magnetic field azimuth in AR10923. The radial structures of the penumbra are well-recovered by the genetic inversion. Note that since the umbral fields are in a nearly-perfect vertical direction, the azimuthal angle is not well-defined here, hence the "static" in the umbral regions.

Figure 5.8 shows the Doppler line-width inferred by the genetic inversion. This parameter represents the broadening of the absorption line due to the thermal motions of the plasma, and takes larger values for higher temperatures. Therefore, the Doppler width is larger in the hotter penumbra than in the cooler umbra, as is evident from the figure. Also, traces of a striated structure at the umbra/penumbra boundary are evident, in agreement with the "uncombed" penumbral model of Solanki and Montavon (1993). The Evershed flow, expected to be products of hot, upflow channels in the inner penumbra which turn into radial outflow channels in the outer penumbra is apparent in this image of Doppler line-width. Because of the dependence on temperature, the hot upflow channels should have

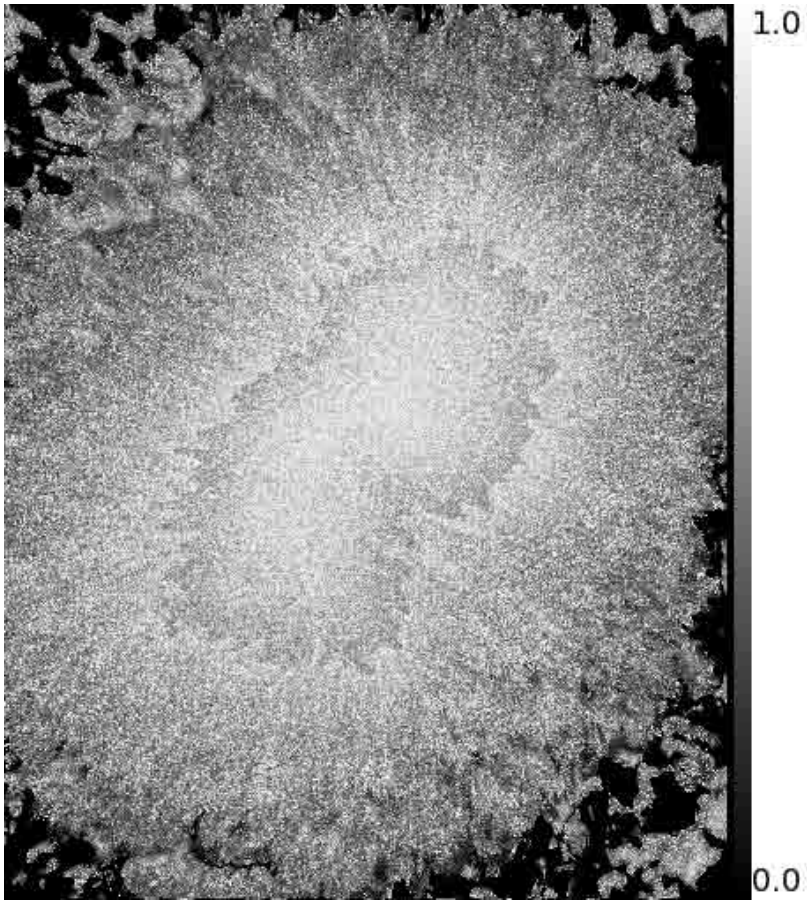

FIG. 5.7 Magnetic fill-fraction in AR10923. The figure shows the magnetic fill-fraction, defined as the fractional pixel area occupied by the magnetic field, inferred by the genetic inversion. As expected, fill-fractions of order unity are found in the umbral region, where the strongest fields are most likely to exist. Similarly, high fill-fractions are recovered near the umbra-penumbra boundary. This is the result of the penumbral field geometry, in which the fluxtubes that have split away from the main "trunk" of the umbra overarch each other. Therefore, the line-of-sight to these regions passes through many superimposed, almost horizontal fields.

a larger Doppler line-width than the surrounding quiescent background, which is evident in the correlation between the higher Doppler widths and more highly-inclined magnetic fields of the inner penumbra.

Shown in Figures 5.9 and 5.10 are typical penumbral and umbral line profile fits to the spectropolarimetric observations, from the center-side of the sunspot. The profiles fits are typically slightly worse in the dark umbral regions, where the signal-to-noise ratio is low.

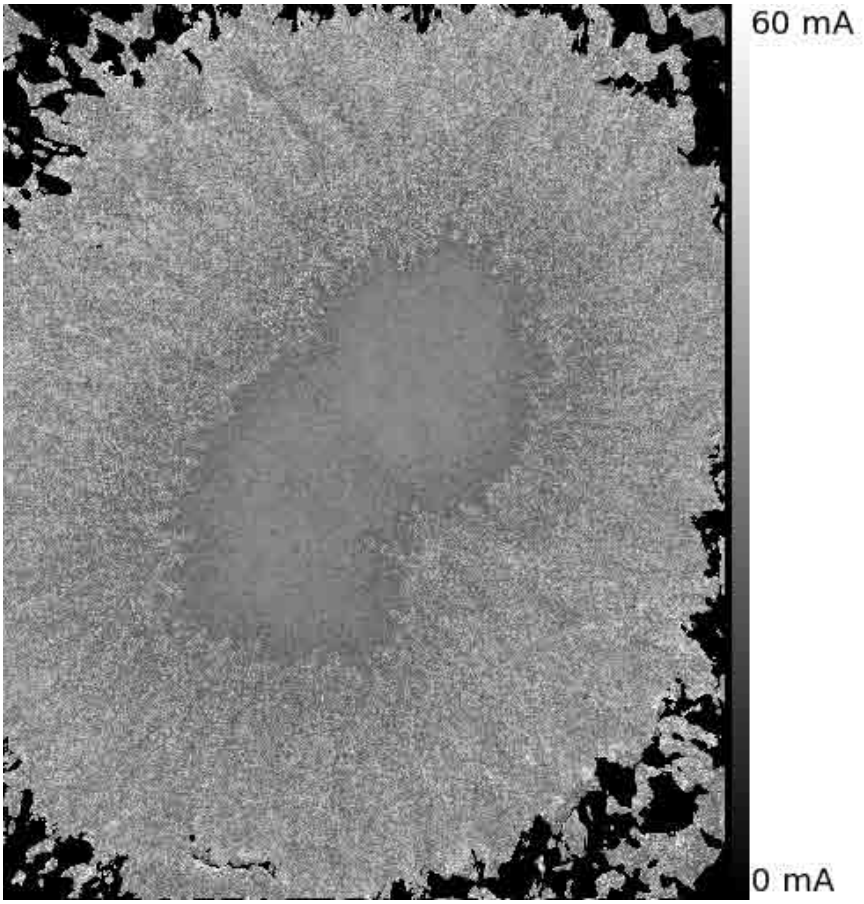

FIG. 5.8 Doppler line-width in AR10923. As expected, the smallest line-widths occur in the umbra, where the magnetic field inhibits convection. As a consequence, no convective overturning occurs, and hot subsurface material is not transported to the surface. Therefore, the temperatures are lower in the umbra, yielding smaller Doppler widths. The filamentary structure of the penumbra is also clearly visible, which contains hot upflow channels embedded in cooler, stationary field regions.

While the weighting scheme used in the genetic inversion to quantify the deviation between the model and the observations assures that deviations in each Stokes profile contribute roughly equally to the χ^2 measure, it does nothing to prevent the noise in the umbral observations from potentially misleading the genetic search. This is evident in the umbral profile fits to Stokes I, where the line-wings of the observations deviate significantly from the continuum background.

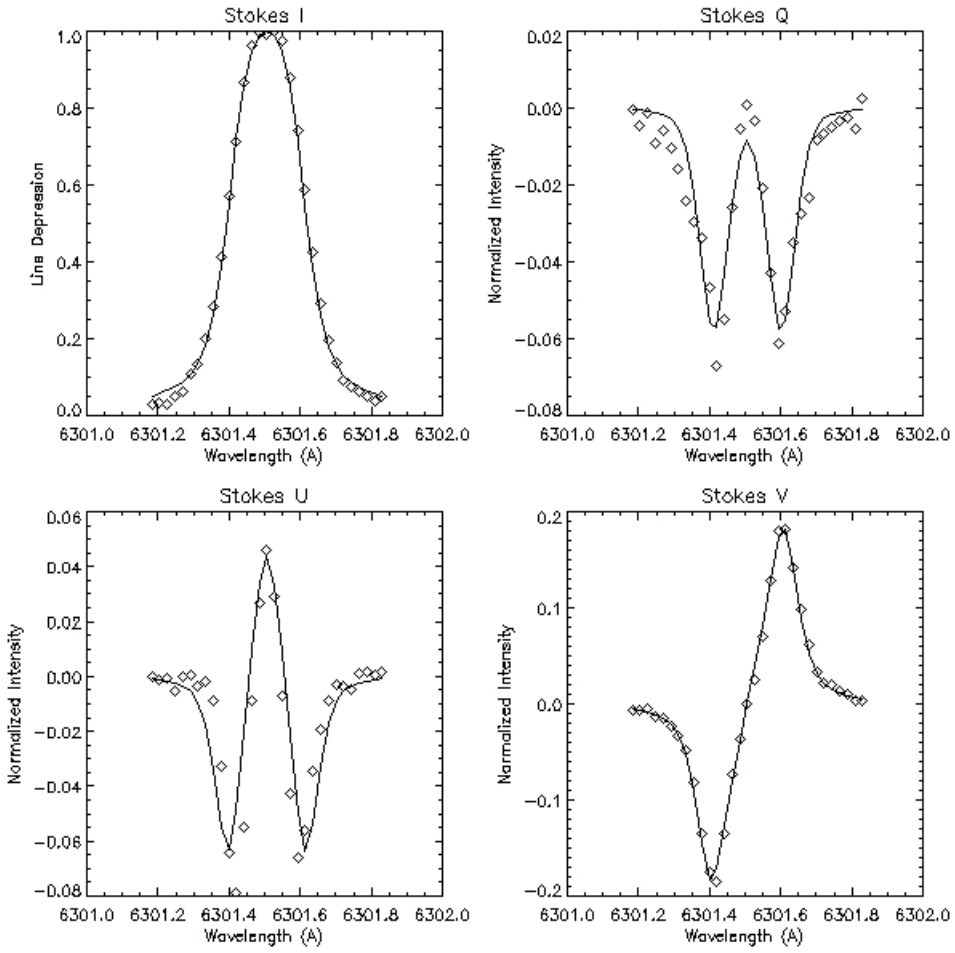

FIG. 5.9 Typical spectral line fits in the penumbra of AR10923. The figure shows the optimal line profiles inferred by the genetic algorithm for a point within the penumbra of AR10923. All four Stokes profiles are well-fit by the genetic algorithm. There are slight asymmetries in the peak values of the Stokes Q and U linear polarization signals, most likely due to the presence of gradients in field strength and/or line-of-sight velocity along the line-of-sight. These gradients are not incorporated into the inversion procedure, and therefore it is unable to fit the asymmetries.

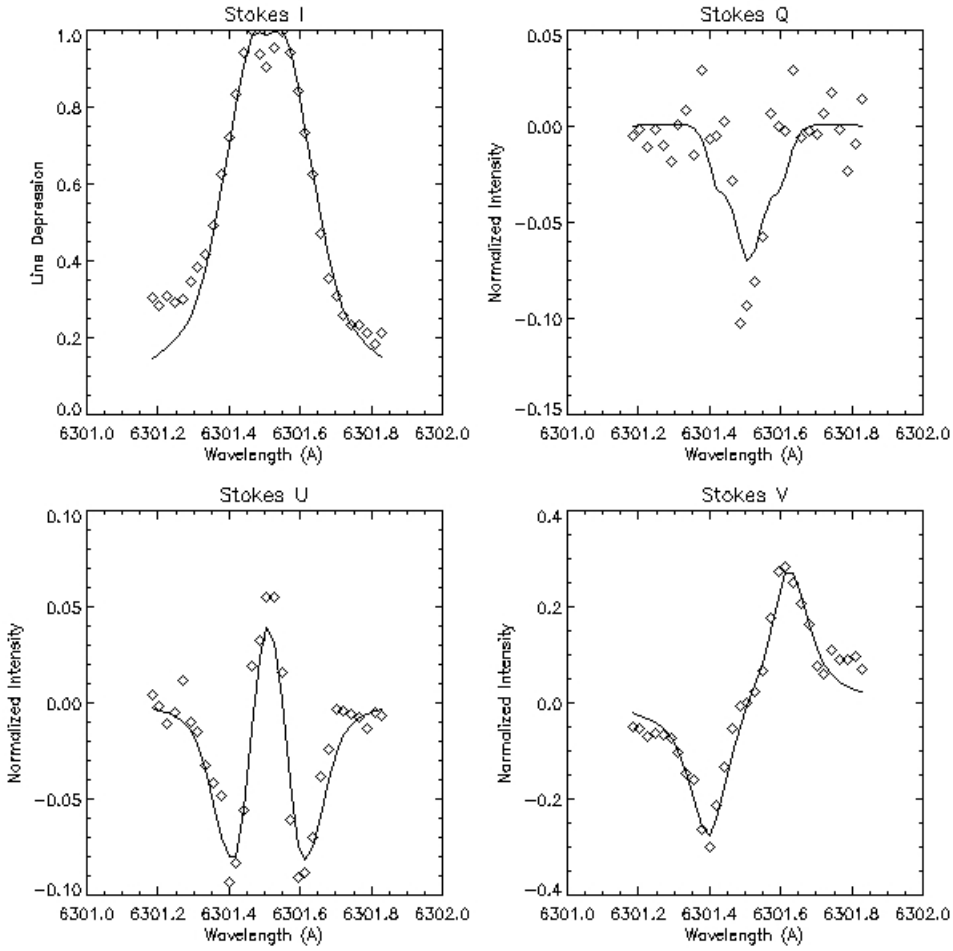

FIG. 5.10 Typical spectral line fits in the umbra of AR10923. The figure displays the optimal line profiles at a point within the center-side umbra of AR10923. The linear polarization signals are expectedly noisier in the umbra, where the azimuth is not clearly defined and the S/N is much lower than in the surrounding structures. The circular polarization also has a larger noise component, but the typical double-lobed structure is still well recovered.

5.2.1 Ambiguity Resolution

The 180°-ambiguity is a problem inherent in any Stokes inversion scheme that infers magnetic field geometries from polarization signals. From the Unno-Rachkovsky solutions to the PRTE, one can see the dependence of the linear polarization elemental profiles $\eta_{Q,U}$ and $\rho_{Q,U}$, on the field azimuth goes as:

$$\eta_Q,\ \rho_Q\ \propto cos2\chi \tag{5.1}$$

$$\eta_U,\ \rho_U\ \propto sin2\chi. \tag{5.2}$$

Therefore, the transformation $\chi \rightarrow \chi\ +\ 180^\circ$ yields:

$$\eta_Q,\ \rho_Q\ \propto cos2(\chi\ +\ 180^\circ) = cos2\chi \tag{5.3}$$

$$\eta_U,\ \rho_U\ \propto sin2(\chi\ +\ 180^\circ) = sin2\chi, \tag{5.4}$$

which implies that the transverse component of a magnetic field modulates the linear polarization signals in exactly the same fashion as a field which points in the opposite transverse direction. Therefore, the inversion algorithm cannot distinguish between a magnetic field oriented at 45°, for example, and one oriented at 225°. Therefore, the genetic inversion is restricted to an azimuth range of $[0^\circ, 180^\circ]$, and the resolution of the ambiguity is left as a post-processing step. Within the boundaries of the penumbra of any sunspot, the assumption of radial magnetic fields is quite good, as can clearly be seen in a continuum (white-light) image. Therefore, within the sunspot boundaries, we resolve the ambiguity by choosing the field orientation which has the better agreement with a radial field centered at the peak-flux location of the umbra. Figure 5.11 shows the disambiguated azimuth for AR10923. While this method works well within a sunspot, outside the boundaries of the penumbra the method is guaranteed to fail. In a general active region containing more than one polarity sunspot, the field in the intermediate regions may be highly twisted and sheared, so that in general, no radial solution can be obtained. In this case, the ambiguity is resolved by making some assumptions about the presence of electric currents in the active

region and comparing the active region field with a synthetically generated potential field, using the line-of-sight magnetic field as a boundary condition (see Metcalf et al. (2006) for an extensive overview of existing algorithms). With the disambiguated field, Figure 5.12 shows the longitudinal (line-of-sight) magnetic field in grayscale overlaid by a vector field representing the transverse field.

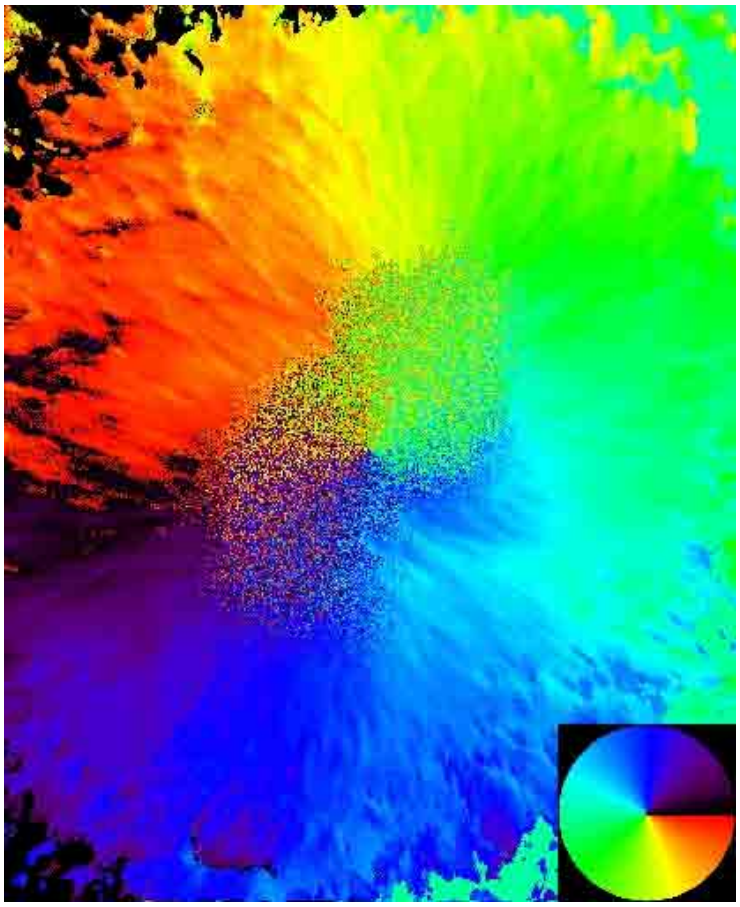

FIG. 5.11 Disambiguated magnetic field azimuth in AR10923. Shown is the magnetic field azimuth for AR10923 after the 180°-ambiguity is resolved. The negative polarity of this spot is evident from the direction of the penumbral magnetic fields, which point in the direction indicated by the color wheel in the lower right of the image

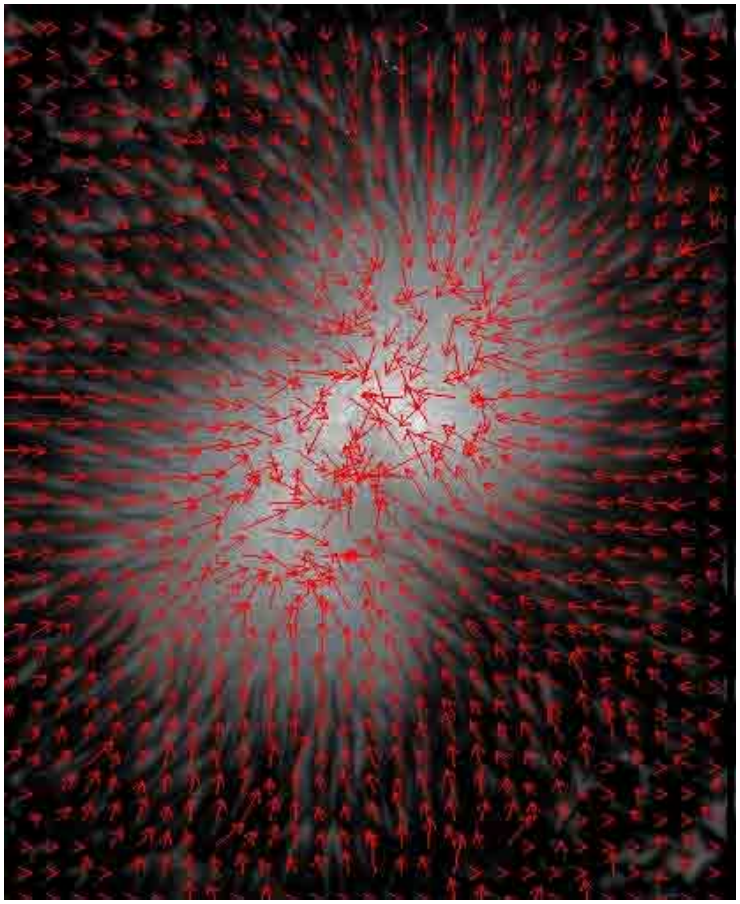

FIG. 5.12 Vector-field representation of the magnetic field in AR10923. The line-of-sight flux in AR10923 is shown in grayscale, overlaid by arrows representing the transverse magnetic field. The length of an arrow is proportional to the field strength at its tail. As can be seen, the field direction correlates very well with the "spiny" structure of the penumbral filaments.

5.2.2 The Penumbra: A Closer Look

Spectropolarimetric data from the *Hinode* satellite has an unprecedented spatial resolution, allowing one to observe the penumbral fine structure at a level never-before seen in ground-based observations. Kitai et al. (2007), Ichimoto et al. (2007), and Katsukawa et al. (2008) have observed and analyzed the structure of penumbral filaments, umbral dots, and moving magnetic features in photospheric active regions, illuminating aspects of the penumbral fine structure and Evershed effect and raising more questions about their origin and evolution. It is widely held that the structure of the penumbra follows an "uncombed"

geometry, whereby more horizontal penumbral magnetic fields are embedded in more vertical background magnetic fields. Furthermore, Rimmele and Marino (2006) have found that the Evershed flow (the radial outflow of photospheric plasma in the penumbra) seems to originate in the bright, inner footpoints of penumbral filaments and is guided outward by the more horizontal component of the uncombed penumbra. This allows us to perform even more tests on the accuracy and physicality of the genetic inversion on this high-resolution *Hinode* data from AR10923. Figure 5.13 shows the correlation between continuum intensity and field inclination in the penumbra. The hot upflows of the Evershed effect have a higher thermal intensity, and these should correlate well with the horizontal fields. Conversely, the more highly-inclined field are representative of the background component of the uncombed penumbra, and should therefore correspond to the darker penumbral filaments. The figure shows exactly this behavior.

To further illustrate this point, Figure 5.14 shows an elliptical path through the penumbra, around which Figure 5.15 displays the continuum intensity and inclination. In the figures, several vertical markers have been placed to identify regions of higher continuum intensity (and hence higher temperature) and their correlations with the more horizontally-inclined fields. The Evershed flow originates at the inner-side of the brighter penumbral filaments, where hot plasma is rising to the surface and is subsequently guided into a radial outflow.

Finally, the Evershed effect is most apparent in a Dopplergram, which shows the line-of-sight component of the plasma velocity based on the observed shift, $\Delta\lambda$ of the line-center wavelength from the laboratory (rest) wavelength, given by:

$$v_{LOS} = \frac{c}{\lambda_0}\Delta\lambda. \tag{5.5}$$

Figure 5.16 shows the LOS velocities derived in this manner from the *Hinode* observations. Black indicates upflows along the line-of-sight, while white indicates downflows.

Figure 5.17 shows the geometry of the line-of-sight projection of the Evershed effect, and how it depends on the observation angle. For active regions at disc-center, the difference

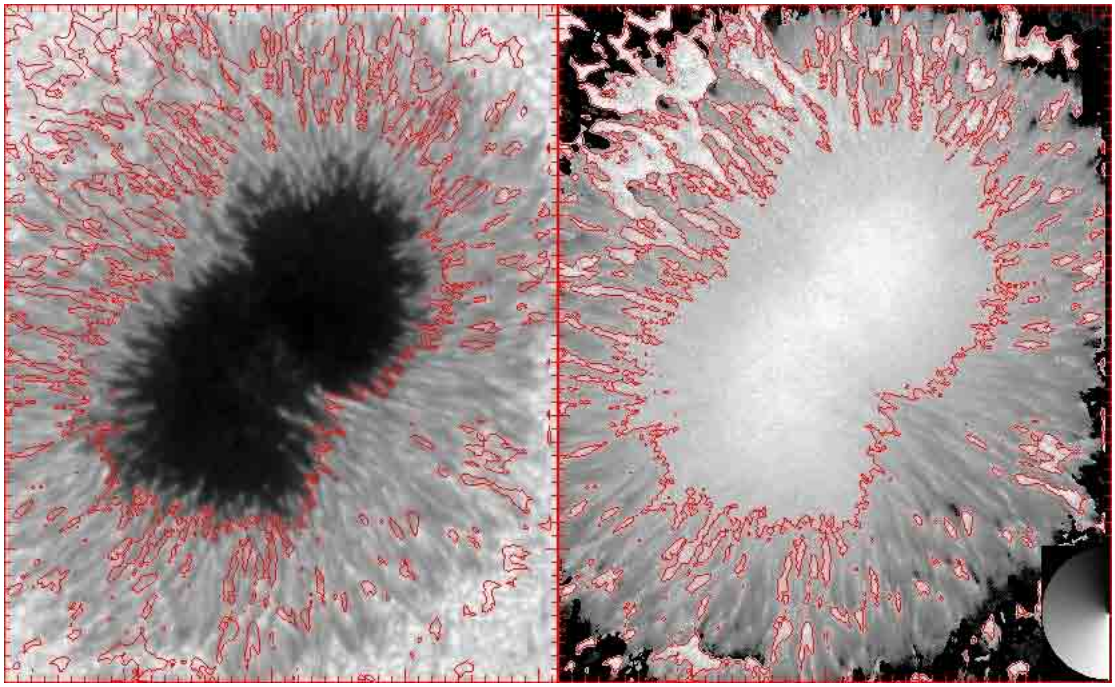

FIG. 5.13 Correlation between brightness and magnetic field inclination. The figure shows the structural correlation between the continuum intensity (left) in the penumbra and the field inclination inferred by the genetic inversion (right). The contours border regions with inclinations of at least 130°. As can be seen in the continuum image, these more vertical regions typically correspond to the darker penumbral filaments. These darker filaments are the background in which the more horizontal penumbral filaments guide the hot Evershed outflow. The recovery of this correlation is good evidence in favor of the physicality of the genetic inversion when used with very high spatial resolution data.

between the disc- and limb-side penumbra is negligible, and the Evershed effect is not observed.

5.3 Derivable Quantities of Physical Interest

While the genetic inversion yields the magnetic field configuration and various thermodynamic parameters of the photospheric plasma, we are not limited to these diagnostics. The most useful parameters for analyzing the non-potentiality or flaring activity/potential of an active region must be derived from the baseline magnetic field. This section presents the most important and relevant of these derivable quantities, which can be of use in diagnosing the complexity of an active region or sunspot.

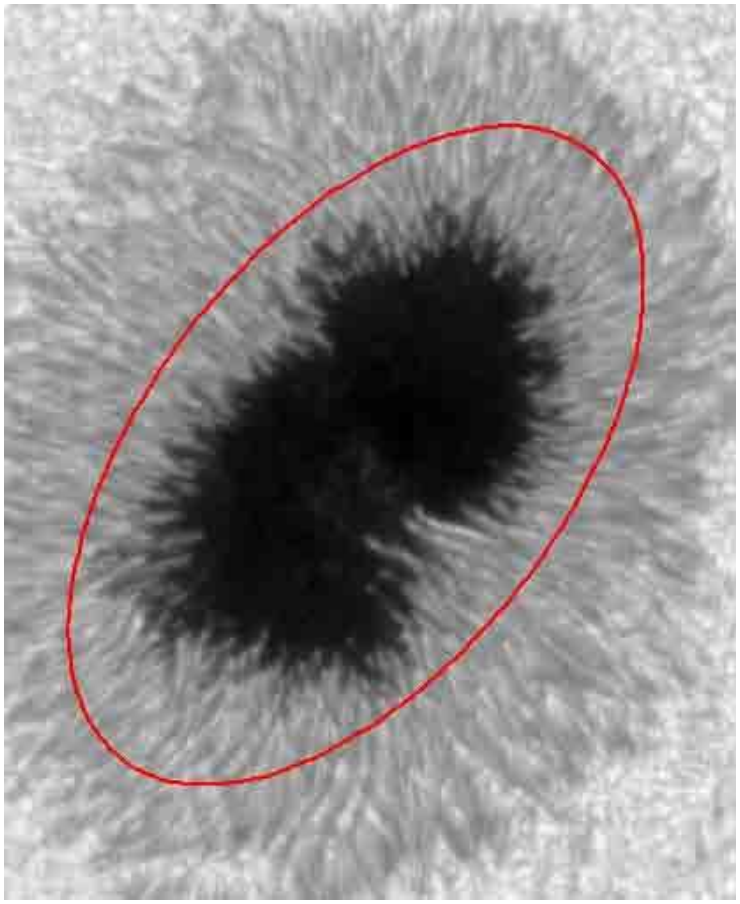

FIG. 5.14 A path around the penumbra of AR10923. This azimuthal path around the penumbra, cutting through bright and dark filaments, is where a comparison between the brightness and inclination of the filaments can be made.

5.3.1 Longitudinal Flux

The longitudinal (line-of-sight) magnetic flux is calculated as

$$\Phi = \vec{B} \cdot \vec{A} = B_z A \tag{5.6}$$

and is shown in Figure 5.18. The vector $\vec{A}$ is an area vector with direction parallel to the local solar normal, and magnitude equal to the pixel area ($A = dxdy$). This is the quantity traditionally measured by longitudinal magnetograms, which calculate the flux as directly proportional to the wavelength separation of the peaks of the positive and negative lobes in the Stokes V circular polarization profile. The total magnetic flux has long been used as a more physical measure of the size of an active region, as opposed to total pixel size in

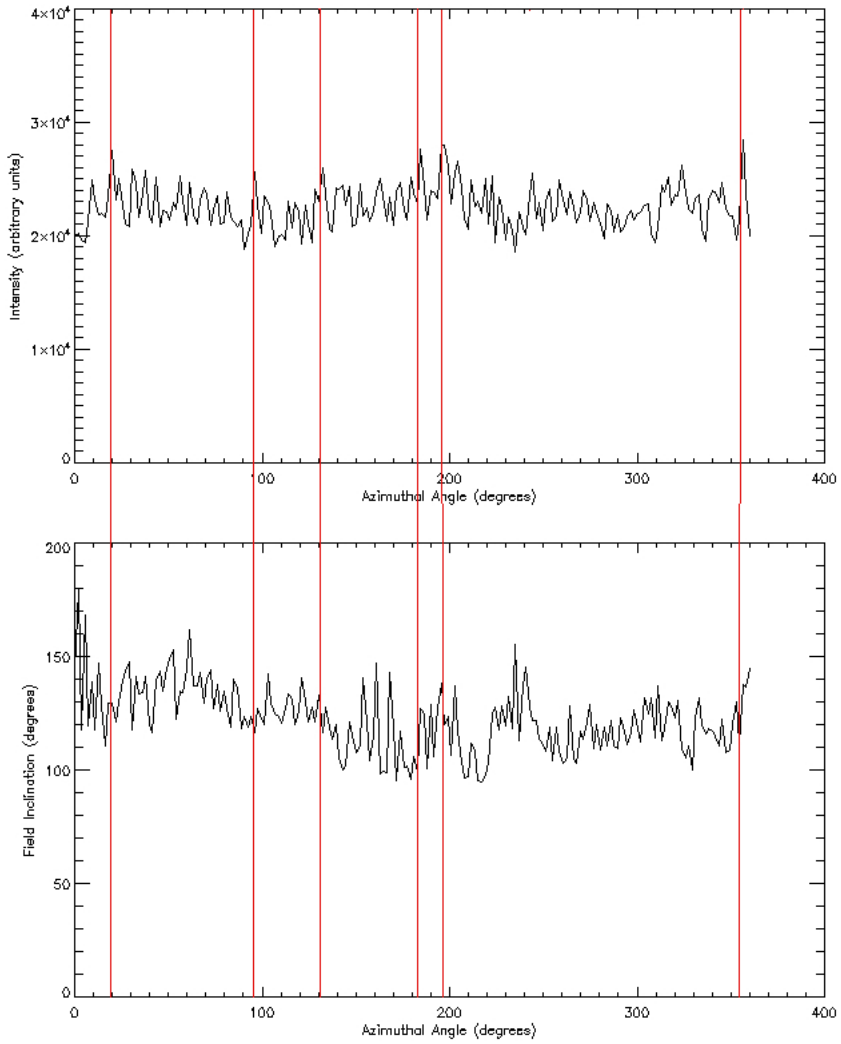

FIG. 5.15 Correlation between penumbral brightness and the horizontal fields which house the Evershed effect. The figure shows a plot of the continuum intensity (top) and field inclination (bottom) as a function of position around the path shown in Figure 5.14. The azimuthal angle now refers to this position, and not the horizontal field orientation. The hot outflow channels of the Evershed effect have a larger brightness which corresponds to the more horizontally-inclined fluxtubes which constrain this outflow. A few of these correlations are marked by the vertical lines through both plots.

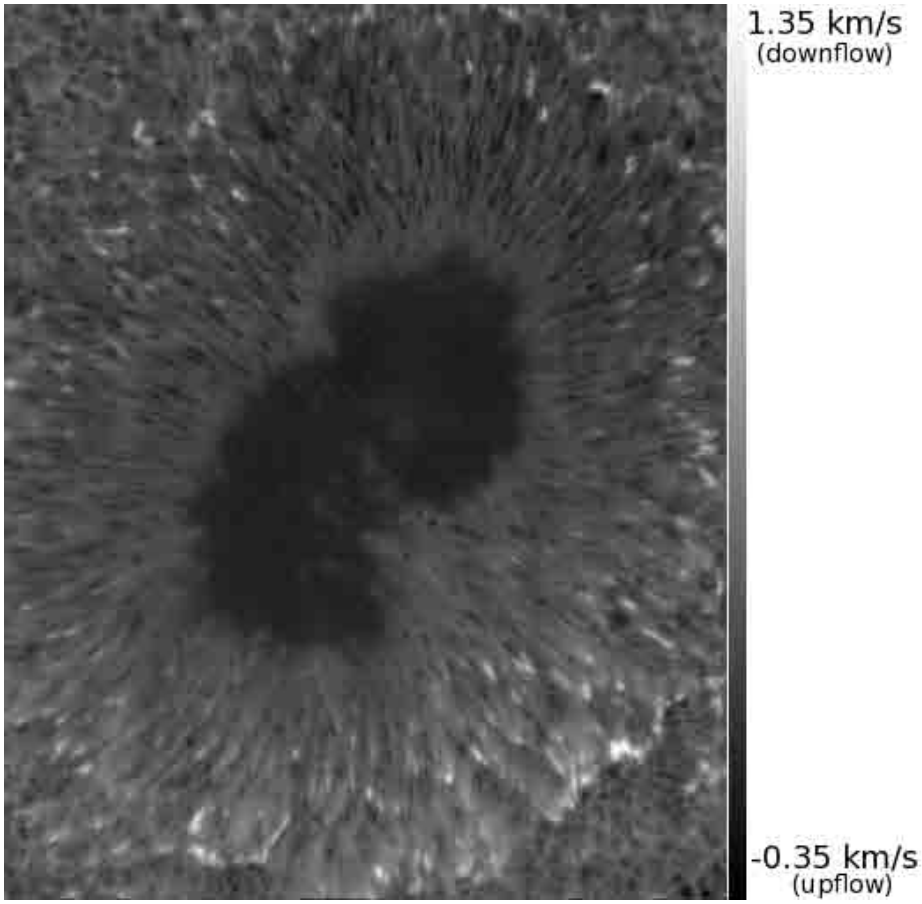

FIG. 5.16 Line-of-sight plasma velocity in AR10923. The figure shows the Doppler velocity derived from the line-center shift of Stokes I profiles. The Evershed effect is clearly identifiable as a difference in the direction of the LOS plasma velocity between the center-side and limb-side penumbra. The signature in this particular observation is relatively weak, since the active region was very close to disc-center. Because of this, there is only a slight difference in the LOS projection of the plasma flow in the horizontal fluxtubes of the penumbra. The effect is more noticeable when the active region is further from disc-center (see section 5.4 below).

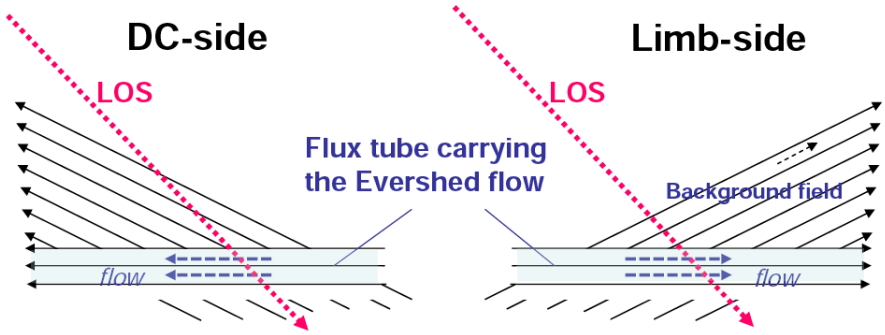

FIG. 5.17 Observation geometry and the Evershed effect. Shown is the effect of observation angle on the resolution and appearance of the plasma flows in the penumbra. On the disc-center side penumbra, the flows have a component moving toward the observer along the line-of-sight, and therefore, the line-center will be blue-shifted to shorter wavelengths. On the limb-side, the reverse is true; the flow appears to be moving away from the observer, and is therefore red-shifted to longer wavelengths. From Ichimoto et al. (2008).

white-light, and larger magnetic flux has historically been correlated with the proclivity for producing flaring events (Canfield et al. [1999], Tian et al. [2002]).

The total flux is $1.76 \times 10^{22} Mx$ which is consistent with Leka and Barnes (2003a), although slightly smaller. This is most likely due to the larger field-of-view used in that work, as well as differences between the active regions they studied; their datasets were large bipolar regions, while AR10923 used in this work is unipolar. The net flux imbalance, defined as

$$|\Phi_{net}| \sim |\sum B_z dA|, \tag{5.7}$$

was calculated to be $1.70 \times 10^{22} Mx$, which is intuitive for this negative-polarity, unipolar sunspot. A larger field-of-view would likely reduce the flux imbalance, since the net flux is highly-dependent on the observed region (Zhang et al. (1994)), and this net flux imbalance, along with its time evolution, is typically associated with flaring activity due to newly-emerging or decaying/dissipating flux.

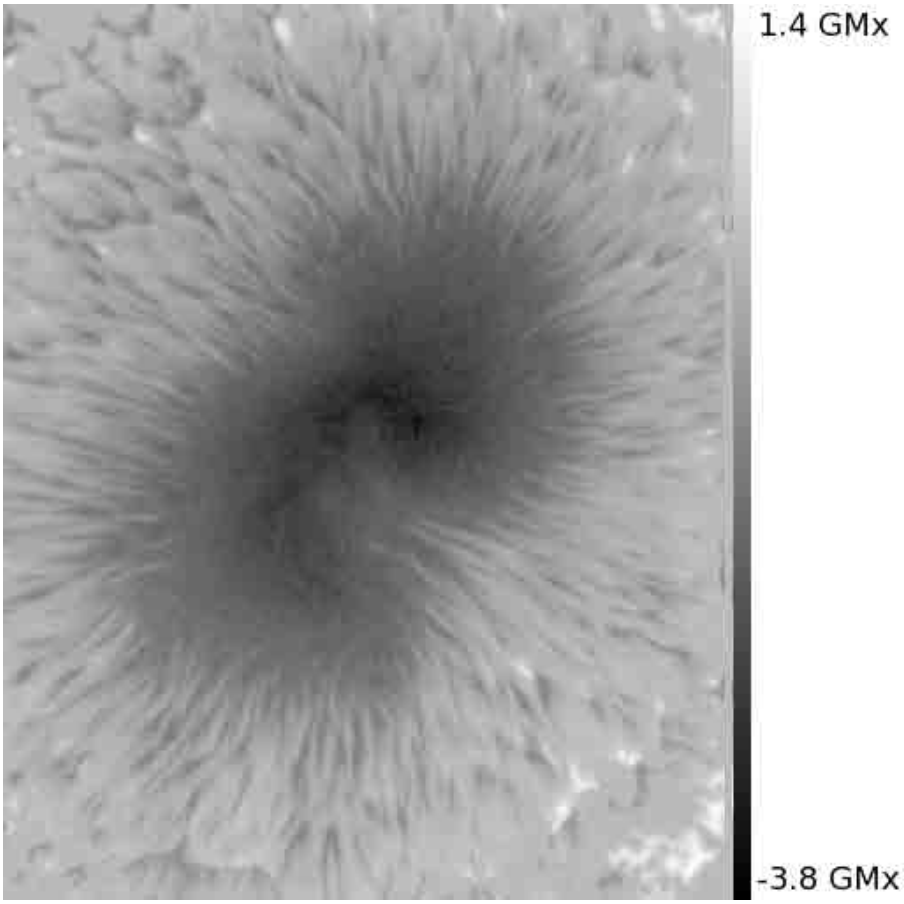

FIG. 5.18 Flux density in AR10956. The flux density, in units of GMx (10^9 Mx), is shown in AR10923. This unipolar sunspot is not flux-balanced within the field-of-view, as is evident in the total negative polarity, with very little surrounding plage of opposite polarity.

5.3.2 Current Density

The presence of electric currents within an active region is a strong indicator of free magnetic energy that is thought to be responsible for allowing magnetic reconnection events, subsequently leading to rapid reconfiguration of the magnetic field topology in a flaring and/or CME event. The electric current density is given by:

$$\vec{J} = \frac{1}{\mu_0}\vec{\nabla} \times \vec{B}, \tag{5.8}$$

where $\mu_0 = 4\pi \times 10^{-3}$ Gm/A is the permeability of free space. Usually, because photospheric vector magnetic fields are typically only available at single heights, one is limited to the

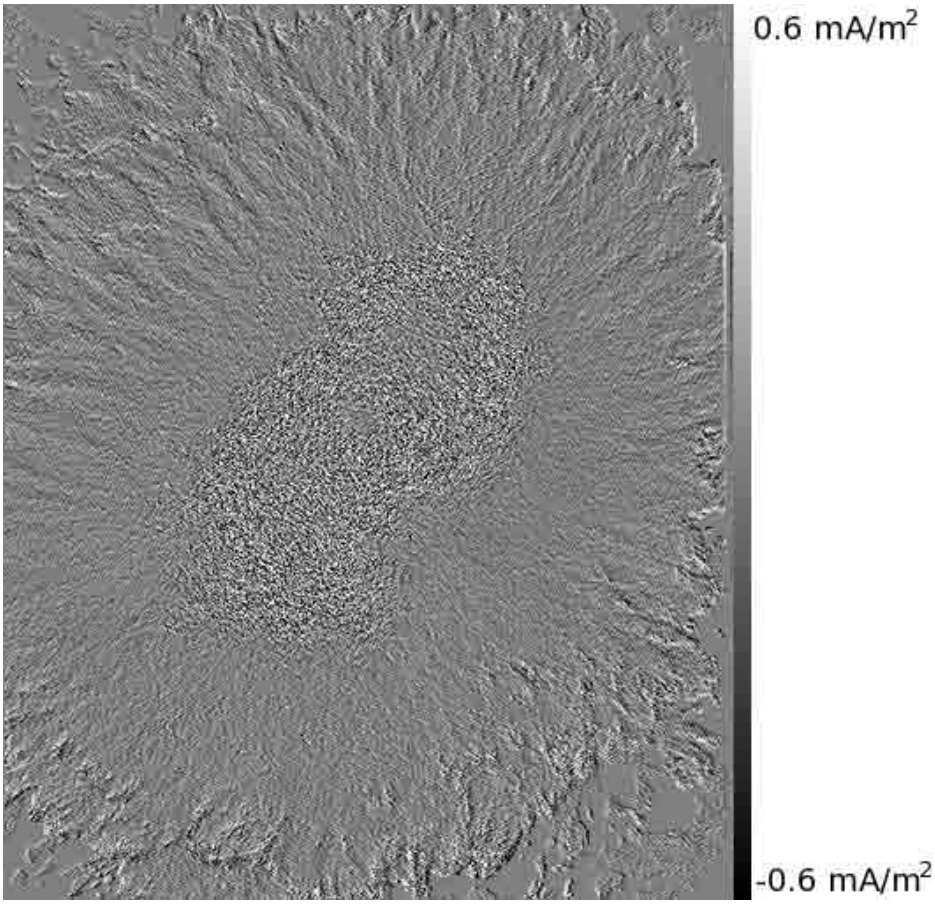

FIG. 5.19 Current density in AR10956. Shown is the vertical component of the current density, in units of $\mathrm{mA/m^2}$, in AR10923.

vertical component (z-component) of the current density

$$J_z = \frac{1}{\mu_0}(\vec{\nabla} \times \vec{B})_z = \frac{1}{\mu_0}\left(\frac{\partial B_x}{\partial y} - \frac{\partial B_y}{\partial B_x}\right). \tag{5.9}$$

This is generally only useful as a lower bound to the magnitude of the full 3D current density, but is nonetheless helpful in establishing (at least qualitatively) where the most "interesting" or "active" parts of the magnetic field lie. This quantity is displayed in Figure 5.19.

The total unsigned current in AR10923 was calculated as 1.26×10^{14} A, which is somewhat larger than the typical values for the three active regions studied in Leka and Barnes (2003a). Wang et al. (1994) hypothesized that strong vertical currents should stimulate flaring activity, although Leka and Barnes (2003a) showed that large currents are

not a sufficient condition for flaring, since their quiet active region showed larger currents than one that was flare-active. The net current, or current imbalance is consistent with Leka and Barnes (2003a), taking a value of $-4.11 \times 10^{11} A$. The average vertical current density is essentially zero for this active region ($\approx -10^{-7}$ mA/m^2), leading to a net current of ≈ -2 A. This is consistent with the fact that the net current imbalance previously measured is a factor of 1000 smaller than the total current. For completely unbalanced systems, the net imbalance should be approximately equal to the total current. This indicates that vertical currents originating within AR10923 ultimately return to a sink somewhere in the field-of-view, most likely near the outer penumbral boundary, much like a fountain.

5.3.3 Magnetic Helicity

The degree of "twist" in a magnetic field configuration can *usually* be associated with the flaring potential of an active region. A well-known example is the *kink instability*, whereby a flux tube can only support so much twist of the field around its axis before the axis itself begins to deform. Because of an inherent limitation in the information gleaned from the active region magnetic field at a single height, typically only the z-component of the total helicity density can be calculated, as the other terms involve vertical gradients in the field, to which we do not have access, i.e.,

$$h^{(c)} = \vec{B} \cdot \vec{\nabla} \times \vec{B} \tag{5.10}$$

$$h_z^{(c)} = B_z \left(\vec{\nabla} \times \vec{B} \right)_z \tag{5.11}$$

Since B_z is largest for low-inclination fields, and $\vec{\nabla} \times \vec{B}$ essentially estimates the direction of rotation of the magnetic field direction along a field line, the quantity h_z^c acts as a proxy for the amount of "twist" in nearly vertical magnetic fields. The vertical component of the helicity is shown in Figure 5.20.

The total current helicity density was calculated as 3029.6 G^2/m, roughly 2–3 times higher than that of Leka and Barnes (2003a), but the fractional helicity imbalance of 131 G^2/m is consistent with their measure. The average helicity density over the field-of-

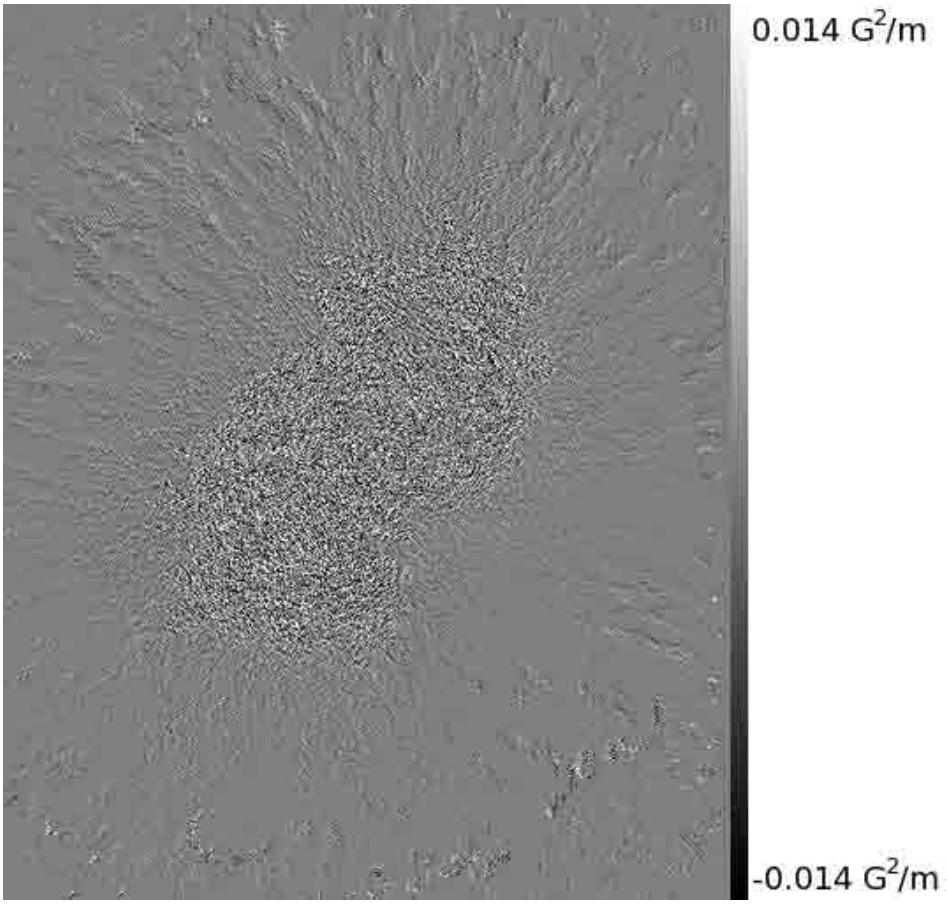

FIG. 5.20 Magnetic helicity density in AR10956. The vertical component of the magnetic helicity, in units of G^2/m, is shown in AR10923.

view is 0.0266 G^2/m, which is roughly one order of magnitude larger than values reported in Bao et al. (1999). The percentage helicity imbalance, defined as

$$100\frac{\sum h_z^{(c)}}{\sum |h_z^{(c)}|}, \tag{5.12}$$

is calculated to be +4.33%, indicating that positive helicity fields occupy slightly more of the field-of-view than negative helicity fields. Of interest is the change of the helicity over time. Several authors (Bao et al. [1999], Liu and Zhang [2002]) have shown a decrease in the magnitude of the average vertical current helicity density prior to both M- and X-class flares, and Leka and Barnes (2003a,b) observed changes in the statistical variance and the kurtosis of the helicity density distributions over their flare-productive active regions. An examination of the temporal behavior of these quantities is presented in section 5.4.

5.4 Time Series Inversions

NOAA AR10923 was observed by the *Hinode* satellite between 10 November and 19 November 2006. I have obtained a full set of the four Stokes parameter maps, one for each day in this range, and have performed the genetic inversion on this time-series. The results previously presented were from the observations on 14 November 2006, when the sunspot was near disc-center. For the observation periods on either side of this window, the angle between the solar normal and the line-of-sight is non-negligible, so there will be projection effects in the inversions of these regions when they are significantly far away from disc-center. Figure 5.21 shows a mosaic of continuum intensity images of AR10923 as it traverses the solar disc, and Figure 5.22 shows the corresponding GOES X-ray flux in two bandpasses for the same period. Nevertheless, the ability of the genetic inversion to track changes in an evolving active region will be very important for future high-speed applications to large datasets. This section presents the preliminary results into this aspect of the investigation.

5.4.1 Field Strength

Figure 5.23 shows the evolution of the magnetic field strength over the time-series. The genetic inversion required approximately 70 minutes per inversion, and somewhat less for the very last dataset, where the foreshortening of the sunspot has significantly decreased the number of pixels in the field-of-view with non-trivial polarization signal. The author cannot make any inferences about the existence of AR10923 prior to 10 November 2006, when the sunspot was not on the visible portion of the solar disc. However, the growth of the sunspot (measured by increasing field strengths in the umbral center) seems to indicate that the sunspot was still relatively young at the time of the observation on 10 November 2006. The umbral field strengths increase during the first few days of the observation, after which the growth of the light-bridge becomes obvious in the reduced field strength that subsequently splits the high umbral field strengths into two distinct regions.

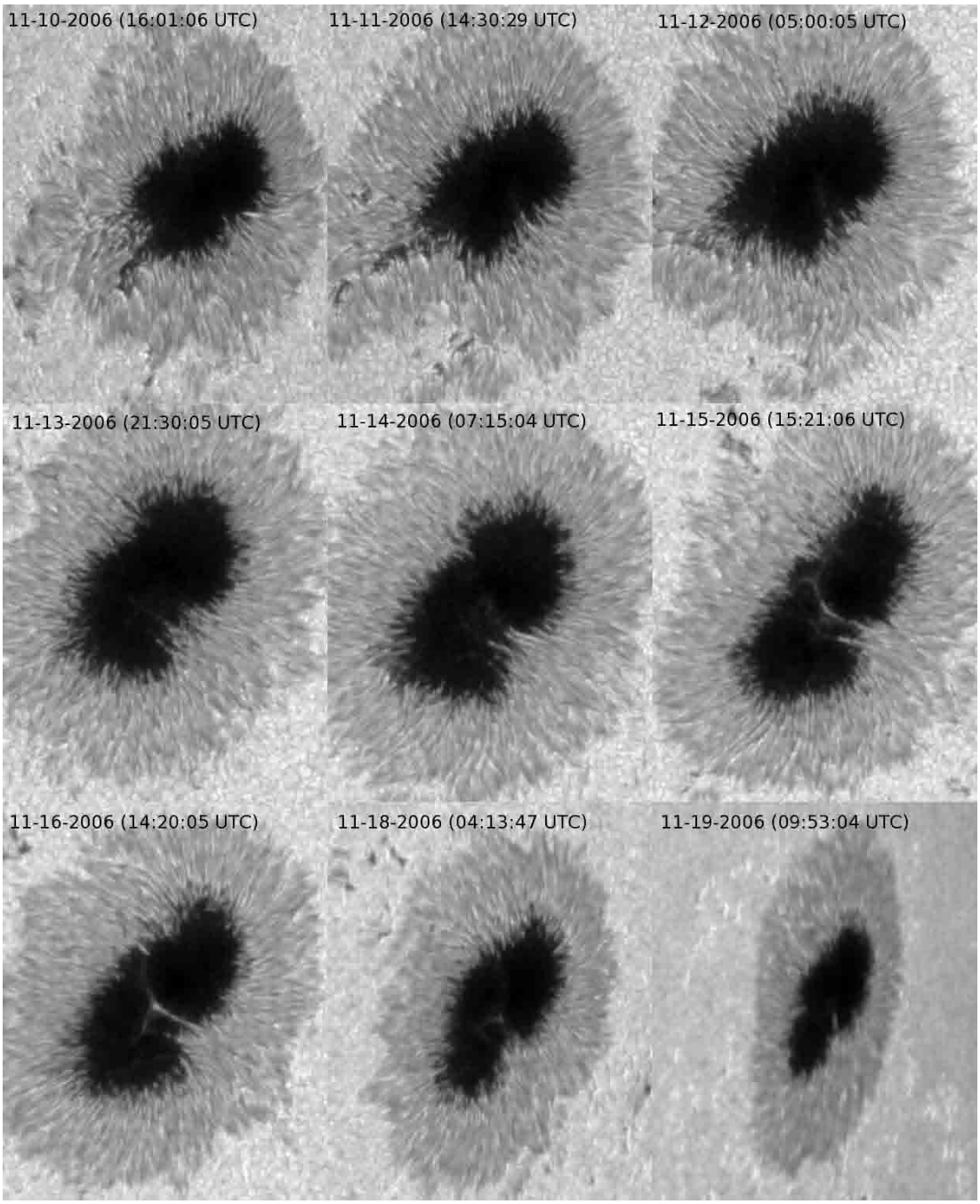

FIG. 5.21 A series of observations of AR10923, spanning nine days. AR10923 was observed between 10 November 2006 (top left) and 19 November 2006 (bottom right) by the *Hinode* satellite. The timestamps of each observation are located in the upper region of each frame. The foreshortening caused by observation near the east and west solar limbs are evident as the sequence is examined from top to bottom, and left to right.

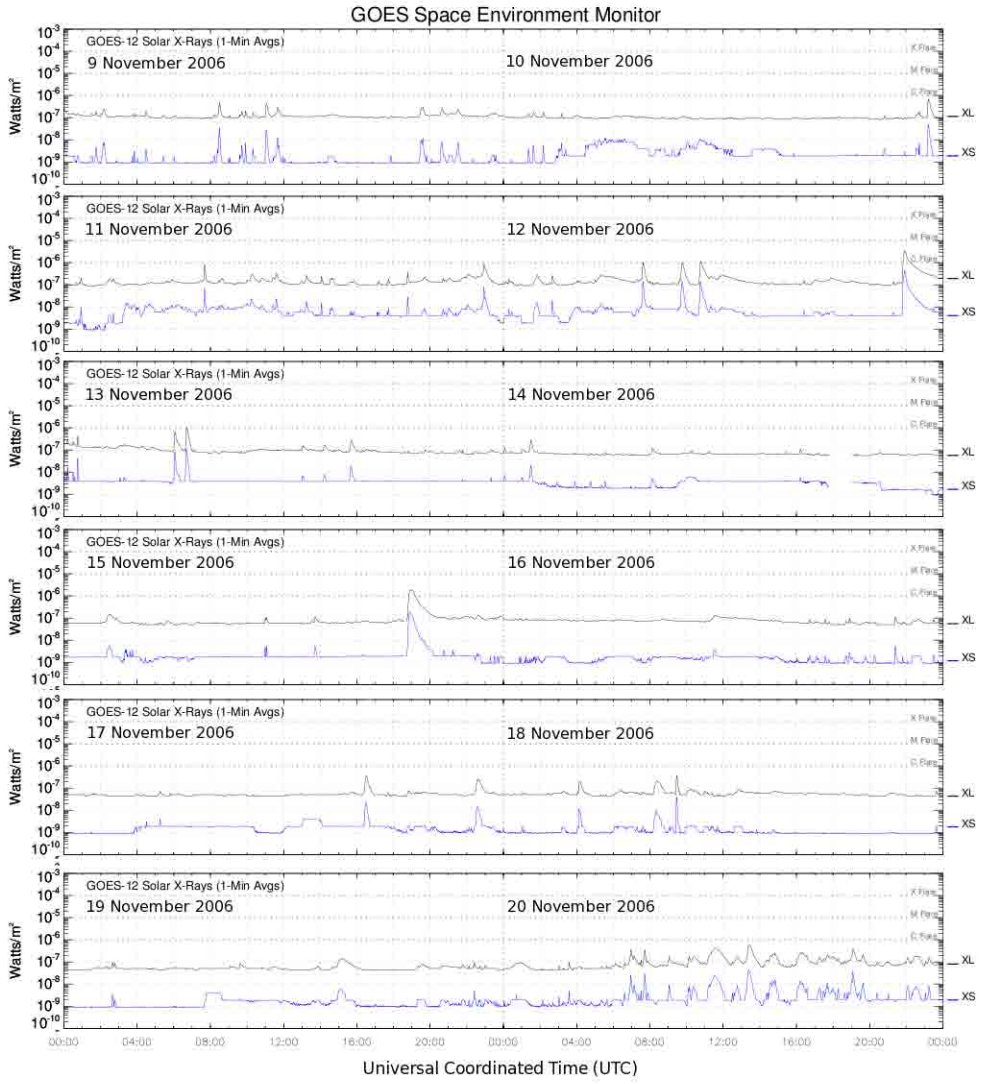

FIG. 5.22 GOES X-ray flux for November 9–20, 2006. The figure shows the X-ray flux (in Watts/m^2) in two wavelength regimes for the period between 9 November 2006 and 20 November 2006, during which time AR10923 was observed by the *Hinode* satellite. The black curve depicts the flux of X-rays with wavelengths between 1 and 8 Å, while the blue curve depicts the same quantity for X-rays with wavelengths between 0.5 and 3 Å. There were two insignificant C-class flares on 12 November 2006 (22:00 UTC) and 15 November 2006 (19:00 UTC), but the vast majority of the observation time was particularly quiet, with little or no flaring activity as AR10923 traversed the solar disc.

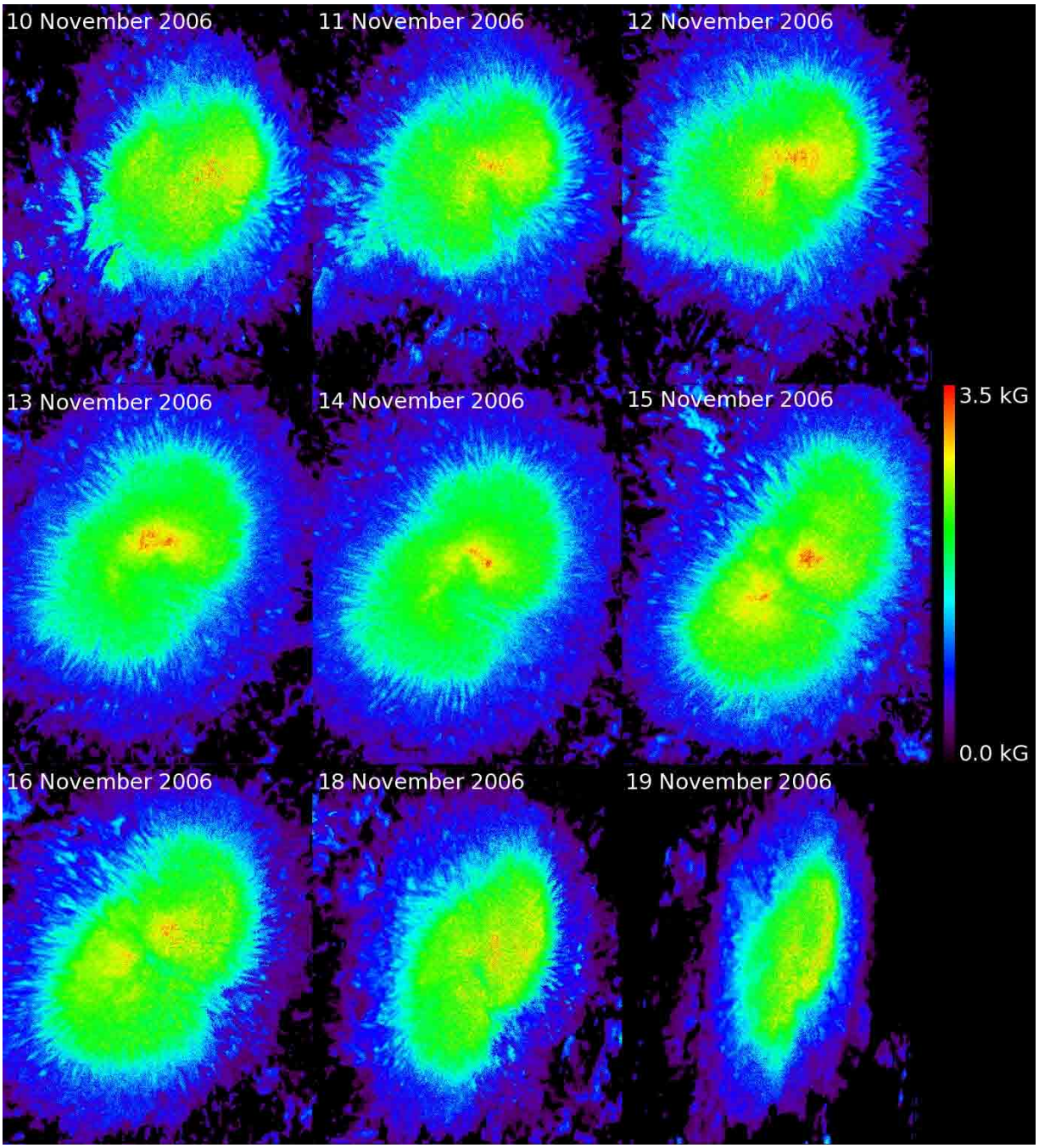

FIG. 5.23 Evolution of magnetic field strength in AR10923. The figure shows the magnetic field strength inferred by the genetic inversion for each dataset in the time-series observation of AR10923. The growth of the sunspot between 10 November 2006 and 14 November 2006 is apparent in this image, and the formation of the light-bridge (as shown in Figure 5.21) is also apparent in the reduction of the field strength in the protrusion into the umbra.

5.4.2 Field Inclination

Figure 5.24 shows the evolution of the magnetic field inclination over the time-series. This is expected to show the greatest projection effects; umbral fields that should be vertical will show a great deal of horizontal orientation, since we are observing them "from the side." From Figure 5.17 on the geometry of the Evershed effect, one can easily see why the penumbral fields during the early and late observations appear to be vertically oriented. These projection effects arise from the observer's line-of-sight falling nearly along the center-side penumbral magnetic field directions, because of the position of the sunspot on the solar disc. For this negative polarity sunspot, the center-side penumbra will always show this type of behavior, hence the change from the high inclinations starting on the right-side of the images and progressing toward the left side as the sunspot moves across to the west solar limb.

5.4.3 Field Azimuth

Figure 5.25 shows the evolution of the magnetic field azimuth over the time series. In the early and late phases of the time-series, when the sunspot is close to either solar limb, we expect to observe that the center-side penumbra will exhibit the same kind of indeterminate behavior in the azimuthal angle as the umbral regions presented earlier in this chapter. Again, for the same reasons we see projection effects in the inclination of the field, the penumbral fields which should be relatively horizontal are now lying mostly along the line-of-sight, where the azimuth is then not well-defined. However, the formation of the light-bridge is also apparent in this image, between 13 November 2006 and 16 November 2006, as the yellow protrusion just below the center of each image moves further into the umbra.

5.4.4 Line-of-Sight Velocity

Figure 5.26 shows the line-of-sight plasma velocity over the time-series. As in a previous section on the penumbral fine structure, the Evershed effect is registered as a difference

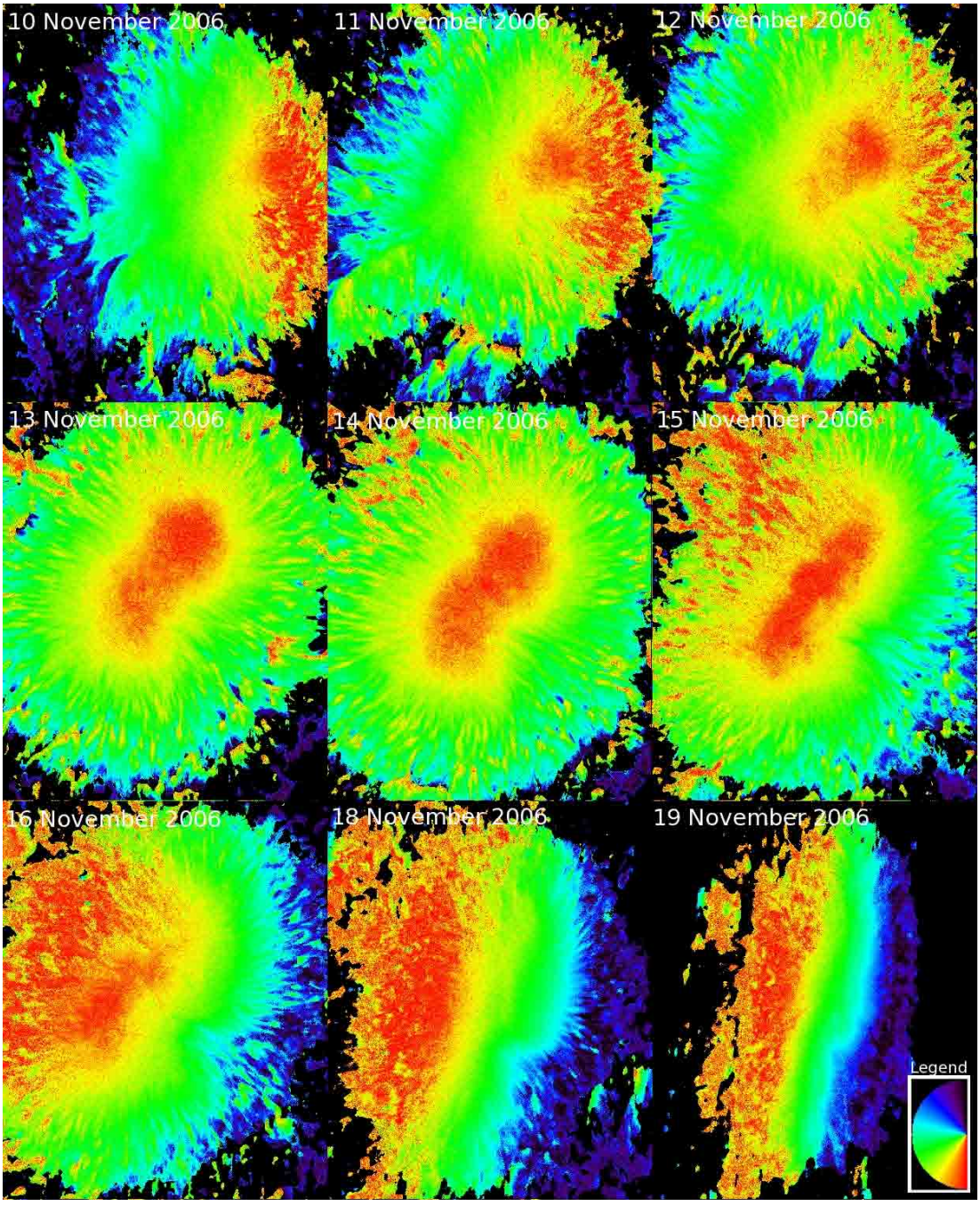

FIG. 5.24 Evolution of magnetic field inclination in AR10923. Shown is the magnetic field inclination from the line-of-sight for each dataset in the time-series observation, as inferred by the genetic inversion. As before, the color wheel legend at the lower right indicates that red points "down" into the plane of the page, while black points "up" out of the page.

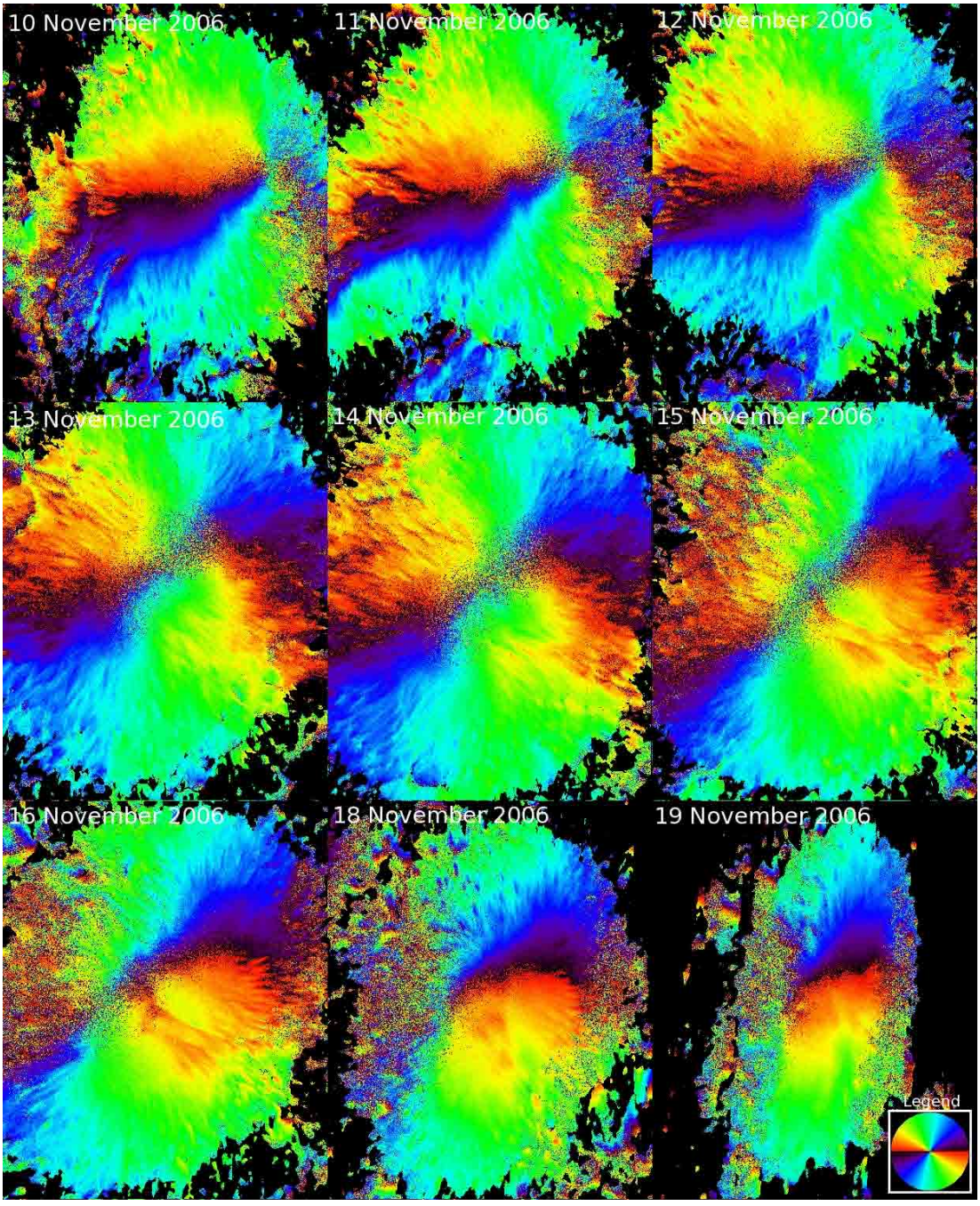

FIG. 5.25 Evolution of magnetic field azimuth in AR10923. The figure shows the magnetic field azimuth (transverse direction) inferred by the genetic inversion for each dataset in the timeseries observations. Again, the ambiguous color wheel at the lower right indicates the direction of the field by color, and the 180° ambiguity is built-in to the legend.

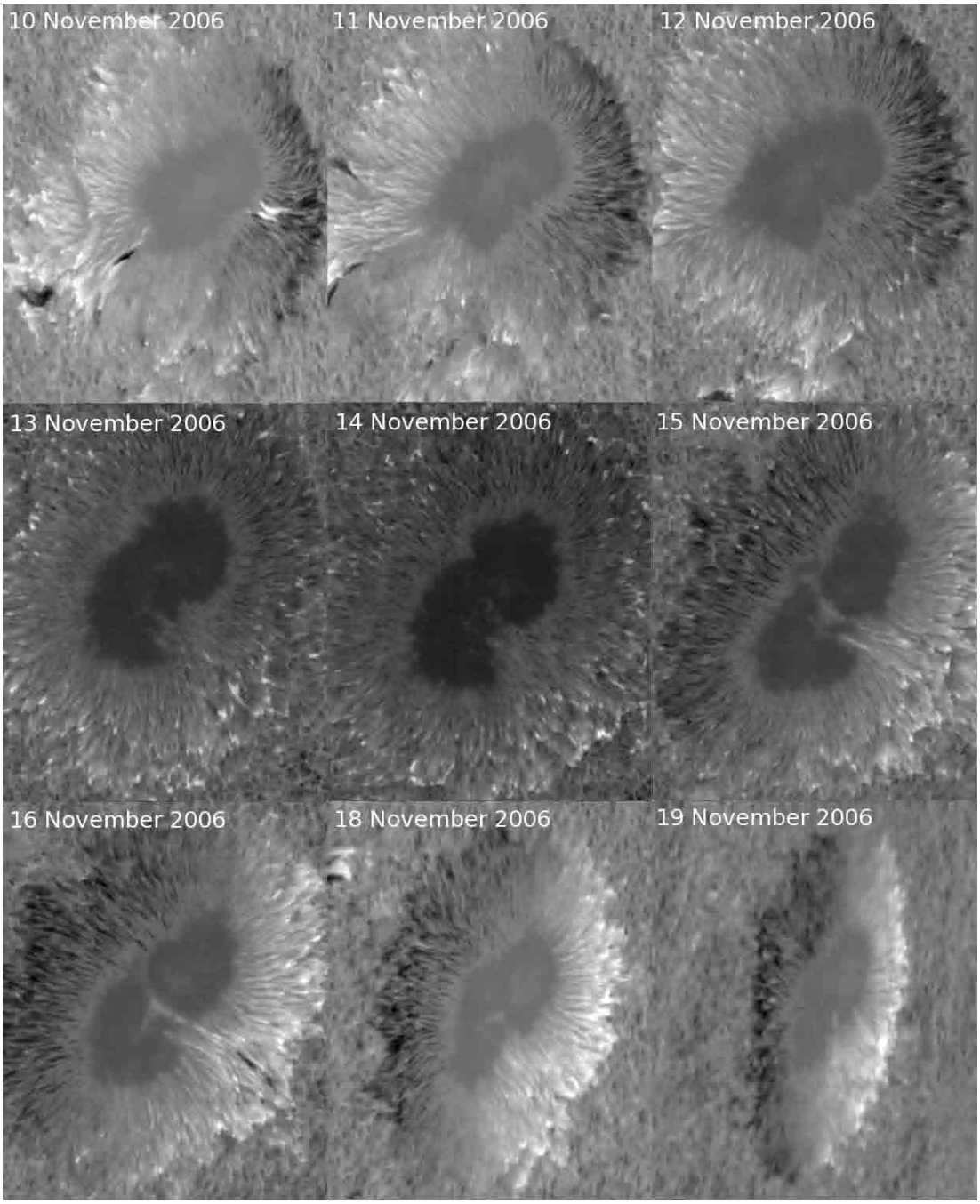

FIG. 5.26 Evolution of the line-of-sight plasma velocity in AR10923. Evidence of the Evershed effect is apparent in the timeseries observations, as differences between the flow velocities on the limb- and center-side penumbra. Black corresponds to negative velocities (toward the observer) while white registers positive velocities (away from the observer. Again, we see the interchange between the limb- and center-side penumbral properties as the sunspot traverses the solar disc.

in flow velocities between the center-side and the limb-side penumbra. With the rather large observation angles, these differences are now very much more pronounced than in the disc-center case previously presented. Line-of-sight velocities are of the order of a few to several km/s, and the convective Doppler shifts are apparent in the quiet regions outside the sunspot. That is, the cores of the convective cells appear black, since they represent the upwelling plasma motion along the line-of-sight, while the intergranular network appears bright due to the downflowing plasma there redshifting the wavelengths.

Ideally, with the inversion results at each step of the time-series in hand, one could apply a rotational transformation to project the limb observations to disc-center, then quantitatively analyze the evolution of the magnetic flux, for example, in the active region. This rotation seems trivial to apply, however to date I have been unable to successfully perform the transformation. The steps of the transformation are as follows:

(1) Obtain the heliocentric coordinates (in arcseconds) of the center pixel in the field-of-view from the *Hinode* SP headers.

(2) Geometrically extrapolate the heliocentric coordinates to every other pixel in the field-of-view.

(3) Find or calculate the solar B_0, L_0, and P angles for the date and time of the observation.

(4) Construct the rotation matrix of Gary and Hagyard (1990).

(5) Convert the magnetic field (in the local reference frame) from spherical coordinates to local Cartesian coordinates.

(6) Apply the rotation matrix to the Cartesian components of the local magnetic field to obtain the Cartesian components of the heliocentric magnetic field.

(7) Convert the heliocentric magnetic field from Cartesian coordinates back to spherical coordinates.

(8) Enjoy success.

To the best of my knowledge, my attempts to perform this transformation falter somewhere between steps (7) and (8). Figure 5.27 shows a sample attempt at transforming the genetic inversion results from 10 November 2006 to disc-center. It is tantalizingly suggestive of a very minor misstep in the algorithm, as the lower half of the image appears to be almost correctly rotated. That is, from the original image, the highly-inclined fields at the extreme right seem to be almost transformed to the proper vertical fields in the umbral regions, as they should be. However, the upper half of the image is a complete mess. All attempts to track down the misstep in the rotation algorithm have resulted in severe frustration and a near-miss with the destruction of certain computing equipment. Nevertheless, this area of research will be explored (and hopefully corrected) in the near future.

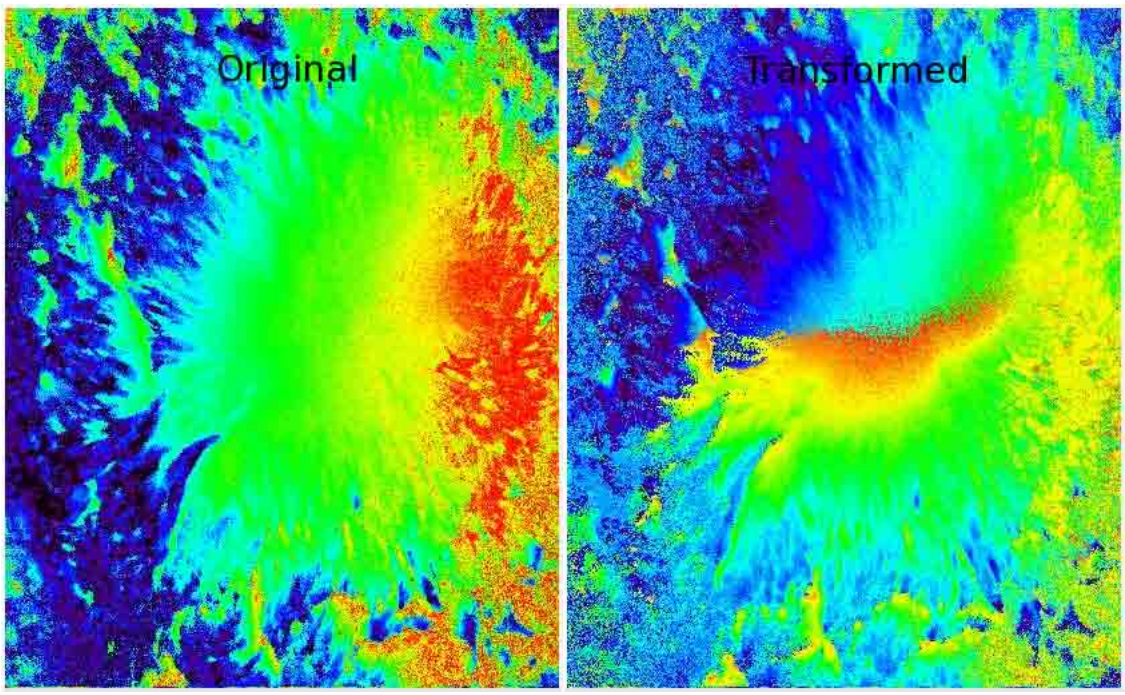

FIG. 5.27 An attempt at magnetic field deprojection. Shown is the magnetic field inclination from the original inversion results (left) as well as the results of applying the rotation transformation of Gary and Hagyard (1990) (right) to extrapolate what the vector magnetic field would look like if the same observation was made at disc-center. It is obvious that there is still much work to be done in this area.

Although we currently lack a properly-functioning transformation routine and therefore, cannot quantitatively analyze the longitudinal flux, the total magnetic field strength is invariant through such a rotation. Therefore, we can still perform at least a degree of quantitative analysis of the field strength in AR10923. Following Leka and Barnes (2003a,b) we calculate the moments of the distribution of field strength within AR10923. We calculate the average field strength, standard deviation, skew, and kurtosis of this distribution and examine its behavior over the observation window. Leka and Barnes (2003a,b) have analyzed several flare-active and flare-quiet regions with this method, and have had limited success in identifying unique pre-flare signatures in the distribution of magnetic field strength. Nevertheless, this method can identify signatures of flux emergence and disappearance, and flux imbalance, so it may be useful for characterizing newly-emerging or decaying active regions. The parametrization of the spatial magnetic field distribution is accomplished by examining the following four equations as a function of time:

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i \tag{5.13}$$

$$\sigma_X = \left[\frac{1}{N}\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2\right]^{1/2} \tag{5.14}$$

$$\xi_X = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{X_i - \overline{X}}{\sigma_X}\right]^3 \tag{5.15}$$

$$\kappa_X = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{X_i - \overline{X}}{\sigma_X}\right]^4 - 3.0. \tag{5.16}$$

These parameters represent the average, standard deviation, skew, and kurtosis of the distribution X, and have the following interpretations:

- The average, $\overline{X}$, is the average value of the distribution.
- The standard deviation, σ_X, is a measure of the spread of the distribution around the average.
- The skew, ξ_X, is a measure of the asymmetry of the tails of the distribution.

- The kurtosis, κ_X, is a measure of the "Gaussianity" of the distribution.

Figure 5.28 shows a histogram of the distribution of magnetic field strength in AR10923 between 10 November and 19 November 2006. Immediately noticeable is the double-lobed shape of these distributions. In the early and late observations, when AR10923 is near the east and west solar limbs, respectively, the strong-field tail is relatively flat, and these tails approach zero at lower field strengths than when the sunspot is near disc-center. This indicates the growth and decay of the sunspot, respectively, and is also clearly visible in the middle row. The middle row shows the emergence of strong field in the umbra as the growth of the strong field lobe near 2000 G. Figure 5.29 shows the four moments (Eqns. 5.13–5.16) of these distributions as a function of observation date. This shows even more clearly the growth and decay of the sunspot in the average field strength over the field of view. The standard deviation of the field strengths remains relatively constant, until the sunspot has almost disappeared from the solar disc on 19 November 2006. The skew of the distribution is well-correlated with the increase in average field strength, primarily to the growth of the strong-field lobe, which causes the distribution to become increasingly asymmetric about its mean. Finally, the kurtosis of the distribution shows strongly non-Gaussian signatures, as is evident from the distributions themselves; they are composed of separate populations, namely the penumbral and umbral fields.

Although initially simplistic, this type of analysis is capable of inferring the growth and decay of the sunspot magnetic fields, and therefore can be helpful in assessing their flare-productivity. However, this moment analysis should be repeated on the longitudinal flux over the active region, calculated with the deprojected fields, and compared to these results. Once I have corrected the flaw(s) in the disc-center transformation procedure, this will be done. Obviously, a longer dataset at higher temporal cadence (particularly at disc-center) would be much more helpful in this type of analysis, and I plan to continue this analysis in the future, working with high-cadence, full-disc SOLIS spectropolarimetry data.

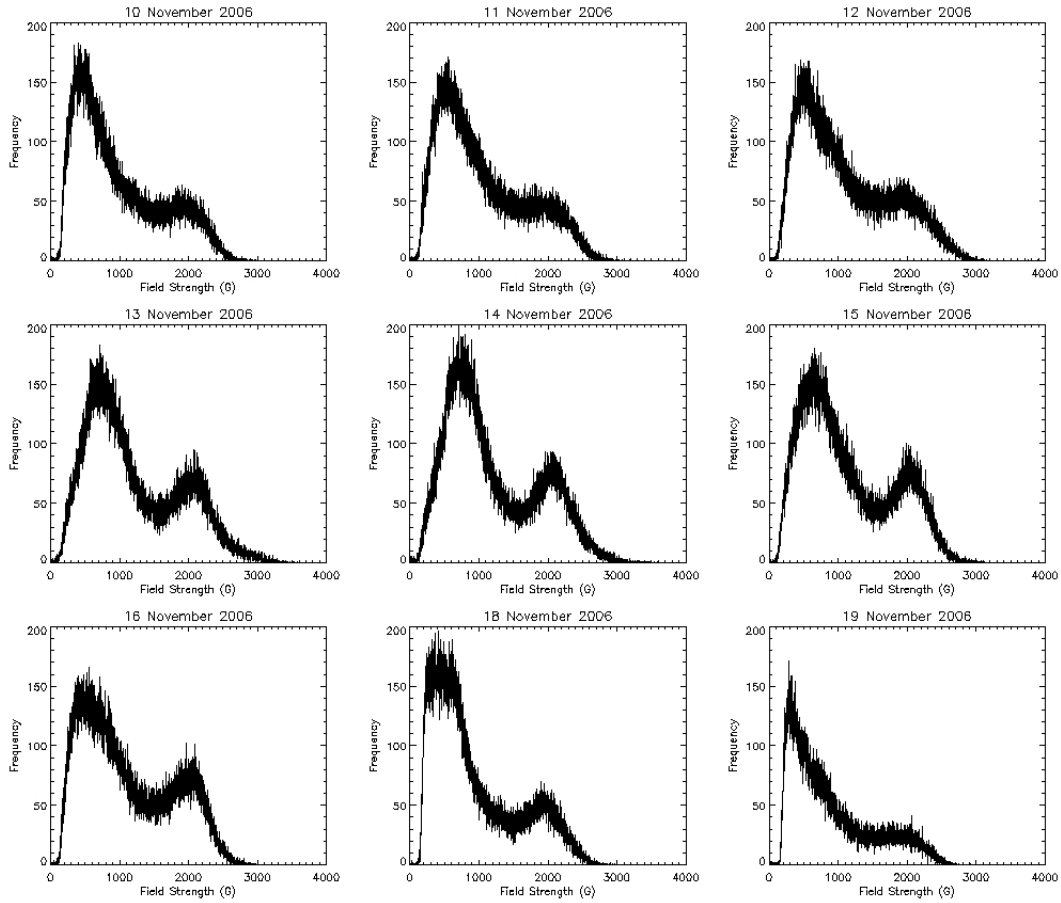

FIG. 5.28 Evolution of the magnetic field strength distribution in AR10923. The histograms show the total field strength distribution in AR10923 between 10 November and 19 November 2006.

5.5 A More Complicated Active Region: NOAA AR10956

As detailed in Chapter 3, NOAA AR10956 was observed by the Diffraction-Limited Spectropolarimeter at the National Solar Observatory, Sunspot, NM, on 20 May 2007. At the time of the observations, the active region was located at 1.6 N latitude and 14.7 W longitude on the solar disc. It is a much more complicated active region than either AR9240 or AR10923, and as such it belongs the "delta" classification, typically meaning a bipolar active region with several umbral regions, many of which share the same penumbral envelope. These "delta" active regions tend to be more flare-productive, owing to their complex plasma motions and field topologies. In fact, AR10956 produced many CMEs

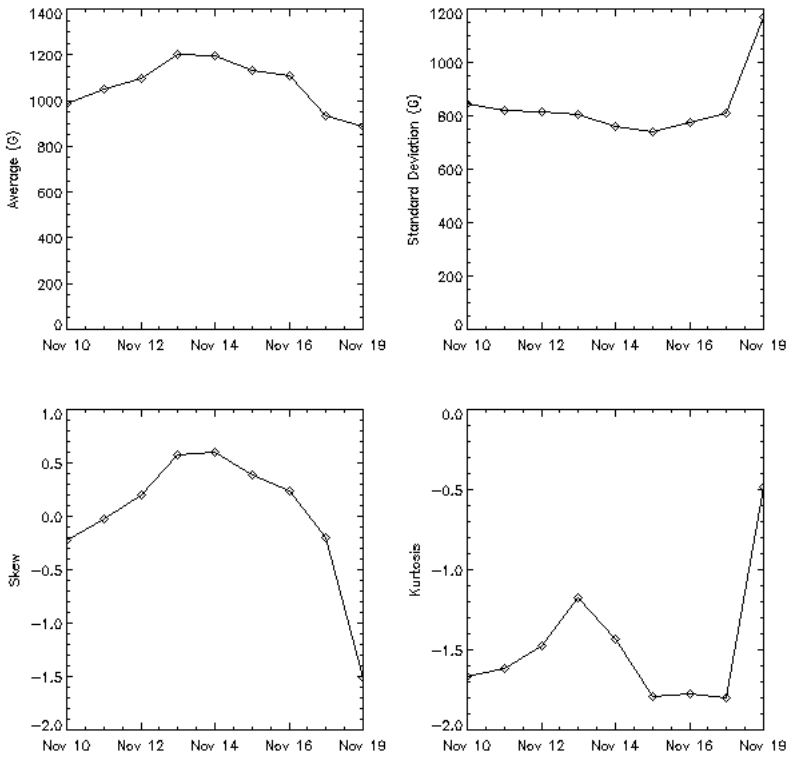

FIG. 5.29 Four moments of the magnetic field strength distribution in AR10923. The average (top left), standard deviation (top right), skew (bottom left) and kurtosis (bottom right) statistics of the distributions in Figure 5.28, are plotted as a function of time.

during its traversal of the solar disc in May 2007. It is unfortunate that I do not have a series of observations of this active region, with which to explore the changes in the vector magnetic field in the pre- and post-flare configurations. However, the active regions is sufficiently more complex than the previous cases, such that we may test the robustness of the genetic inversion in active regions which show large departures from circular, radial symmetry.

Figure 5.30 shows the magnetic field configuration inferred by the genetic inversion of this active region. The diffraction-limited angular resolution of the DLSP is approximately 0.2", giving this particular field-of-view linear dimensions of 69.0 Mm in the horizontal direction and 55.2 Mm in the vertical. The complex, bipolar nature of the active region is

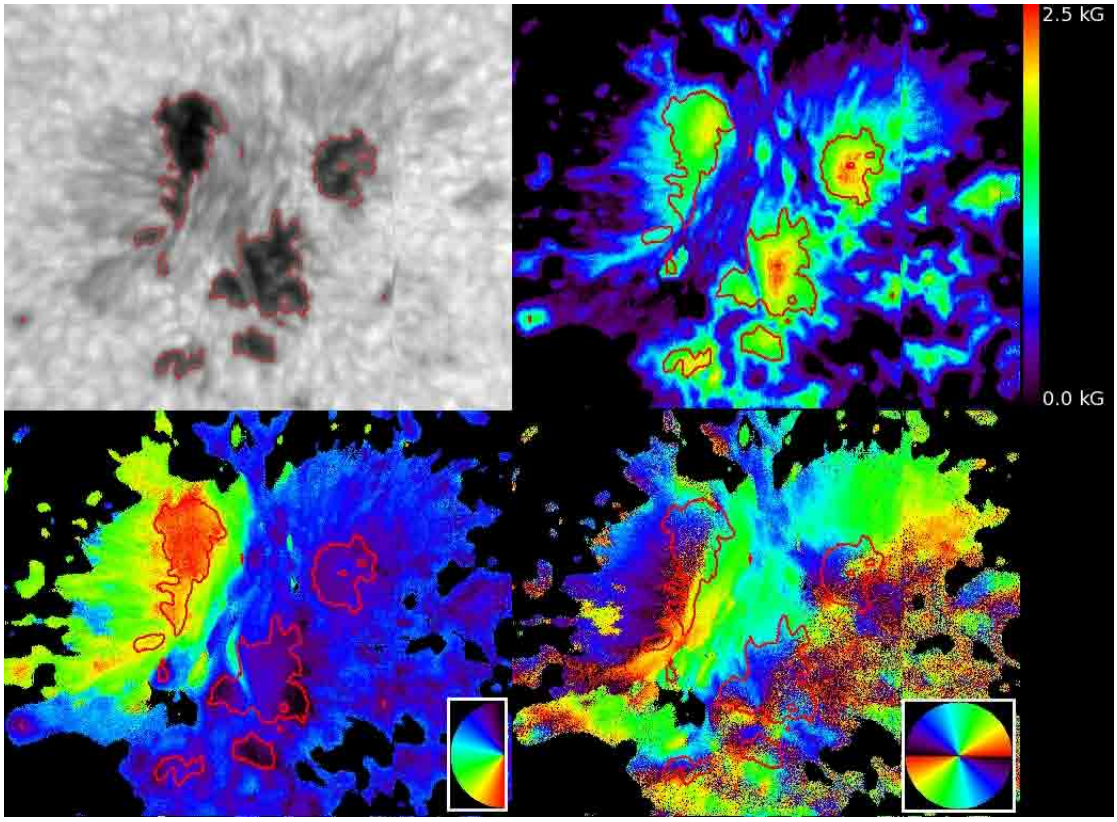

FIG. 5.30 Continuum image and magnetic field geometry of AR10956. The figure shows a continuum (white-light) image of AR10956 as well as the strength, inclination, and azimuthal angle of the magnetic field, as inferred by the genetic algorithm. The red countours outline the umbral regions of AR10956.

well-recovered by the genetic inversion, and the light-bridge cutting into the lower umbral regions is easily identifiable in the lower field strength and higher inclination. Because of this complex nature, there is no well-defined radial structure, and so the 180° ambiguity cannot be resolved by assuming radial fields. Disambiguation must be done by comparison to a potential field calculation, or by a local structure minimization algorithm, as in Georgoulis et al. (2004).

Figure 5.31 shows the longitudinal flux, the quantity measured by traditional line-of-sight magnetograms. Since this active region is bipolar, we may now test the flux balance in the region. One should always be wary when calculating the flux budget of an active region, since there are field-of-view effects that exclude, for example, plage field in the

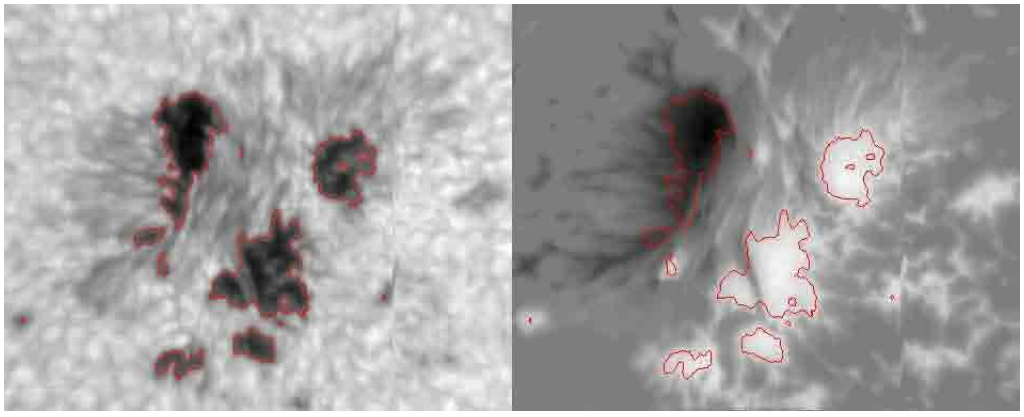

FIG. 5.31 Longitudinal magnetic flux in AR10956. The longitudinal magnetic flux in AR10956, showing both positive and negative polarity magnetic fields, is similar to what a traditional line-of-sight magnetogram would measure.

surrounding areas, which will make the region appear as if the flux is unbalanced. However, from the asymmetry between the physical areas occupied by positive- and negative-polarity field, one might already expect the region to be flux-unbalanced. The total (unsigned) flux in the active region was determined to be 1.49×10^{22} Mx, which is consistent with the magnitudes of typical sunspots, although slightly smaller than the estimates in Leka and Barnes (2003a). This is to be expected, since the higher spatial resolution of the DLSP means a smaller (total) field-of-view. Refer to Figure 3.9 for a comparison of the size of this DLSP field-of-view with that of the ASP. The net fluxes of each polarity were 1.17×10^{22} Mx for the positive areas and -3.26×10^{21} Mx for the negative areas, therefore the net flux imbalance is 8.41×10^{21} Mx. As expected, this region is not flux balanced. Further evidence of this can be seen in the geometry of the neutral line (also known as the polarity inversion line). Figure 5.32 shows a null contour of the line-of-sight flux, which displays the regions where the line-of-sight component of the magnetic field is zero. This linear region is of particular interest, since several authors have found that the flare-productivity of an active region is typically correlated with the complexity of the neutral line, as well as the shear of the magnetic field near the neutral line. Wang (2006) analyzed five flare-producing δ-class sunspots (like AR10956) and found significant increases (and decreases) in the transverse field gradients across the neutral line prior to and following flares. Yang

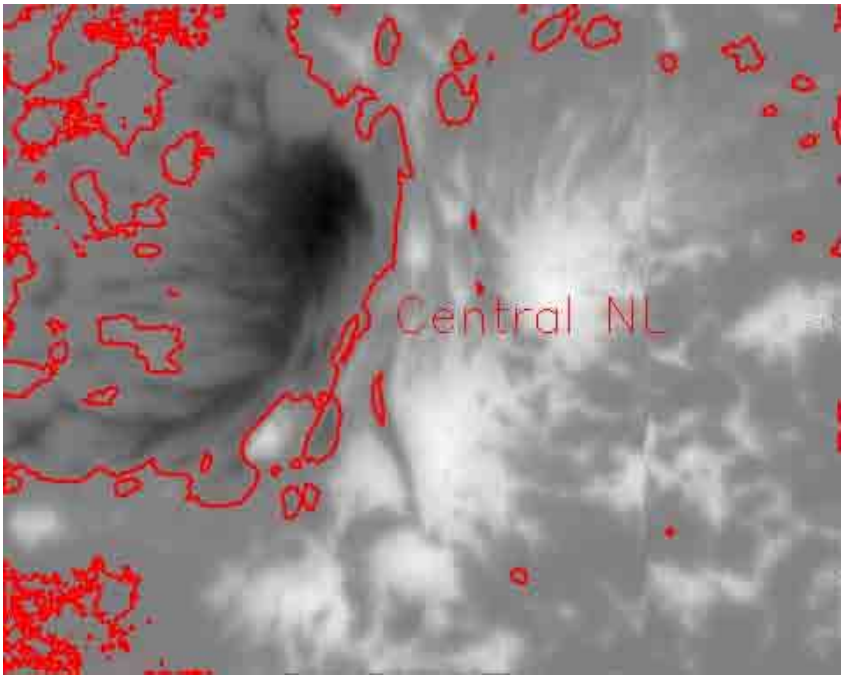

FIG. 5.32 Neutral line geometry in AR10956. The neutral line (polarity inversion line) is inferred from the null points in the line-of-sight magnetic flux in AR10956. The longitudinal magnetic field vanishes at these points, and is interpreted as the top of magnetic arches spanning the region between the two polarities.

et al. (2004) analyzed high-resolution data obtained at the National Solar Observatory and found horizontal flows across the neutral line as well as shear flows traveling in opposite directions on opposite sides of the neutral line, prior to an X10 flare which had a significant white-light component. It is important to note that the *magnetic* shear is defined as the difference between the vector magnetic field inferred from observations and a potential (current-free) field extrapolated from the longitudinal component of the observed vector magnetic field, while the more general term shear simply represents a flow with the same oppositely-pointing characteristics as shearing (deforming) forces applied to a solid body. A simplified schematic of neutral line geometry in a bipolar active region is shown in Figure 5.33.

The neutral line is identified by first smoothing the longitudinal flux over a small window, to eliminate pixel-to-pixel variations that might lead to a fragmented or disjoint neutral line. Pixels belonging to the neutral line are then identified as zero-flux pixels, and continuity is

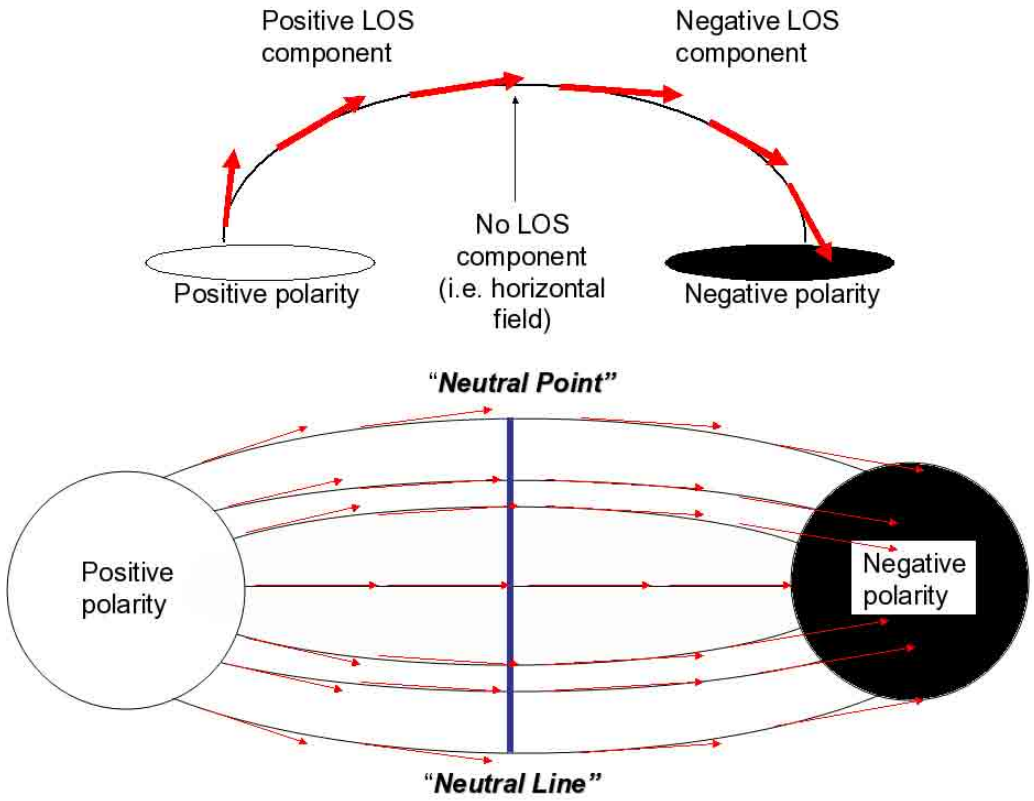

FIG. 5.33 Schematic of a simple neutral line geometry. From the continuity and divergence-free conditions of MHD, the neutral line must be continuous and must not cross itself.

enforced by bitwise shifting of each pixel to identify the next pixel on the neutral line. As can be seen from the figure, the central neutral line (expectedly) lies between the regions of positive and negative flux, acting as a topological separator between opposite flux regions. There are several other neutral line regions, as can be seen outside the boundaries of the main sunspot group. These belong to plage regions in the quiet-sun at the upper- and lower-left of the image. Other neutral line contours can be seen near the main-polarity lobes of the active region. These smaller contours typically outline small patches of polarity opposite to the that of their nearest umbral regions. Without subsequent images in a time-series, it is difficult to assess the character of these small patches of magnetic flux. However, they do appear to be similar to the moving magnetic features (MMFs) presented in Chapter 1. That is, they could be the results of arching penumbral magnetic fields that dive below

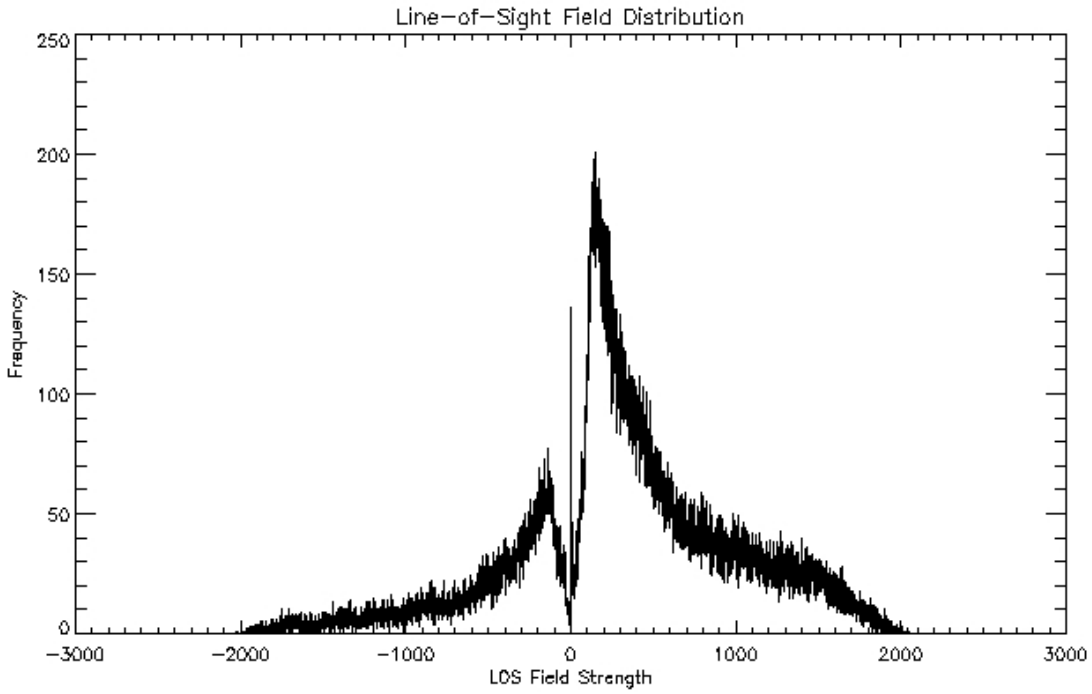

FIG. 5.34 Distribution of longitudinal magnetic flux in AR10956. The distribution of positive- and negative-polarity flux in AR10923 is plotted as a histogram (no binning). As can be seen, flux imbalance in this active region is evident in the asymmetric positive and negative lobes of the distribution.

the surface and subsequently re-emerge. The fact that these patches of flux are seen at the outer edges of the penumbral regions reinforces the idea of the "sea-serpent" model of MMFs which are carried outward by the moat flow surrounding active regions.

Since the central neutral line does not close on itself within the active region, there must be field-occupied regions outside the field-of-view. Therefore, we cannot make any strict judgements on the flux balance in this region, since the field-of-view effects are important in this case. However, we can still examine the distribution of flux within the active region. Figure 5.34 shows the flux distribution, and Table 5.1 displays the statistical parameters as moments of this distribution, as before.

As can be seen from Figure 5.34, not only is the flux imbalance evident in the grayscale flux image, but in the large positive lobe of the distribution as well. Further, the non-zero skew of the distribution indicates that there are large asymmetries around the mean field

TABLE 5.1 Statistical moments of the longitudinal flux distribution in AR10956.

Average (G)	Standard Deviation (G)	Skew	Kurtosis
321.6	660.1	-0.265	0.648

strength, which is evident from the figure. The variation of the distribution around the mean field strength of 321.6 G is very asymmetric, not only from the flux-imbalance but also from the presence of strong-field "bumps" in the tail of the positive lobe. A relatively symmetric (in terms of flux) bipolar sunspot would have a nearly Gaussian distribution of flux, and hence the kurtosis would tend to zero. Since it is obvious that the distribution is composed of two separate positive and negative distributions, the large kurtosis value is expected. All these characterizations help to identify unbalanced flux regions which may be flare-productive, and similar analyses could be carried out for other parameters of interest, as in Leka and Barnes (2003a,b). In the future, I plan to include such analyses in near real-time active region monitoring algorithms.

This chapter has presented the results of several genetic inversions on high spatial resolution data from the *Hinode* satellite. I have shown the method is not only capable of resolving the fine structure (particularly in the penumbra), but also can be employed to repeatedly perform inversions on sequential datasets to recover the evolution of the magnetic field strength and geometry. This is a key aspect for the continued application of this method to current and future large datasets, such as those produced by the SOLIS vector spectromagnetograph, located at Kitt Peak National Observatory, as well as is expected to be produced by the Advanced Technology Solar Telescope. However, the high spatial resolution and fast time cadence of these advanced instruments makes the current incarnation of the genetic inversion algorithm unsuitable for such an application. What is needed is a strategy to speed up the genetic inversion to much smaller timescales. Any reduction in execution speed will be beneficial, but the ultimate goal of such a strategy will be to increase the speed by such a factor as to make it comparable to the full field-of-view integration time needed by the spectropolarimeter, such that the current dataset of Stokes parameters

over the entire field-of-view can be inverted before the next dataset is obtained. This will open the door for near real-time "measurements" of the vector magnetic field, and will be extremely useful for characterizing the potential for flare activity in solar active regions.

CHAPTER 6

HARNESSING THE PARALLEL GENETIC ALGORITHM

The ultimate goal of this research is to produce high time-cadence views of the surface-level solar magnetic field. Unfortunately, the algorithm as I have described it in this dissertation is not currently capable of performing such a task. However, the genetic algorithm is perfectly suited for parallel processing, and it is this fact that we exploit to achieve a drastic speedup of the inversion process.

Assigning a single pixel to a single CPU node in a cluster is perhaps the most efficient way of parallelizing the inversion algorithm. Neglecting the overhead due to internode communication (master-slave architecture), if there are N pixels and n CPUs in the cluster, each running a serial version of the genetic algorithm, a theoretical speedup factor of

$$s = \frac{N}{n}$$

can be obtained for the time it takes to invert the whole field-of-view. For a slit-scan spectrograph like the ASP, the most straightforward way of doing this is to have one CPU for every pixel in the slitjaw. This gives a modest investment of about 230 CPUs. The field-of-view of the ASP is 300×230 pixels, giving a speedup factor of 300! The serial code takes approximately 3 hours to invert the entire ASP field-of-view, giving a rough estimate of approximately 30 seconds execution time for the parallelized version. This is competitive with the fastest known inversion techniques (PCA, ANN), discounting the lengthy time needed ($\sim$) to calibrate/train these methods (Socas-Navarro [2005]). Although a trained ANN can invert the ASP field-of-view in less than a minute (Socas-Navarro [2005]), these methods lack mathematical rigor and are essentially glorified interpolation routines, using finite-sized sets of either synthetic or real observations that have *already* been inverted as the interpolation support points. If the support points are biased or flawed in some way, the recovered parameters will be similarly biased. Furthermore, the training set may not include sufficient support for (e.g.) various line-of-sight flows or other quantities that vary wildly between two different active regions, and so the ANN inversion may not be able to

recover the details of different active regions. Furthermore, the PCA technique does not yield any physically-real interpretation of the PCA components, and the LOS magnetic field and velocity can only be obtained by a rather esoteric manipulation of the weak-field regime approximation within the framework of the PCA (Skumanich and López Ariste (2002)). Therefore, I am of the opinion that the PCA and ANN techniques belong more to the realm of the traditional LOS magnetogram, and their calibration/training times prevent them from being used in a practical multi-observation analysis in any reasonable time. Furthermore, there is no guarantee that an ANN trained on one active region can recover the magnetic structures of a different active region. They may require separate training epochs for every attempted inversion, and this would be prohibitively slow.

The other great advantage to the approach in this work is that the genetic algorithm control parameters (population size, number of generations, etc.) can be adjusted so that the average runtime matches the integration time (per slitjaw position) of the ASP. This should, in principle, allow near real-time inversions. That is, the data from one slitjaw position can be inverted in roughly the same amount of time as the integration time needed for the *next* slitjaw position.

6.1 Stage 1 Parallelization

This "stage" of course-grained parallelization involves splitting the workload among the nodes in a cluster, so that each node inverts the polarization profiles from a single pixel. The ASP has a spectrograph slit of 230 usable pixels, which means that a modestly-sized cluster of 230 nodes could invert the entire array of slit pixels in approximately the same time as a single node could invert a single pixel. In reality, this method will incur some communication overhead, but can be tuned to give the best results. For example, the average runtime for the inversion of a single pixel is about 0.5 seconds for a modestly-sized genetic population. This is somewhat smaller than the total integration time for the slit array, which means that we could potentially use much larger populations (and hence be much more confident in the results), in order to "fill up" the time between exposures.

6.2 The Message Passing Interface (MPI)

To construct the genetic inversion algorithm within the framework of phase 1 parallelization, I have employed a classic master-slave algorithm (or master-worker, for more politically-correct nomenclature) using MPICH2 libraries. This Message Passing Interface contains the framework for sending and receiving packets of data between simultaneously executing processes. The strength of this approach lies in its ability to transfer data to a separate process (executing program) which actually does the computationally expensive calculations, returns the result, and waits for more work to be assigned.

The core system calls of MPI are the MPI_SEND and MPI_RECV subroutines which, as the names suggest, actually effect the transfer of information from one process to another. In FORTRAN 90, the MPI_SEND subroutine is called as:

MPI_SEND(variable, size, MPI_VARIABLE_TYPE, destination, tag,
MPI_COMM_WORLD, error)

where "variable" is the information of type "MPI_VARIABLE_TYPE" to be sent and has dimension "size." The parameter "destination" identifies which process the information should be sent to, and "tag" conveys a little extra piece of information that may be useful in various bookkeeping tasks. The parameter "MPI_COMM_WORLD" conveys that the information could be sent to all detected processes, and "error" is an exit-code flag which returns 0 if everything went smoothly. For example, to have a process send a double-precision variable named "vec" of dimension 3 (a 3-element vector) to process 4, with a tag of 1, the following subroutine call should be made:

CALL MPI_SEND(vec, 3, MPI_DOUBLE_PRECISION, 4, 1, MPI_COMM_WORLD,
error)

In this case, the tag could be interpreted as denoting that "vec" is actually the first row of a matrix. The tag is therefore vitally important when splitting up a workload, as it identifies which particular chunk of the workload a process is working on. However, in order for the

communication to take place, a process that is being sent a piece of information must post a *matching* MPI_RECV call. The MPI_RECV subroutine is called as:

MPI_RECV(variable, size, MPI_VARIABLE_TYPE, source, tag,
MPI_COMM_WORLD, status, error)

where "variable", "size", "MPI_VARIABLE_TYPE", and "tag" *must* match the values specified in the MPI_SEND subroutine call. The parameter "source" identifies which process sent the information to be received. The parameter "status" is a vector of supplementary information akin to the tag, and again "error" is an exit-code flag. For example, if process 0 sent the vector "vec" to process 4 as in the above example, the transfer would be complete only if process 4 made the following subroutine call:

CALL MPI_RECV(vec, 3, MPI_DOUBLE_PRECISION, 0, 1, MPI_COMM_WORLD,
status, error)

This incarnation of the parallel genetic inversion algorithm belongs to the realm of the EPC; that is, Embarrassingly Parallel Computation. That is, the very nature of the inversion algorithm developed in this work demands to be implemented in a parallel fashion. Every pixel in the field-of-view can be inverted without regards to its nearest neighbors, and so should be done in parallel. The acronym EPC applies here, since it would be embarrassing to use such an inherently parallel algorithm in a strictly serial implementation. Specifically, the parallel genetic inversion is accomplished as follows: a single process is dubbed the "master" process, and is in charge of dispatching the proper pixel coordinates to the many "slave" processes, each of which performs a separate genetic inversion. Each slave process then returns the resulting vector of model parameters to the master process, which properly assembles them into maps that display the vector magnetic field as well as the secondary thermodynamic parameters. After a slave process sends the resulting vector back to the master process, a new set of coordinates are dispatched to that slave process. Operating in this asynchronous fashion, there is very little idle time, as the transmission of a completed

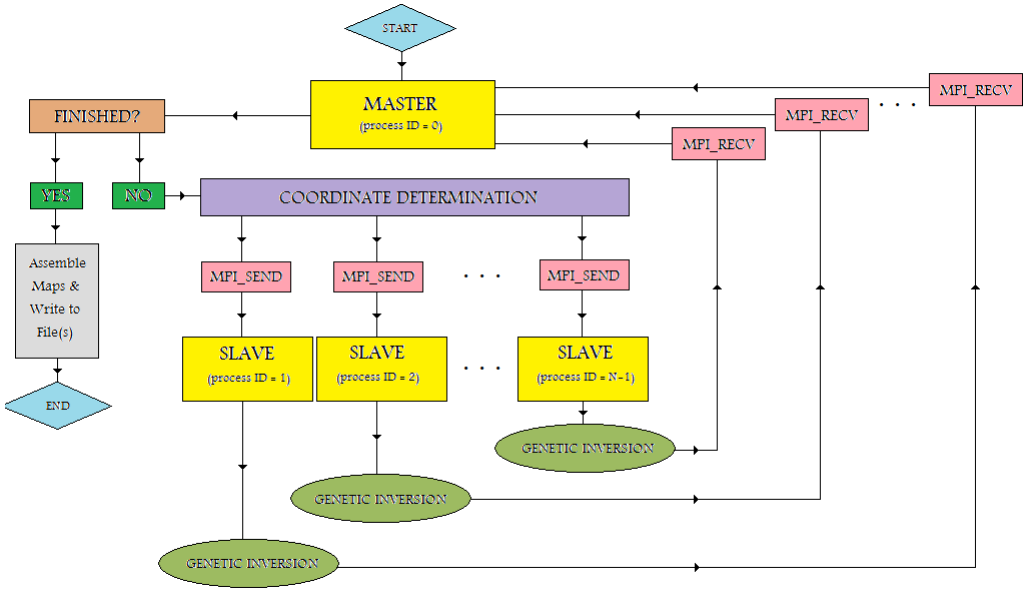

FIG. 6.1 Flowchart for the parallel genetic inversion. The figure shows the conceptual design and operation of the parallelized genetic inversion algorithm. There are N total nodes/processors/CPUs/cores, each of which runs its own process (program). Process 0 is the Master node, and is responsible for distributing the data and collecting the results, and there are $N-1$ Slave nodes to actually perform the inversions. The varying arrow lengths to and from the "MPI_RECV" calls are meant to represent the asynchronous nature of the parallel code, whereby each pixel is inverted at a completely independent rate, and Slave processes can accept new work at their own paces.

piece of work is immediately followed by a request for more work. A flowchart demonstrating the operation of the parallel algorithm is shown in Figure 6.1.

6.3 Results and Timing

Amdahl's Law (Amdahl (1967)) is frequently cited as an estimate of the speedup one can hope to obtain by parallelizing a particular computation. Let S be the fraction of the computation (or computation time, if you prefer) that is *inherently* serial. This may involve reading data from a file, various bookkeeping tasks, etc. Then $1-S$ is the fraction of the computation that can be parallelized, or carried out in parallel among several different

processes. If one defines the rate of computation for a single process (i.e., serial code) as

$$R_1 = \frac{Work}{T_s + T_p}, \tag{6.1}$$

where $T_{s,p}$ represents the amount of time the program spends doing serial and potentially parallel calculations, respectively, then the theoretical rate at which the same work can be done on N processors is

$$R_N = \frac{Work}{T_s + \frac{T_p}{N}}. \tag{6.2}$$

If we define $S \equiv T_s/T_{tot}$, then the expected speedup to be obtained by using N parallel processes can be shown to be

$$Speedup \equiv \frac{R_N}{R_1} = \frac{1}{S + \frac{(1-S)}{N}}. \tag{6.3}$$

For EPCs, $S \rightarrow 0$, such that the the parallel program executes faster by a factor of N, which is the theoretical maximum speedup that scales linearly with the number of parallel processes.

To profile the timing of the parallel genetic inversion, I have included in the algorithm various stopwatch subroutine calls to measure the time taken to send the result of one inversion, from the slave that performed it, back to the master process. It should be noted that these times are not general, but *only apply to the specific architecture (AMD Turion64X2) used in this work*. After adding all the measured time intervals required to send the results back, and dividing by the number of pixels, the average communication time was determined to be 0.088 ms, i.e., extremely fast! To complement this measure, I have also profiled the average computation time for a single-pixel inversion using the genetic algorithm control parameters adopted for this work. The average runtime for a single pixel was determined to be 20 ms, or a factor of 227 times longer than the actual communication requires. Finally, we can note that the typical runtime for the 415 x 509 pixel field-of-view from the *Hinode* data is about 80 minutes, while the time spent communicating the inversion results back to the master process totals only 16.90 seconds. Therefore, it is very clear that the total communication time for sending and receiving the results of all the

genetic inversions performed by the slaves is very much less than the time it takes to actually perform the inversions. Although not unexpected, this is a very nice result that provides the verification that a parallel genetic inversion of entire spectropolarimetric datasets may be very efficient and expedient.

The application of this analysis to the parallel genetic inversions is especially promising; given a field-of-view of N_p *independent* pixels over which to produce a vector magnetogram, we expect to observe the nearly linear scaling when utilizing N parallel processes, each hosted on a separate CPU or core. In order to test the effectiveness of partitioning the workload over multiple processes, I begin by performing the parallel genetic inversions on a single workstation utilizing an AMD Turion 64X2 processor, with various numbers of slave processes. This particular processor is a dual-core CPU, so that the limit of my parallel testing will be 2 slave processes, each executing on a single core. I expect that the runtime for these various trials should be nearly identical if the workload is being partitioned effectively among the different processes. For example, a run with only a single slave is identical to the serial version of the code, and that slave would utilize 100% of the processing power of a single core. With 2 slaves, each would be able to utilize 100% of each core, and so should perform the inversion in about half the time of the serial case. If this turns out to be correct, we may be able to extrapolate the speedup to larger numbers of cores. However, we must be sure that the workload is being effectively partitioned among the processes, and so I continue this analysis beyond the 2 maximally operating processes available to my workstation. With 3 slaves, each only has 33% of the *total* computing power since we are overscheduling the cores of the processor (i.e., more than one process per core), so by extension, a parallel run with $N > 2$ slave processes only allocates a fraction $1/N$ of the total computing power to each slave, such that the time to complete those N inversions should be about the same as the 2-slave parallel code performing inversions on N pixels. Table 6.1 shows the scaling behavior for various numbers of slave processes for the parallel code executing *on the single processor*.

As can be seen from Table 6.1, the runtimes for both the purely serial algorithm and the

TABLE 6.1 Efficiency of the division of labor within the parallel genetic inversion. The inversions were all performed with a population size of 50 evolving over 20 generations on a single workstation. The serial version of the code is, for all intents and purposes, identical to the parallel version without the top-level MPI driver program.

Serial Execution Time:	151.32 minutes
Number of Slave Processes	Execution Time (minutes)
1	157.28
2	79.99
3	80.97
4	80.55
5	80.34
6	80.69
7	79.93
8	79.89
9	79.80
10	79.86
20	80.23
50	79.51

parallel algorithm utilizing only a single slave are essentially identical. As expected, when two slave processes are used to do the actual work of the inversions, we reduce the runtime by about a factor of 2. Although there is danger in extrapolating from such a limited set of data, this gives the impetus to hypothesize that a cluster of N processors/cores will indeed yield an implementation of the parallel genetic inversion that is a factor N faster than the serial genetic inversion. Table 6.2 presents some extrapolations of the speedup to be gained in this manner. Further evidence in support of this prediction is the constancy of the runtime for multiple slave processes. Since these runtimes are essentially equal, this indicates efficient division of the workload, further strengthening the assumption of a linear

TABLE 6.2 Projected runtimes for the parallel genetic inversion. These projections pertain to the parallel genetic inversion runtime for execution on a cluster with variable numbers of independent computing cores. The quoted figures would represent the time to complete an inversion of the 415 × 509 pixel *Hinode* dataset from 14 November 2006, presented in Chapter 5. However, the runtimes for other datasets in the series would be slightly shorter, since the other sets contain more quiet-sun pixels with negligible polarization, and hence are ignored in the inversion.

Single Core Execution Time:	157.28 minutes
Number of Cores	Execution Time (minutes)
2	79.99
3	52.43
4	39.32
5	31.46
6	26.21
7	22.47
8	19.66
9	17.48
10	15.73
15	10.49
20	7.86

speedup factor.

While these tests show that the genetic inversion can be efficiently split into distinct parallel inversions, it is important to note that the tests were carried out on a single workstation. Therefore, a true quantitative analysis of the speedup factor for a parallel genetic inversion on a cluster of multiple CPUs has yet to be performed. However, I am confident that the tests prove that we can expect a near-linear scaling of the speedup factor with the number of processors. The next step in this work is to actually perform these cluster-based tests, and I now present an efficient model design for a small, inexpensive, but powerful desktop cluster for which I hope to be able to secure the funding to build.

6.4 An Efficient Design Model for a Desktop Cluster

Modern compute clusters are typically built from off-shelf components (processor, memory, network, power, cooling). While a convenient way to accomplish this is to purchase several CPU towers with all this hardware already in place, this could be rather expensive for high-end computing solutions. This section presents several theoretical design considerations for building a cheap but powerful "desktop supercomputer."

The PlayStation 3 (PS3) gaming console is perhaps considered one of the best gaming platforms on the market today. Its powerful processor runs highly parallelized, multithreaded codes for the graphics processing of its games. Internally, the PS3 sports a very advanced Cell processor. Composed of 8 separate cores operating at 3.2 GHz, it is essentially a radically-advanced version of a standard graphics processor unit (GPU) that is found in every desktop or laptop computer. GPUs are designed for handling floating-point computations, and so may also be used for scientific computation.

PS3 also has its own internal onboard memory of 512MB, which is hard-docked to the system, and therefore cannot be easily upgraded. Although more memory would be desirable, this presents no technical challenge to the operation of the genetic inversion; even with large population sizes, the genetic algorithm rarely uses more than 256 MB of memory. It also has its own internal hard disk drive that is large enough to store many spectropolarimetric datasets, and has its own network controller. It should be fairly obvious by now that all these components essentially describe a typical computer. In essence, the PS3 is a state-of-the-art computer disguised as a video game platform. What is even more impressive is the following: the PS3s internal architecture was designed in partnership between Sony, IBM, and Toshiba to be an *open* platform. That is, software development on/for its standard architecture is not shut off from the public through the use of proprietary hardware and software drivers. As such, the systems of the PS3 can be *hacked* to function as a normal desktop computer! Even better, with some effort, several PS3s can be setup to function as a powerful compute cluster. The PS3 is capable of running several Linux-

based operating systems (e.g., Fedora Core 5, Yellow Dog Linux) that make it ideal for software development and parallel programming. With their small form-factor, a cluster of many PS3s could easily fit on an office desk, and they require *no* special power or cooling considerations. Keeping in mind that each PS3 has 7 functional cores, a small desktop cluster of 8 PS3 consoles provides 56 independent processing cores for scientific computation, all for a grand total of around $3000 USD. An equivalent set-up, designed and built from scratch by industry professionals, is likely to cost over $100,000 USD, making this design a very attractive alternative to high-performance computing. It is my aim to lobby for NSO funding to build such a desktop cluster, and to use it for executing blazing-fast inversions of large spectropolarimetric datasets. For example, a liberal estimate of the inversion time for the *Hinode* datasets used in this work on the 56-core PS3 cluster (not including communication overhead) is 2.8 minutes. Granted, ignoring communication overhead in a cluster of more than roughly 10 nodes is not justified, as the overhead typically increases with increasing numbers of nodes, but since SOLIS full-disc spectropolarimetric scans require on the order of 20 minutes, near real-time, model-based vector magnetic field diagnostics could be a reality.

CHAPTER 7
CONCLUSIONS

7.1 Summary

This thesis has investigated the application of genetic algorithm techniques to the inversion of spectropolarimetry datasets in an effort to recover the vector magnetic fields of sunspots and active regions at the photospheric level. I have presented several perspectives on solar magnetic fields and their manifestations, as well as the pertinent radiative transfer physics used in the inversion of data obtained at the National Solar Observatory, as well as with the spectropolarimeter aboard the *Hinode* satellite. Through the exploration of the genetic algorithm mechanism in Chapter 4, I have shown that they are indeed suitable for solving the inverse problem, and present unique opportunities for fine-tuning their highly problem-dependent performance. Chapter 5 presented the inversion results for a unipolar sunspot, and it was shown to recover the expected physics of the umbra and penumbral fine structure. The major conclusions of this work are as follows:

- Genetic algorithms are an attractive alternative to the Stokes inversion problem, and are capable of accurately and repeatedly inferring the magnetic properties of sunspots and active regions.

- The implementation of the genetic inversion in this work is robust and stable, and produces results that are consistent with other inversion techniques applied to the same data.

- The genetic inversion is capable of producing such measurements of the magnetic field on timescales that are several factors smaller than other inversion techniques.

- The current incarnation of the genetic inversion is capable of handling data from a variety of different instruments, with minimal change to the control parameters of the algorithm.

- The genetic inversion is highly-parallel, and I have developed a parallel algorithm, using the MPICH2 libraries, which shows the potential for making near real-time measurements of the vector magnetic field a reality.

7.2 Future Work

Although this work already shows very promising results, there are some key places where I believe improvements can be made. The areas I have targeted for improvement are as follows:

- Extension to full Fe I multiplet at 6301.5 and 6302.5Å for Hinode data. The 6302.5Å line as observed by ground-based solar telescopes is contaminated in its red wing by a terrestrial absorption feature. Since *Hinode* is in Earth-orbit, no such contamination is present in the data. Since both absorption lines of this multiplet are formed in roughly the same atmospheric conditions, the genetic inversion should be better able to locate the optimal vector magnetic field, given two absorption lines of different magnetic sensitivity. However, the 6302.5Å line is formed in a slightly higher photospheric layer than the 6301.5Å line, and so there will necessarily be a larger parameter space, i.e., one must include different Doppler line-widths for each line.

- Extension of the Milne-Eddington inversion to a fully depth-dependent inversion with a genetic implementation of the DELO method. This is an iterative solution to the polarized radiative transfer equations, which will require a larger encoded genotype in the genetic algorithm, in order to specify the atmosphere at multiple optical depths. The DELO method is the workhorse of the most sophisticated depth-dependent inversion routine to date, named SIR (**S**tokes **I**nversion by **R**esponse Functions). Specifically, this will allow the asymmetries in the Stokes polarization profiles caused by gradients of the model parameters along the line-of-sight to be accounted for.

- Extension of the parallel genetic inversion to encompass more fine-grained parallelism. Since the genetic algorithm requires a large number of fitness function evaluations, it

may be beneficial to parallelize the evaluation of the population at each generation. That is, instead of a evaluating the population serially (one individual at a time within a DO loop), the entire population could be evaluated in parallel. One final level of parallelism would be to parallelize the fitness function itself. The current incarnation of the genetic inversion calculates the entire line profile over all the wavelengths of interest. Parallelizing the fitness function would require each slave process to be responsible for a single wavelength bin only, and, provided the communication time is less than the line-profile computation time, even more speed could be gained from this approach.

As a final note, I wish to mention that the genetic inversion code developed and presented in this work will be placed in the public domain, as part of the High Altitude Observatory's Community Inversion Code (CIC) project. Furthermore, I am currently in the process of cleaning up the genetic algorithm code, commenting it heavily, and producing a "User's Guide". The genetic algorithm itself is a massively-customizable optimization algorithm, with many potential genetic operators, control settings, and modes of genetic search. As such, it will require quite a bit of documentation for users that are not familiar with its customizability. Once complete, the general-purpose genetic algorithm function optimizer will also be placed in the public domain and will be freely-available on the Internet.

REFERENCES

Amdahl, G.M., 1967, in Proc. AFIPS Conf. (Vol. 30; Reston, VA: AFIPS Press)

Arny, T.T., & Schneider, S.E. 2008, in Explorations: An Introduction to Astronomy (New York, NY: McGraw-Hill), 374

Auer, L.H., Heasley, J.N., & House, L.L., 1977a, ApJ, 216, 531

———. 1977b, Solar Phys., 55, 47

Bäck, T., 1996, Evolutionary Algorithms in Theory and Practice (New York, NY: Oxford University Press)

Bao, S.D., Zhang, H.Q., Ai, G.X., & Zhang, M., 1999, Astron. Astrophys. Suppl. Series, 139, 311

Baron, E., & Hauschildt, P.H., 2007, arXiv preprint (astro-ph/0703437v1)

Basu, S., & Antia, H.M., 2003, ASP Conference Series, 293, 250

Beasley, D., Bull, D.R., & Martin, R.A., 1993, in Evolutionary Computation (Cambridge, MA: MIT Press)

Bellot Rubio, L.R., Balthasar, H., & Collados, M., 2004, Astron. Astrophys., 427, 319

Bernasconi, P.N., & Solanki, S.K., 1996, Solar Phys., 164, 227

Borrero, J.M., & Bellot Rubio, L.R., 2002, Astron. Astrophys., 385, 1056

Brandt, P.N., Mauter, H.A., and Smartt, R., 1987, Astron. Astrophys., 188, 163

Canfield, R.C., Hudson, H.S., & McKenzie, D.E., 1999, Geophys. Res. Let., 26, 627C

Carlson, S.E., & Shonkwiler, R., 1998, in Proc. IEEE Int'l Conf. on Systems, Man, and Cybernetics, 4, 3931

Carroll, T.A., & Staude, J., 2001, Astron. Astrophys., 378, 326

Charbonneau, P., 2002, NCAR Technical Note: Release Notes for PIKAIA 1.2

Charbonneau, P., Christensen-Dalsgaard, J., Henning, R., Larsen, R. M., Schou, J., Thompson, M. J., & Tomczyk, S., 1999, ApJ, 527, 445

Chatterjee, P., Nandy, D., & Choudhuri, A.R., 2004, Astron. Astrophys., 427, 1019

Darwen, P., & Yao, X., 1995, in Proc. IEEE Int'l Conf. on Evolutionary Computation, 1, 166

Deb, K., & Goldberg, D.E., 1989, in Proc. Third Int'l Conf. Genetic Algorithms (San Mateo, CA: Morgan Kaufmann Inc.)

Dick, G., 2005, in 17^{th} Annual Colloq. of the Space Information Research Center

Elmore, D.F., Lites, B.W., Tomczyk, S., Skumanich A.P., Dunn, R.B., Schuenke, J.A., Streander, K.V., Leach, T.W., Chambellan, C.W., Hull, H.K., & Lacey, L.B., 1992, in SPIE Vol. 1746: Polarization Analysis and Measurement

Eydenberg, M.S., Balasubramaniam, K.S., & López Ariste, A., 2005, ApJ, 619, 1167

Gan, J., & Warwick, K., 2000, in Proc. Genetic & Evolutionary Computation Conf., 1, 96

———. 2001, in Proc. of 2001 Congress on Evolutionary Computation, 1, 215

———. 2002, in Proc. of 2002 Congress on Evolutionary Computation, 1, 43

Gary, G.A., & Hagyard, M.J., 1990, Solar Phys., 126, 21

Georgoulis, M.K., LaBonte, B.J., & Metcalf, T.R., 2004, ApJ, 602, 446

Gingerich, O., Noyes, R.W., Kalkofen, W., & Cuny, Y., 1971, Solar Phys., 18, 347

Goldberg, D.E., 1989, Genetic Algorithms in Search, Optimization, and Machine Learning (Reading, PA: Addison Wesley)

Gullixson, C., 2007, DLSP Operation Manual (Sunspot: NSO/Sacramento Peak)

Hagyard, M.J., 1971, Solar Phys., 16, 286

Hirzberger, J., Stangl, S., Gersin, K., Jurcák, J., Puschmann, K.G., & Sobotka, M., 2005, Astron. Astrophys., 442, 1079

Holland, J.H., 1975, Adaptation in Natural and Artificial Systems (Ann Arbor: University of Michigan Press)

Ichimoto, K., Shine, R.A., Lites, B., Kubo, M., Shimizu, T., Suematsu, Y., Tsuneta, S., Katsukawa, Y., Tarbell, T.D., Title, A.M., Nagata, S., Yokoyama, T., & Shimojo, M., 2007, Publ. Astron. Soc. Japan, 59, S593

Ichimoto, K., Tsuneta, S., Suematsu, Y., Katsukawa, Y., Shimizu, T., Lites, B.W., Kubo, M., Tarbell, T.D., Shine, R.A., Title, A.M., & Nagata, S., 2008, Astron. Astrophys, 481, L9

Jackson, J.D., 1998, Classical Electrodynamics, 3^{rd} Edition (New York, NY: Wiley Publishing)

Jones, T. & Forrest, S., 1995, in Proc. of the Sixth Int'l Conf. on Genetic Algorithms (San Francisco, CA: Morgan Kaufmann, Inc.)

Katsukawa, Y., Yokoyama, T., Berger, T.E., Ichimoto, K., Kubo, M., Lites, B.W., Nagata, S. Shimizu, T., Shine, R.A., Suematse, Y., Tarbell, T.D., Title, A.M., & Tsuneta, S., 2008, arXiv preprint (arXiv:0709.2527v1)

Kichatinov, L.L., & Rüdiger, G., 1996, Astronomy Letters, 22 (2), 279

Kitai, R., Watanabe, H., Nakamura, T., Otsuji, K., Matsumoto, T., UeNo, S., Nagata, S., & Shibata, K., 2007, Publ. Astron. Soc. Japan, 59, S585

Kosugi, T., Matsuzaki, K., Sakao, T., Shimizu, T., Sone, Y., Tachikawa, S., Hashimoto, T., Minesugi, T., Ohnishi, A., Yamada, T., Tsuneta, S., Hara, H., Ichimoto, K., Suematsu, Y., Shimojo, M., Watanabe, T., Shimada, S., Davis, J.M., Hill, L.D., Owens, J,K., Title, A.M., Culhane, J.L., Harra, L.K., Doschek, G.A., & Golub, L., 2007, Solar Phys., 243, 3

Kubo, M., Shimizu, T., & Tsuneta, S., 2007a, ApJ, 659, 812

Kubo, M., Ichimoto, K., Shimizu, T., Tsuneta, S., Suematsu, Y., Katsukawa, Y., Nagata, S., Tarbell, T.D., Shine, R.A., Title, A.M., Frank, Z.A., Lites, B., & Elmore, D., 2007b, Publ. Astron. Soc. Japan, 59S, 607

Lagg, A., Woch, J., Solanki, S.K., & Gandorfer, A., 2006, in ASP Conference Series: Solar Polarization 4, Vol. 358

Landi Degl'Innocenti, E., 1976, Astron. Astrophys. Supp., 25, 379

Landi Degl'Innocenti, E., & Landi Degl'Innocenti, M., 1977, Astron. Astrophys., 56, 111

———. 1985, Solar Phys., 97, 239

Landolfi, M., & Landi Degl'Innocenti, E., 1982, Solar Phys., 78, 355

Leka, K.D., 2001, in ASP Conference Proceedings: 20^{th} NSO/Sac Peak Summer Workshop, 236, 571

Leka, K.D., & Barnes, G., 2003a, ApJ, 595, 1277

———. 2003b, ApJ, 595, 1296

Leka, K.D., & Skumanich, A., 1998, ApJ, 507, 454

Lites, B.W., Elmore, D.F., Streander, K.V., Sankarasubramanian, K., Rimmele, T.R., & Sigwarth, M., 2003, ASP Conference Proceedings, Vol. 307, 324

Liu, Y., and Zhang, H., 2002, COSPAR Colloquia Series, Vol. 14, 55

López Ariste, A., & Semel, M., 1999, Astron. Astrophys., 350, 1089

Love, J.J., 1999, Astron. & Geophys., 40, 6.14

Mahfoud, S.W., 1994a, in Proc. of the First IEEE Conf. on Evolutionary Computation

———. 1994b, in Third Workshop on Foundations of Genetic Algorithms

———. 1995, in Proc. Sixth Int'l Conf. on Genetic Algorithms (San Francisco, CA: Morgan Kaufmann Inc.)

Maresky, J., Davidor, Y., Gitler, D., Aharoni, G., & Barak, A., 1995, in Proc. Sixth Int'l Conf. on Genetic Algorithms (San Francisco, CA: Morgan Kaufmann Inc.)

Maselli, A., Ferrara, A., & Ciardi, B., 2003, arXiv preprint (astro-ph/0307117v1)

Matsui, K., 1999, in Proc. IEEE Int'l Conf. on Systems, Man, and Cybernetics, 1, 625

Metcalf, T.R., Leka, K.D., Barnes, G., Lites, B.W., Georgeoulis, M.K., Pevtsov, A.A., Balasubramaniam, K.S., Gary, G.A., Jing, J., Li, J., Liu, Y., Wang, H.N., Abramenko, V., Yurchyshyn, V., & Moon, Y.-J., 2006, Solar Phys., 237, 267

Michalewicz, Z., 1995, in Proc. Sixth Int'l Conf. on Genetic Algorithms (San Francisco, CA: Morgan Kaufmann, Inc.)

Miller, B.L., & Shaw, M.J., 1995, in IlliGAL Report No. 95010

Moore, C., 1945, A Multiplet Table of Astrophysical Interest (National Bureau of Standards Technical Note)

Moriyasu, S., Kudoh, T., Yokoyama, T., & Shibata, K., 2004, ApJ, 601, L107

Parker, E.N., 1974, ApJ 189, 563

Pierce, A.K., & Breckinridge, J.B., 1973, The Kitt Peak Table of Photographic Solar Spectrum Wavelengths (Kitt Peak National Observatory: National Science Foundation)

Pogorelov, N.V. & Zank, G.P., 2004, in 35^{th} COSPAR Scientific Assembly, 18 - 25 July 2004, 3192

Press, W.H., Teukolsky, S.A., Vetterling, W.T., & Flannery, B.P., 1986, Numerical Recipes in Fortran 77 (New York, NY: Cambridge University Press)

Rachkovsky, D.N., 1962, Izv. Krymsk. Astrofiz. Obs., 27, 148

Ravindra, B., 2006, Solar Phys., 237, 297

Rees, D.E., Murphy, G.A., & Durrant, C.J., 1989, ApJ, 339, 1093

Rees, D.E., López Ariste, A., Thatcher, J., & Semel, M., 2000, Astron. Astrophys., 355, 759

Rijkhorst, E.-J., Plewa, T., Dubey, A., & Mellema, G., 2005, arXiv preprint (astro-ph/0505213v1)

Rimmele, T., & Marino, J., 2006, ApJ, 646, 593

Roberts, B., & Webb, A.R., 1978, Solar Phys., 56, 5

Ronald, E., 1995, in Proc. Sixth Int'l Conf. on Genetic Algorithms (San Francisco, CA: Morgan Kaufmann, Inc.)

Roudier, Th., Bonet, J.A., & Sobotka, M., 2002, Astron. Astrophys., 395, 249

Ruiz Cobo, B., & del Toro Iniesta, J.C., 1992, ApJ, 398, 375

Rumelhart, D.E., Hinto, G.E., & Williams, R.J., 1986, in Parallel Distributed Processing (Cambridge, MA: MIT Press)

Rybicki, G.B., & Lightman, A.P., 1979, Radiative Processes in Astrophysics (New York, NY: John Wiley & Sons, Inc.)

Sánchez Almeida, J., 1992, Solar Phys., 137, 1

Sánchez Almeida, J., and Lites, B.W., 1992, ApJ, 398, 359

Sareni, B., & Krähenbühl, L., 1998, in IEEE Transactions on Evolutionary Computation 2, 3, 97

Schlichenmaier, R., 2003, in ASP Conference Series, 286, 211

Schlichenmaier, R., Jahn, K., & Schmidt, H.U., 1998, Astron. Astrophys., 337, 897

Shir, O.M., & Bäck, T., 2006, in Parallel Problem Solving from Nature - PPSN IX (Berlin/Heidelberg: Springer)

Sidlichovsky, M., 1975, Bull. Astron. Inst. Czech., 27, 71

Skumanich, A., 2001, in 20^{th} International Sacramento Peak Summer Workshop on Advanced Solar Polarimetry - Theory, Observation, and Instrumentation (ed. M. Sigwarth), 236, 543

Skumanich, A., & Lites, B.W., 1987, ApJ, 322, 473

Skumanich, A., Lites, B.W., Martínez Pillet, V., & Seagraves, P., 1997, ApJ Supp. Series, 110, 357

Skumanich, A., & López Ariste, A., 2002, ApJ, 570, 379

Socas-Navarro, H., 2005, ApJ, 621, 545

Solanki, S.K., & Montavon, C.A.P., 1993, Astron. Astrophs., 275, 283

Spruit, H.C., 1979, Solar Phys., 61, 363

Spruit, H.C., & Roberts, B., 1983, Nature, 304, 401

Stenflo, J., 1994, Solar Magnetic Fields: Polarized Radiation Diagnostics (Dordrecht: Kluwer Academic Publishers)

Tian, Y., Liu, Y., & Wang, J., 2002, 34th COSPAR Scientific Assembly (meeting abstract), 48T

Tian, L., Wang, J., & Wu, D., 2002, Solar Phys., 209, 375

Tomassini, M., Vanneschi, L., Collard, P., & Clergue, M., 2005, in Evolutionary Computation, Vol. 13, Issue 2 (Cambridge, MA: MIT Press)

Unno, W., 1956, Pub. Astr. Soc. Japan, 8, 108

Wang, H., 2006, ApJ, 649, 490

Wang, T., Xu, A., & Zhang, H., 1994, Solar Phys., 155, 99

Wilson, R.C., 1980, Science, 207, 177

Yang, G., Yan, X., Cao, W., Wang, H., Denker, C., & Rimmele, T.R., 2004, ApJ, 617, L151

Zhang, H., Zi, G., Yan, X., Li, W., & Liu, Y., 1994, ApJ, 423, 828

Zhi-Hua, C., and Jian-Chao, Z., 2002, Proc. of 1^{st} International Conf. on Machine Learning and Cybernetics, 1429

CURRICULUM VITAE

BRIAN J. HARKER

164 Pioneer Avenue, Apt. 102, Logan, UT 84321 *(435) 512-1482; brian.harker@gmail.com*

Professional Interests

- Spectropolarimetry and the measurement of solar magnetic fields from Stokes polarization absorption profiles.
- Space weather forecasting, with emphasis on increasing the speed with which solar photospheric magnetic field diagnostics can be performed.
- Machine-learning and artificial intelligence systems (e.g. genetic algorithms, genetic programming, artificial neural networks) for autonomous data reduction and analysis.
- Design, construction, and use of low-cost, high-performance compute clusters (a.k.a. "desktop supercomputers").
- Cluster-computing and the design and implementation of parallel algorithms to bolster the efficiency of (or replace) current serial algorithms.

Education

- **Utah State University**—Logan, UT—*Graduate Education: 08/2003-present*
  - Ph.D. Program, Physics, Cumulative GPA: 3.96
  - Graduation Date: 1/2009
  - Coursework: spectropolarimetry, polarized radiative transfer, solar magnetic fields, atomic physics & the Zeeman effect, electrodynamics & plasma physics,

methods of mathematical & computational physics, quantum mechanics, classical mechanics, statistical mechanics

  - **Thesis Title: "On the Applicability of Genetic Algorithms to Fast Solar Spectropolarimetric Inversions for Vector Magnetography"**

    My Ph.D. thesis work involves the measurement of solar magnetic fields, using a novel genetic algorithm approach to solve the inverse problem by effectively and efficiently utilizing the solution to the forward problem. I have shown that the method is fast, accurate, and robust, and the results agree well with established inversion procedures that take over twice as much time for the same dataset. I have extended the applicability of the algorithm to datasets from a variety of different instruments, and have developed several parallel versions to be used in a cluster-computing environment, where I expect dramatic speedup factors to occur.

- **The University of Arizona**—Tucson, AZ—*Undergraduate Education: 08/1999-05/2003*
  - Bachelor of Science, Physics
  - Bachelor of Science, Astronomy
  - GPA: 3.641 (Cum Laude)

Experience

- **Research Assistant, Department of Physics, Utah State University**—Logan, UT—*08/03-present*
  - Developed sophisticated genetic algorithm techniques for real-world applications.
  - Performed computational work on inferring solar photospheric magnetic fields from Zeeman-induced polarization.

  - Extended previous serial computational work into parallel-programming regime using the MPICH2 and openMPI open-source implementations.
  - Performed theoretical and computational work involving extrapolation of magnetic field-guided flows in solar active regions.

- **Graduate Student Professor, Department of Physics, Utah State University**—Logan, UT—*06/2008-08/2008, 09/2007-12/2007*
  - USU 1040: Introduction to Astronomy (Fall 2007, Summer 2008)
  - Designed lecture material, course website, homework assignments, and project assignments.
  - Designed, implemented, and graded class-time examinations.
  - Mediated student excursions to the Physics Department Observatory.

- **Teaching Assistant, Department of Physics, Utah State University**—Logan, UT—*09/05-present*
  - PHYS 2200: Elements of Mechanics
  - PHYS 2210/2220: General Physics - Science & Engineering I/II
  - Lead two general review, homework workshops, exam preparation sessions, and two laboratory experiment sessions per week.
  - Helped set up and execute relevant physics demonstrations/experiments in classroom setting.
  - Designed, implemented, and graded class-time examinations.
  - Assumed lecture responsibilities when professor otherwise unavailable.

- **Summer Research Assistant, Department of Physics, Montana State University**—Bozeman,MT—*06/02-08/02*

  - Developed computational methods for examining properties of interplanetary magnetic fluxropes and progenitor active regions.
  - Executed analysis of data from WIND & ACE satellites, including the development of algorithms to identify variable fluxrope geometries.

- **Internship, Imaging Technology Laboratory, University of Arizona**—Tucson, AZ—*05/01-08/01*
  - Developed new machine methods for automating the protection of vital CCD components during the acid-etching process.
  - Performed assigned tasks in a sterile, clean-room environment.

Posters & Presentations

- Harker-Lundberg, B.J., Sojka, J.J., Balasubramaniam, K.S., "Extracting Sunspot and Flare Physics from Solar Spectropolarimetry", USU Graduate Research Symposium, 2 April 2008, Utah State University, Logan, UT 84321
- Harker-Lundberg, B.J., "Parallel Genetic Algorithms for Vector Magnetography", Invited Talk, Solar Data Assimilation Workshop, Utah State University, 25-26 October 2007, Logan, UT 84321
- Harker-Lundberg, B.J., Sojka J.J., Balasubramaniam, K.S., "Mapping Sunspot Magnetic Fields via Genetic Spectral Inversions", SHINE 2007 Workshop, 30 July 2007 - 3 August 2007, Whistler, BC, Canada
- Harker-Lundberg, B.J., Sojka J.J., Balasubramaniam, K.S., "A Niching Genetic Algorithm for Milne-Eddington Spectral Line Inverions", American Physical Society 4-Corners Meeting, 6-7 October 2006, Utah State University, Logan, UT 84321
- Harker-Lundberg, B.J., Balasubramaniam, K.S., Sojka, J.J., "Niching Genetic Algorithms for Milne-Eddington Spectral Line Inverions", SHINE 2006 Workshop,

31 July 2006 - 4 August 2006, Midway, UT 84049

- Harker-Lundberg, B.J., "Closing the Dynamo Loop: Sustained Magnetic Field Generation in the Sun", Public Outreach Presentation, Las Cruces Museum of Natural History, 25 July 2006, Las Cruces, NM 88011

- Harker-Lundberg, B.J., "Line-of-Sight Velocities in Newly Emergent Active Regions", Invited Talk, Utah State University Department of Physics Colloquium, 7 September 2004, Logan, UT 84321

Papers & Publications

- Leamon, R.J., Canfield, R.C., Jones, S.L., Lambkin, K., Lundberg, B.J., Pevtsov, A.A., *Journal of Geophysical Research*, Volume 109, Issue A5, CiteID A05106, 2004.

- Harker-Lundberg, B.J. "An Improved Genetic Algorithm for Stokes Spectropolarimetric Inversions" (in preparation)

- Harker-Lundberg, B.J. "On Fast Stokes Inversions for Vector Magnetography with Parallel Genetic Algorithms" (in preparation)

Honors & Awards

- Awarded the "Tomorrow Fellowship" in solar physics, jointly by the Department of Physics, Utah State University and the Space Dynamics Laboratory, Utah State University, 2003.

- Awarded the Lawrence R. and Abelina Megill Scholarship by the Department of Physics, Utah State University, 2006.

Skills

- Mathematical and computational problem-solving.
- Proficiency in IDL, FORTRAN, shell-scripting, and command-line interfaces.
- Working (but limited) knowledge of C/C++.
- Handling, reduction, and analysis of large datasets.
- Use of Microsoft Windows platforms, UNIX/Linux systems and their administration/maintenance, and Sun Solaris systems.
- Independent functionality with minimal direction.
- Public speaking/speaking to large audiences.
- Clear, concise, and effective communication, both orally and in writing.
- Moderately fluent in Spanish, with limited knowledge of other Latin-based languages.

References

- **Dr. Jan J. Sojka**—Department Head—*Department of Physics, Utah State University*
  - Logan, UT 84322
  - *Phone:* (435) 797-2964
  - *Email:* sojka@gaim.cass.usu.edu
- **Dr. K.S. Balasubramaniam**—Senior Research Astrophysicist—*USAF/AFRL, Solar Disturbances Prediction*
  - Sunspot, NM 88349
  - *Phone:* (575) 434-7134

  - *Email:* bala@nso.edu

- **Dr. Eric D. Held**—Associate Professor—*Department of Physics, Utah State University*
  - Logan, UT 84322
  - *Phone:* (435) 797-7166
  - *Email:* eheld@cc.usu.edu