2016

POINT SPREAD FUNCTION ESTIMATION AND UNCERTAINTY QUANTIFICATION

Kevin Thomas Joyce

POINT SPREAD FUNCTION ESTIMATION AND UNCERTAINTY QUANTIFICATION

By

Kevin Thomas Joyce

B.S., Montana State University, Bozeman, MT, 2006
M.S., Montana State University, Bozeman, MT, 2009

Dissertation

presented in partial fulfillment of the requirements
for the degree of

Doctorate of Philosophy
in Mathematics

The University of Montana
Missoula, MT

May 2016

Approved by:

Sandy Ross, Associate Dean of the Graduate School
Graduate School

Dr. Johnathan Bardsley, Co-Chair
Mathematical Sciences

Dr. Aaron Luttman, Co-Chair
Signals Processing and Applied Mathematics
National Security Technologies, LLC, Las Vegas, NV

Dr. Jon Graham
Mathematical Sciences

Dr. Peter Golubtsov
Department of Physics
Lomonosov Moscow State University

Dr. Leonid Kalachev
Mathematical Sciences

Joyce, Kevin, Doctorate of Philosophy, May 2016 Mathematics

Point Spread Function Estimation and Uncertainty Quantification

Committee Chairs: Johnathan Bardsley, Ph.D. and Aaron Luttman, Ph.D.

An important component of analyzing images quantitatively is modeling image blur due to effects from the system for image capture. When the effect of image blur is assumed to be translation invariant and isotropic, it can be generally modeled as convolution with a radially symmetric kernel, called the point spread function (PSF). Standard techniques for estimating the PSF involve imaging a bright point source, but this is not always feasible (e.g. high energy radiography). This work provides a novel non-parametric approach to estimating the PSF from a calibration image of a vertical edge. Moreover, the approach is within a hierarchical Bayesian framework that in addition to providing a method for estimation, also gives a quantification of uncertainty in the estimate by Markov Chain Monte Carlo (MCMC) methods.

In the development, we employ a recently developed enhancement to Gibbs sampling, referred to as partial collapse. The improved algorithm has been independently derived in several other works, however, it has been shown that partial collapse may be improperly implemented resulting in a sampling algorithm that that no longer converges to the desired posterior. The algorithm we present is proven to satisfy invariance with respect to the target density. This work and its implementation on radiographic data from the U.S. Department of Energy's Cygnus high-energy X-ray diagnostic system have culminated in a paper titled "Partially Collapsed Gibbs Samplers for Linear Inverse Problems and Applications to X-ray Imaging."

The other component of this work is mainly theoretical and develops the requisite functional analysis to make the integration based model derived in the first chapter rigorous. The literature source is from functional analysis related to distribution theory for linear partial differential equations, and briefly addresses infinite dimensional probability theory for Hilbert space-valued stochastic processes, a burgeoning and very active research area for the analysis of inverse problems. To our knowledge, this provides a new development of a notion of radial symmetry for L^2 based distributions. This work results in defining an L^2 complete space of radially symmetric distributions, which is an important step toward rigorously placing the PSF estimation problem in the infinite dimensional framework and is part of ongoing work toward that end.

Acknowledgments

This work, although ultimately channeled through me, is truly the product of the best advising, guidance, and overall support that anyone could hope for. I have been very lucky to have had mentors, teachers, colleagues, friends, and family that have guided and shaped the journey through my graduate studies and have given me much more than I deserve. There are too many to name here, and if you are omitted, know that it is not because I have forgotten you, but because the task of thanking everyone who deserves it is more than I am capable of.

I must first thank my two advisors, Dr. John Bardsley and Dr. Aaron Luttman, whose vision, generosity with ideas, and guidance have not only shaped this work into what it is, but also me as a mathematician. They set me right when I was lost and pushed me forward and encouraged me when the ideas were good. They graciously and patiently guided me along the path to becoming a mathematical researcher each in their own distinctly valuable way, and I have a unique skill set and perspective because of them, which I am truly grateful for.

The members of my committee deserve much thanks, not only for their input in this work, but also, and likely more importantly, for their advice and tutelage over my time at the University of Montana. Dr. Jon Graham instilled in me a solid appreciation for statistical thought and his unrelenting dedication to teaching was an inspiration. Dr. Peter Golubtsov was very generous with ideas and advice that have had a great influence on the direction of the work, and his insights and answers to my questions in courses and discussions were invaluable. Lastly, Dr. Leonid Kalechev guided me not just a teacher and mentor, but also as a diligent and expertly capable department chair for three of my four years at UM.

I must also thank Dr. Jen Brooks, whose early influence and encouragement taught me to be ever vigilant in the face of theoretical difficulties, for the fruits that they bear have been truly worth the effort. Many thanks to the other members of the analysis group, Dr. Greg St. George, Dr. Karl Stroetoff, and the late Dr. Thomas Tonev for the useful insights I gained in their seminar and in preparation for the analysis preliminary exam. Also, thanks to the current department chair and former graduate chair, Dr. Emily Stone and to the current graduate chair, Dr. Cory Palmer without whose leadership and help I would not have made it through.

I am very gratefule to have been financially supported from a Site Direct Research and Development grant from National Security Technologies LLC and the Mathematical Sciences department at the University of Montana. This support made this work possible. Also, the friendly staff at the Kettlehouse Southside provided many a beverage and listening ear when emotional currency was low.

Finally, the most thanks is due to those who supported me emotionally and personally through this journey. I am eternally grateful, as it is the relationships that I've built these last four years that I value more than anything else that I have gained. Thanks to those who were at one time or another, a fellow student with me for all the helpful conversations, mathematical or

otherwise, but mostly, the camaraderie and collegiality that made the experience of graduate school in Missoula truly unique. In particular, to Dr. Cody Palmer, Nhan Ngyuyen, Charlie Katerba, Dr. Marylesa Howard: with you, I have made life-long friends. I cannot express the magnitude of my gratitude to my close friends and family outside of mathematics – your value has and ever shall be beyond my ability to quantify – Dayne Robinson, Lora Weitzel, Alexis Billings, Corinne Casolar, Justin van Almelo, and Christa Carleton. To my family, and especially Mom, Dad, and Abbey: you are my core and my everlasting support; without you I would be lost.

This manuscript is dedicated to the memory of Paul Joyce, whose brilliance and grace were taken from us too early.

Notations

Image Model

p ..... Point spread function
b ..... Observed data
$\mathcal{G}$ ..... Forward edge-blur operator on a radial profile

Sets

$\mathbb{Z}$ ..... Integers
$\mathbb{R}$ ..... Real numbers
$\mathbb{C}$ ..... Complex numbers
$\mathbb{R}^k$ ..... k-dimensional vector space over real numbers

Functional Analysis

$(\cdot, \cdot)$ ..... Inner product
$\| \cdot \|$ ..... Norm
$\mathscr{X}^*$ ..... Space of continuous linear functionals on $\mathscr{X}$
$\langle \cdot, \cdot \rangle$ ..... Action of linear functional
$\mathscr{D}$ ..... Compactly supported smooth test functions
$\mathscr{D}^*$ ..... Space of distributions
L^2 ..... Square integrable distributions
$\mathscr{H}^n$ ..... Sobolev space of order n
$\mathscr{K}$ ..... Space of radially symmetric PSFs
$\mathscr{P}$ ..... Space of radial profiles
$\arg\min_{x \in A} \Phi(x)$ ..... The $x \in A$ such that $\Phi(x)$ is minimized
$\| \cdot \|_{TV}$ ..... Total variation semi-norm

Probability

$\mathbb{P}$ ..... Probabilty measure
$\mathbb{E}$ ..... Expectation operator
K ..... Transition kernel
$\mathcal{K}$ ..... Transition operator
$\pi(\cdot)$ ..... Probability density
$\pi(\cdot|\cdot)$ ..... Conditional probability density
$X \sim \pi$ ..... Random variable X has distribution p

Contents

List of Tables

List of Figures

List of Algorithms

Chapter 1

Images and Blur

1.1 Introduction

In addition to being a rich source of artistic and creative value, images (or more precisely, visual information from projections of light) are an important source of scientific information. Even the word 'observation' generally connotes visual perception, and its use as a catch-all for the measured verification of a hypothesis exemplifies the central role of vision in science. Many important scientific results have used visual information to discover and explain natural phenomena; for example, visual observations such as the color and shape of various plant organs in Gregor Mendel's experiments on hybridized peas formed the primary source of data for developing his model for genetic inheritance [Magner, 2002]. Arthur Eddington's 1919 image of the gravitationally lensed path of a comet during a solar eclipse provided the first experimental evidence supporting Albert Einstein's general theory of relativity [Dyson et al., 1920]. In these cases, only the *qualitative* components of the visual response and their relationship to the experiment were relevant. With the advent of the camera, photosensitive chemistry, and later digital imaging technology, high-fidelity recording of visual observations

as data became possible, allowing the potential to *quantitatively* analyze visual information. Digital images when viewed quantitatively, can be described as the response of the incidence of light, and the subsequent exchange of energy, on a grid of regularly spaced grid elements, which we refer to as pixels. The ever-progressing technologies in optical science and engineering are rapidly increasing the amount of data that can be measured in an image. Images are quickly becoming a source of 'big data' from which new methods are rapidly being developed to extract information from this rich data source. Yet there is a dichotomy between the amount of information available in image data and the complexity of how to quantitatively analyze it. That is, despite having a large volume of data to extract information, their spatial nature makes measurements at each pixel highly dependent on measurements at adjacent pixels, and an objective extraction of information taking this structure into account is not obvious. A quantitative analysis of the image cannot assume that measured values are independent, because it is precisely this lack of independence that makes an image interesting – independent image data is white noise (perhaps more appropriately, 'gray noise') – from which one can only infer the average of the measured pixels.

In addition to the difficulties related to spatial dependence and high resolution, the structure of the data may have more components than spatial aspects since light is measured on a spectrum. That is, the energy response of light is not univariate since it is frequency dependent, and in many cases, is measured at several fixed bands of frequency. For example, astronomical images captured by the interstellar robotic probes Voyagers I and II measured 5 bands in the visible spectrum on a pixel grid with dimension 800×800 [Showalter et al., 2006]. In medical imaging, computed tomography (CT) is an imaging process where a series of axial measurements of attenuated electromagnetic radiation are used to reconstruct a cross-section of a scanned object, and the geometry of the measurement system is a primary source of complexity for quantitative analysis since data are organized in a non-Cartesian parameterization [Epstein, 2008]. The primary focus of this work is analyzing pulsed X-ray measurements, referred to as a radiographs, which are used as a diagnostic of high-energy

physics experiments. Ultimately, analysis of the image is used to inform physical properties of the scene of interest, yet this measurement is indirect in several ways. For example, in several high-energy radiographic imaging systems, collimated X-rays are pulsed through a scene of interest, then the attenuated X-rays excite a crystal that responds by luminescing visible light at an intensity related to the energy of the attenuated wave-front. The light is then focused and measured on a high resolution array (on the order of 1000×1000 or more) of charge coupled devices (CCD) calibrated to count photons at a specified spectral band. Each of these examples highlights the potential depth and complexity of modeling image capture, and models must be sufficiently flexible in order to be realistic and reliable. Our approach will be to remove the constraints of a parametric form for describing the process of image blur so that we have more flexibility in capturing the effects of aggregate complexity, as well as allowing data to better characterize the form of the model.

A general treatment for quantitatively modeling image capture is to consider image data as response to a signalling system an ideal image or signal is filtered by some process forward modeled by the physics of the system and stochastic measurement effects. The ideal image is then estimated indirectly by 'reversing' the forward filter. When this process is modeled as an operator between function spaces, tools from functional analysis can be used to 'solve' the system. See one of the books [Vogel, 2002; Epstein, 2008] for an in-depth discussion and broad set of examples.

Relatively recently, techniques and ideas from formal data science have been used to not only model stochastic effects from measurement, but also the unknown ideal image. Bayesian techniques provide a natural way to incorporate prior knowledge as well as uncertainty in that knowledge. Formally, an image is modeled as a random field, and its correlation structure is the primary focus of study. Under the assumption of a Gaussian Markov random field, the correlation and mean completely describe the stochastic properties of the image [Rue and Held, 2005]. The books [Cressie, 1993; Rue and Held, 2005] provide an overview of the history and current methods for statistical methods for spatial data. The study of random fields, and

more specifically image data, has seen much development over the past half-century and has inspired a wealth of theory and computational tools, but is far from complete. Moreover, a broad field of scientific disciplines has considered image-like data in one way or another; fields such as astronomy, astrophysics, biology, medicine, geology, computer science, and nuclear physics to name a few. Each have a unique perspective on the problem, and a vast literature on the subject has been accumulated. Although much work has been done, it is still a very active research area and is far from achieving the level of consensus and understanding that analysis of independently sampled data has achieved. The aim of this work is to develop and adapt current models and methods for estimation and quantifying uncertainty to the specific component of image analysis related to blur introduced by the system for capturing images. Understanding this is an important component to the development and analysis of methods for quantitatively analyzing the images themselves.

1.1.1 Organization

This dissertation consists of four, more or less independent chapters that address novel contributions to aspects of modelling and estimating the point spread function for an imaging system that exhibits translation invariant and radially symmetric blur.

The remaining sections of this chapter provide the general background for modeling image blur and introduces how blurred images can be modeled with integral convolution with a point spread function (PSF). The main goal of this work will be to estimate this quantity in a rigorous way that also incorporates uncertainties of the estimate. The model we develop is completely novel, and the development is substantially more general than other parametrically dependent methods currently used. The modeling perspective will provide several mathematical forms (e.g. by changing variables) for describing the blurring of an opaque edge, with each providing a different insight into the process. We emphasize that each of these forms does not change the physical assumptions of the model and are used to prove and understand various

aspects of the underlying, intrinsic operation. Moreover, each of these different mathematical formulations provides several methods for analysis; e.g. one formulation shows that a blurred edge is sufficient for estimation, and another is useful for seeing that the problem is ill-posed. In this process, we will also show connections with other operator based models of active interest; specifically, the Radon and Abel transforms Epstein [2008].

In the second chapter, we derive the necessary theory and technical definitions to formally define the operator between separable Hilbert spaces. Again, this development takes several perspectives but does not change the fundamental action. The analysis provides a formal notion of the interaction of radially symmetric objects with non-radially symmetric objects. Chapter 2 is mainly theoretical, but the explicit forms for the discrete model of the forward operator and prior information are motivated and derived there. This chapter will use tools from the theory of distributions for partial differential equations (PDEs) to develop a novel notion of symmetry that retains the powerful structure of a complete inner product space.

For estimation, we take a Bayesian approach that allows for uncertainty quantification of the estimate, and Chapter 3 will lay out prerequisite Markov Chain theory and the general algorithms based on it that are used in this work. We derive and describe from first principles a recently developed enhancement to Gibbs sampling, which has been independently derived and studied in several other works [Van Dyk and Park, 2008; Agapiou et al., 2014]. Although stated without proof in [Van Dyk and Park, 2008], we provide a novel argument from first principles to prove an essential condition that has been overlooked in the literature for the invariance and ergodicity of the algorithm. The algorithms presented in this chapter are not specific to PSF reconstruction and can be generally applied to statistical inference problems where simulating samples from a complicated density are required.

Chapter 4 further develops these notions in order to carry out PSF estimation on a computer, and gives detailed descriptions of the algorithms and associated probability densities so that they can easily be adapted for a standard scientific or statistical software suite. We will

also deal with how to discretely represent each of the necessary components in the estimation problem. Finally we present results from an implementation of the methods using the MATLAB computing software library on synthetically derived and measured radiographic data. The measured data are from a high-energy X-ray imaging system at the U.S. Department of Energy's Nevada National Security Site. We will end with a discussion of conclusions and possible future work.

1.2 Modeling blur with a PSF

One major component of the spatial relationship of neighboring pixels of an image is due to blur from the imaging instrumentation. That is, under the assumption that arbitrary images are consistently measured by the modeled system, what contribution does this system have on how pixels are related, and how can we quantify this relationship? A widely used model for blurring [Hansen, 2010; Jain, 1989; Vogel, 2002; Epstein, 2008] expresses this relationship as a linear filter that maps an ideal image f to a blurred image b by integrating

$$b(x,y) = \iint_{\mathbb{R}^2} k(x,y;s,t) f(s,t)\, dsdt, \tag{1.1}$$

where $b(x,y)$ represents the intensity of the blurred image at (x,y); $f(s,t)$ represents the intensity of the ideal un-blurred image at (s,t); and k is the kernel of the filter, which characterizes the blurring process. For the purpose of modeling, we assume that each function is sufficiently regular so that each integral and change of variable can be interpreted in the Riemann sense.

Informally, the effect of blur can be viewed point-wise by observing the system response of a 'point-source' at $(\overline{x},\overline{y})$, then $b(x,y) = k(x,y;\overline{x},\overline{y})$ represents the "spread" of the signal f at the point source. The function k is referred to as the *point spread function* (PSF) of the system at $(\overline{x},\overline{y})$. When the effect of blurring does not depend on the location of this

point, that is, translating the ideal image f by $(\overline{x}, \overline{y})$ results in the blurred image b also being translated by $(\overline{x}, \overline{y})$, we say that the blur is *spatially invariant*. This means

$$\begin{aligned} b(x-\overline{x}, y-\overline{y}) &= \iint_{\mathbb{R}^2} k(x,y;s,t) f(s-\overline{x}, t-\overline{y}) ds dt \\ &= \iint_{\mathbb{R}^2} k(x,y;s'+\overline{x}, t'+\overline{y}) f(s',t') ds' dt'. \end{aligned} \tag{1.2}$$

On the other hand, applying the translation to (1.1) implies

$$b(x-\overline{x}, y-\overline{y}) = \iint_{\mathbb{R}^2} k(x-\overline{x}, y-\overline{y}, s, t) f(s,t) ds dt. \tag{1.3}$$

Since (1.2) and (1.3) hold for all f, we have for each $x, y, \overline{x}, \overline{y}, s$, and t that

$$k(x,y;s+\overline{x}, t+\overline{y}) = k(x-\overline{x}; y-\overline{y}, s, t) \tag{1.4}$$

and, in particular, when we fix $(s,t) = (0,0)$,

$$k(x,y;\overline{x},\overline{y}) = k(x-\overline{x}, y-\overline{y}; 0, 0). \tag{1.5}$$

Hence, the PSF in the blurring operation is independent of translation by $(\overline{x}, \overline{y})$. Let us denote $k(x,y;\overline{x},\overline{y}) = k(x-\overline{x}, y-\overline{y})$, then the linear filter in (1.1) reduces to

$$b(x,y) = \iint_{\mathbb{R}^2} k(x-s, y-t) f(s,t)\, ds dt. \tag{1.6}$$

Equation (1.6) is called the *convolution* of f by k. In fact, any translation invariant linear filter described as an operator between L^p spaces can be expressed through convolution with some generalized function k [Grafakos, 2014]. In any case, when blur is assumed to be spatially invariant, it results in solving the convolution equation (1.6). Mathematical methods that estimate f given b and k are referred to as *deconvolution* techniques.

Note that a change of variables by $s' = x - s$ and $t' = y - t$ results in a convolution of k by

f, which is to say that convolution, as an operation, is symmetric. That is

$$b(x,y) = \iint_{\mathbb{R}^2} k(s,t) f(x-s, y-t)\, dsdt. \tag{1.7}$$

This dual relationship between the PSF and the image will allow us to use the framework and many of the tools of deconvolution for the problem of PSF estimation. That is, we will use a known ideal f to estimate the PSF k.

Typically, deconvolution methods assume that the form of the PSF can be accurately described by modeling the imaging system [Jain, 1989; Hansen, 2010], but for X-ray radiography this is not realistic. If, instead, the imaging system is designed so that repeated images can be taken under consistent conditions, then by convolution symmetry in (1.7), the blurring of a known calibration image can be cast as deconvolving the PSF from the ideal f corresponding to the known image.

Recall that the PSF models the blurring response of a single point, so a direct estimate of k can be obtained by imaging a bright point-source, which approximates the impulse response to (1.7). In astronomical imaging, the point-source can be a bright distant star, or in a controlled setting where visible light is measured, a focused laser provides a good point-source estimate [Tomaney and Crotts, 1996]. However, in the spectral regime of X-rays, focusing the high-frequency light is notoriously difficult and usually is impractical in situations of interest, so a point-source estimate of the PSF is usually unavailable. Instead, the system response of a uniformly opaque calibration object with a simple geometry can be measured. When modeling the image of an object that is sufficiently thick so that X-rays are completely attenuated on its profile, we assume that the function for the ideal image is given by an indicator function on a set $E \subseteq \mathbb{R}^2$ determined by the object's profile. Calibration objects typically have simple geometry and reduce the complexity of solving the deconvolution problem in (1.7). For example, the object could be a circular aperture or two perpendicular edges aligned with the imaging plane [Doering et al., 1992; Watson, 1993]. As will be seen, the additional assump-

tion of radial symmetry on k is sufficient to estimate the PSF from the calibration image of a straight edge. Specifically, if the calibration object completely attenuates X-rays along a vertical edge at the fixed location at $s = 0$ in the imaging plane, then $E = \{(s,t) : s \geq 0\}$ and $f_E(s) = f(s,t) = 1$ if $s \geq 0$ and $f_E(s) = 0$ if $s < 0$; see Figure 1.1 for a schematic of the calibration object in the measurement system and Figure 1.2 for an example of recorded intensity data. The model for blur in (1.7) reduces to

$$b(x,y) = \iint_{\mathbb{R}^2} k(s,t) f_E(x-s) ds dt. \tag{1.8}$$

Note that b does not depend on y in (1.8), so denoting $b(x) = b(x,0)$, (1.8) reduces to

$$b(x) = \iint_{\mathbb{R}^2} k(s,t) f_E(x-s) ds dt. \tag{1.9}$$

In general, estimating k from b in (1.9) is underdetermined, since there are many distinct k that can result in the same output b. To see this, note that

$$\int_{-\infty}^{\infty} t e^{-t^2-s^2} dt = 0 \tag{1.10}$$

for all s since the integrand is odd in t, and given any solution k, $k(s,t) + te^{-t^2-s^2}$ also satisfies (1.9). Observe that these PSFs are not radially symmetric. It will be seen in the next section that the assumption of radial symmetry on k is sufficient for a unique solution to (1.9). Solving the integral equation in (1.9) is the primary focus of this work.

1.3 The Abel transform and a deterministic solution

This section is devoted to explicitly deriving a solution to (1.9) and showing that radial symmetry is sufficient to guarantee a unique solution. This deterministic method offers useful insight into the problem, but will prove to be inadequate for practically obtaining estimates

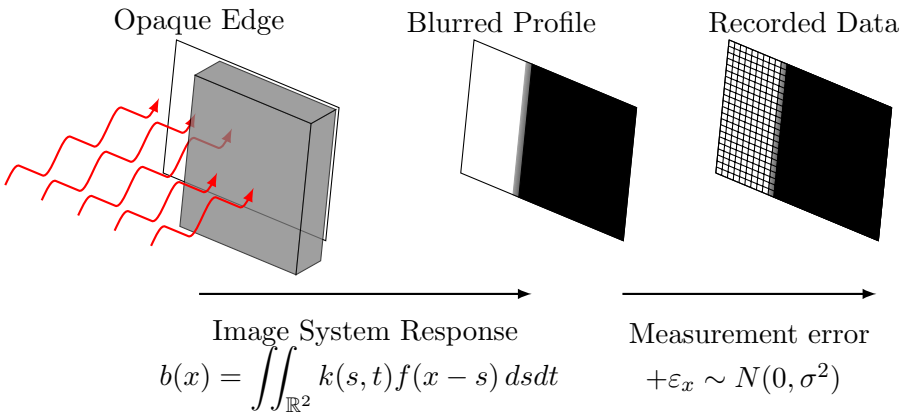

Figure 1.1: A schematic of the measurement model for an X-ray image of an edge. An opaque block aligned with the imaging plane blocks light on the half plane to produce a blurred edge.

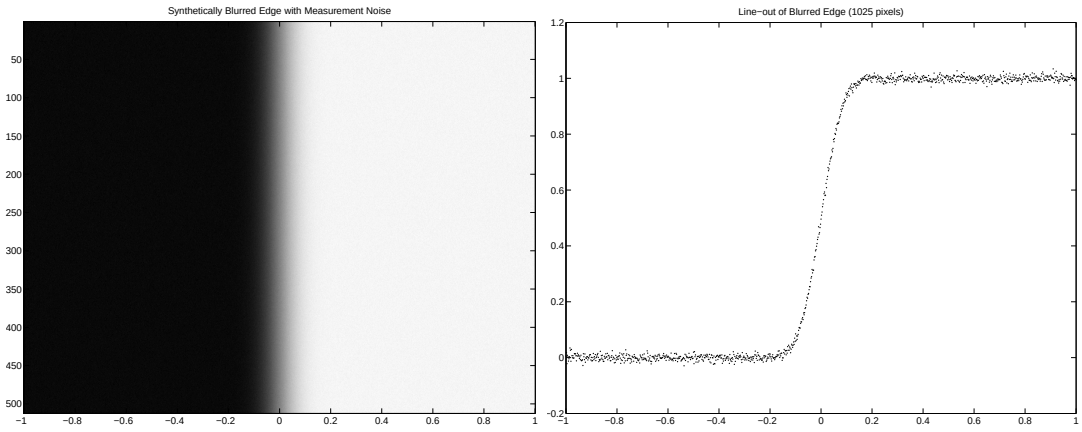

Figure 1.2: A synthetically blurred edge with simulated measurement error and a line-out (horizontal cross-section) from the data.

when the data are subject to measurement error.

Often, limitations due to physical laws put a lower bound on the measurement precision so that even an optimal design cannot ignore the effect of blur. Although arbitrary resolution may be impossible, it is often of interest to design a system that does not introduce error that is biased. A system that is not biased with respect to spatial orientation is said to exhibit *isotropic blur*, so that, in the convolution model, the PSF is radially symmetric. In fact, many parametrically modeled PSFs assume radially symmetry [Doering et al., 1992; Jain, 1989; Kundur and Hatzinakos, 1996; Watson, 1993].

When one assumes that the PSF of their system is a radially symmetric continuous function on $\mathbb{R}^2$, then it has a unique representation on $\mathbb{R}$; i.e., for $k : \mathbb{R}^2 \to \mathbb{R}$, there exists a function $p : [0, \infty) \to \mathbb{R}$ so that $k(s,t) = p\left(\sqrt{s^2+t^2}\right)$. The function p is referred to as the radial profile of k. Define

$$\ell(s) \stackrel{\text{def}}{=} \int_{-\infty}^{\infty} p\left(\sqrt{s^2+t^2}\right)\, dt, \tag{1.11}$$

which is integration along a line perpendicular to the edge E, a form commonly encountered in other imaging applications with radial geometry, such as tomographic imaging science. The transformation that takes p to ℓ is known as the *Abel transform*; for its study in imaging science, see [Bracewell, 1965; Epstein, 2008; Knill et al., 1993].

Viewing (1.9) as iterated integration first in t allows $f_E(s-x)$ to be factored out of the inner integral. Then substituting (1.11) for k into (1.9) and changing the bounds of integration in s according to $f_E(x-s)$, results in

$$\begin{aligned} b(x) &= \int_{-\infty}^{\infty} f_E(x-s)\left(\int_{-\infty}^{\infty} p\left(\sqrt{s^2+t^2}\right)\, dt\right) ds \\ &= \int_{-\infty}^{x} \left(\int_{-\infty}^{\infty} p\left(\sqrt{s^2+t^2}\right)\, dt\right) ds \\ &= \int_{-\infty}^{x} \ell(s) ds. \end{aligned} \tag{1.12}$$

The graph of b exhibits reflection symmetry about $(0, b(0))$. That is, for any $x > 0$, the point $(0, b(0))$ is the mid-point between the points $(x, b(x))$ and $(-x, b(-x))$. To see this explicitly, let $\widetilde{b}(x) \stackrel{\text{def}}{=} b(x) - b(0)$, then for $x > 0$, equation (1.12) implies

$$-\widetilde{b}(x) = -\left(b(x) - \int_{-\infty}^{0} \ell(s)ds\right) = -\int_{0}^{x} \ell(s)ds = \int_{0}^{-x} \ell(s')ds' = b(-x) - b(0) = \widetilde{b}(-x). \tag{1.13}$$

Hence $\widetilde{b}(x)$ is odd, so the graph of $b(x)$ has reflection symmetry about $(0, b(0))$. This means that data defined on either $x \in (-\infty, 0]$ or $x \in [0, \infty)$ are sufficient for estimating p, since the other half is determined by symmetry. This observation will be important in the next section.

The Abel transform has an explicit expression for its inverse [Epstein, 2008] given by

$$p(r) = -\frac{1}{\pi r}\frac{d}{dr}\left(\int_{r}^{\infty} \frac{\ell(s)sds}{(s^2 - r^2)^{1/2}}\right). \tag{1.14}$$

The following calculations verify (1.14).

Proposition 1.3.1. *Suppose that $p(r)$ is such $\lim_{r\to\infty} rp(r) = 0$, $\ell(s)$ in (1.11) is point-wise defined, and the integral in (1.14) is finite for each r. Then equation (1.14) holds.*

Proof. We can express the inner integral in (1.12) as

$$\ell(s) = 2\int_{|s|}^{\infty} \frac{p(t)t}{(t^2 - s^2)^{1/2}}\, dt \tag{1.15}$$

by symmetry of the integrand (it is even) and a change of variable by $r = s^2 + t^2$. Now, interchanging the order of integration in (1.14) results in

$$\begin{aligned}\left(\int_{r}^{\infty} \frac{\ell(s)sds}{(s^2 - r^2)^{1/2}}\right) &= \int_{r}^{\infty}\int_{s}^{\infty} \frac{2p(t)ts}{(s^2 - r^2)^{1/2}(t^2 - s^2)^{1/2}}\, dtds \\ &= \int_{r}^{\infty} p(t)t \int_{r}^{t} \frac{2s}{(s^2 - r^2)^{1/2}(t^2 - s^2)^{1/2}}\, dsdt. \end{aligned} \tag{1.16}$$

In the second step, we have interchanged variables and the integral's support can be expressed

$$\{(s,t) : r \leq s, s \leq t\} = \{(s,t) : r \leq s \leq t, r \leq t\}. \tag{1.17}$$

Another change of variables by $s^2 = \tau t^2 + (1-\tau)r^2$ (note $s \geq 0$) results in $2sds = (t^2 - r^2)d\tau$, so that the inner integral in (1.16) is

$$\int_r^t \frac{2s}{(s^2-r^2)^{1/2}(t^2-s^2)^{1/2}}\, ds = \int_0^1 \frac{1}{\tau^{1/2}(1-\tau)^{1/2}}\, d\tau = \pi, \tag{1.18}$$

where the last equality is given by an integral identity involving the gamma function. Collecting these results and applying the fundamental theorem of calculus with the assumption that $\lim_{r\to\infty} rp(r) = 0$ to (1.16) implies

$$\begin{aligned} -\frac{\pi}{r}\frac{d}{dr}\left(\int_r^\infty \frac{\ell(s)sds}{(s^2-r^2)^{1/2}}\right) &= -\frac{\pi}{r}\frac{d}{dr}\int_r^\infty p(t)t\pi dt \\ &= \frac{1}{r}\left(p(r)r - \lim_{r'\to\infty} p(r')r'\right) \\ &= p(r), \end{aligned} \tag{1.19}$$

which proves the identity in (1.14). □

With this result, p can be recovered from b in (1.12) as follows: given $b(x)$, the fundamental theorem of calculus gives $\ell(x) + \lim_{x'\to\infty} \ell(x')$ by differentiating $b(x)$. Since $\lim_{r\to\infty} tp(t) = 0$, symmetry and the change of variables in (1.15) implies $\lim_{x'\to-\infty} \ell(x') = 0$. Then, applying the inversion formula in (1.14) to $b'(x)$ gives the radial profile $p(r)$. That is,

$$p(r) = -\frac{1}{\pi r}\frac{d}{dr}\left(\int_r^\infty \left(\frac{d}{ds}b(s)\right)\frac{sds}{(s^2-r^2)^{1/2}}\right). \tag{1.20}$$

Hence, the assumption of radial symmetry gives a unique solution to (1.6), and hence, sufficiently constrains the problem to uniquely determine the PSF from an edge calibration object illustrated in Figure 1.2.

In theory, we have outlined a solution to the problem, but there is one more component to the model that has not been addressed – random effects due to measurement error – for which a direct application of the outlined method on measured data will fail spectacularly, due to the estimation problem being ill-posed, which we address in the next section.

1.4 PSF reconstruction as an ill-posed inverse problem

The solution given by (1.20) will not be sufficient when measurement errors are introduced. One clue that indicates this insufficiency is that it requires taking derivatives of measured data, which is known to be problematic [Hanke and Scherzer, 2001]. In this section, we will return to (1.9), and perform a different variable transformation to explicitly illustrate the instability, and in doing so, will derive a form that is more suitable for analysis and numerical discretization.

The measurements of the imaging system are generally not deterministic and are subject to measurement noise. Precisely modeling the stochastic effect of measurement error is system dependent and can be quite complicated. In X-ray radiography, uncertainty can enter into the system at the luminescing crystal response, in the counting process of the CCD array, or through the electrical transmission of the signal. In order to be broadly applicable, and appealing generally to various central-limit-theorem-like results in probability [Durrett, 2010], we model the stochastic measurement effect in aggregate as an additive, independent Gaussian noise process with zero mean and unknown variance. For now, this assumption can be viewed as a small perturbation from the model, but its form will be important for the inference techniques developed in subsequent chapters.

Estimating a quantity of interest, in our case k, from indirect and noisy measurements, b, with a model where an operator takes k to b (referred to as the *forward operator*) is called an inverse problem. The problem is called well-posed when the forward operator is invertible, and

the inverse is continuous. These famous conditions were laid out in the early 20th century by [Hadamard, 1902], but a number of important applications have arisen (among those computational imaging) where these conditions are violated; enough to the extent that the term 'inverse problems,' as it refers to the mathematical research area, is exclusively devoted to solving ill-posed problems. In particular, most cases of interest exhibit a model where the inverse of the forward operator is discontinuous.

Returning to (1.9), a variable transformation by $(s,t) = T(r,v) = (r\cos v, r\sin v)$, has $|dT(r,v)| = r$ and

$$\begin{aligned} b(x) &= \int_0^\infty p(r)\left(\int_{-\pi}^{\pi} f_E(x - r\cos v)dv\right) rdr \\ &= \int_0^\infty p(r)g(x,r)rdr, \end{aligned} \tag{1.21}$$

where

$$g(x,r) \stackrel{\text{def}}{=} \begin{cases} 0 & x < -r \\ 2(\pi - \operatorname{acos}(x/r)) & |x| \le r \\ 2\pi & x > r \end{cases}. \tag{1.22}$$

To see that g has this form, note that integrating $f_E(x - r\cos v)$ is the radian measure of the set $\{v \in (-\pi,\pi) : r\cos v \le x\}$; see Figure 1.3.

There are three key observations to make. From this viewpoint, the forward model is now a one-dimensional integral equation on the radial profile as opposed to the two-dimensional problem in (1.9). Second, note that $g(x,r)$ is continuous (although it has a discontinuity in its partial derivatives across $r = s$). Finally, recall that the graph of $b(x)$ exhibits reflection symmetry about $(0, b(0))$. So, $b(x)$ defined on either $(-\infty, 0]$ or $[0,\infty)$ completely determines p.

The discussion thus far has been somewhat informal, as we have assumed that all functions are continuous and have not defined a space for the PSF, the domain of the forward operator

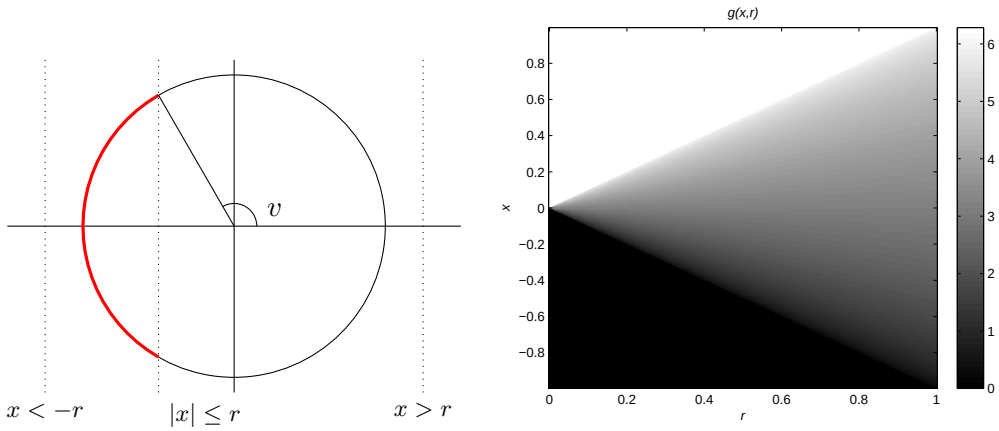

Figure 1.3: The PSF forward integral operator kernel $g(x, r)$ represented as the arc measure of v in $(-\pi, \pi)$ where $x \geq r \cos v$.

denoted $\mathscr{P}$, or the ambient space for the data, the codomain of the forward operator. Defining these spaces rigorously is the main focus of Chapter 2, and it will be shown that both of these spaces are separable Hilbert spaces and, in particular, that the data are a subspace of $L^2((-\infty, 0])$. For now, let $\mathcal{G} : \mathscr{P} \to L^2((-\infty, 0])$ be given and let the action of the operator be consistent with the integral equation in (1.21), i.e.,

$$[\mathcal{G}p](x) = \int_0^\infty p(r)g(x, r)rdr. \tag{1.23}$$

The operator $\mathcal{G}$ is a compact Hilbert-Schmidt operator since g is continuous. Moreover, $\mathcal{G}$ is an injective operator since we showed that $b = \mathcal{G}p$ has an explicit solution. The spectral theorem for such operators implies that $\mathcal{G}$ has a countable spectrum which has zero as a limit point, so its inverse $\mathcal{G}^{-1}$ is unbounded. See one of many texts on functional analysis such as [Bachman and Narici, 1966; Rudin, 1991] for the spectral theorem regarding Hilbert-Schmidt operators, and [Tikhonov, 1963; Vogel, 2002; Morozov and Stessin, 1993] for its role in inverse problems.

Solving ill-posed inverse problems requires prior assumptions about p that regularize the unbounded inverse. Recall that the original formulation of the problem is cast in terms of

deconvolution, and much of the literature of inverse problems is devoted to this subject. This work draws heavily from techniques for that purpose, but has many nuanced differences that make the analysis challenging, yet also novel and interesting. We will utilize a Bayesian approach to analyzing the inverse problem, since in addition to estimating k, these techniques provide uncertainties in the resulting estimate and can be quantified by analyzing the so-called posterior distribution. These methods have been the subject of much recent research (see the books [Calvetti and Somersalo, 2007; Kaipio and Somersalo, 2005; Stuart, 2010]), and the problem of PSF reconstruction fits neatly into that framework once the space $\mathscr{P}$ and the forward operator have been well-defined, which we address in the next chapter.

Chapter 2

Radial Symmetry for Sobolev Spaces

The domain and geometry of an image and a point spread function are inherently two-dimensional. While the assumption of radial symmetry reduces the representation dimension, the geometry of the problem is still two-dimensional and must be appropriately translated to the radial representations. The question of how regularity assumptions like square summability of derivatives translates to a radial representation does not have an obvious resolution if we allow generalised functions and generalized derivatives. The goal of this chapter is to rigorously address these issues by developing the necessary mathematical tools to encapsulate prior notions of regularity and radial symmetry within the structure of separable Hilbert spaces and concretely establish how that affects a radial representation. The development is done within the theory of distributions, developed initially for the analysis of linear partial differential equations (PDEs). This framework provides sufficient generality to encapsulate radial symmetry and sufficient regularization via generalization of differential operators and function composition, all within the rich structure of a separable Hilbert space. Hence, tools for analyzing inverse problems that require radial symmetry are readily applicable to the PSF

estimation problem.

2.1 Distribution spaces

In this section we establish the preliminary definitions and main results from distribution theory. There are several treatments of the subject in varying levels of generality, and this work draws primarily from [Richtmyer, 1978; Hörmander, 1983; Rudin, 1991; Griffel, 2002; Strichartz, 2003].

2.1.1 The space of test functions and distributions

Let $\mathscr{D}(\Omega)$ denote the space of compactly supported smooth functions defined on an open set $\Omega \subseteq \mathbb{R}^N$. Endow $\mathscr{D}(\Omega)$ with the topology such that convergence of the sequence $(\phi_n) \subset \mathscr{D}(\Omega)$ happens whenever there exists a compact set K such that

$$\bigcup_{n=1}^{\infty} \operatorname{supp} \phi_n \subseteq K \quad \text{and} \quad \sup_{m \geq n} |\partial^\alpha (\phi_n - \phi_m)| \to 0 \text{ as } n \to \infty \tag{2.1}$$

for any multi-index such that $|\alpha| \leq k$. That is, α is a k-tuple of non-negative integers $(\alpha_1, \ldots, \alpha_k)$, such that $\sum \alpha_i \leq k$ and $\partial^\alpha = \prod \left(\frac{\partial}{\partial x_i} \right)^{\alpha_i}$. In distribution theory, these are called *test functions* on Ω. The space of continuous linear functionals, denoted $\mathscr{D}^*(\Omega)$, are the *distributions* on Ω. We adopt the notation $\langle f, \phi \rangle$ for the action of a linear functional f on $\phi \in \mathscr{D}(\Omega)$ and freely use the natural inclusion of functions $g \to \widetilde{g} \in \mathscr{D}^*(\Omega)$ by $\langle \widetilde{g}, \phi \rangle = \int g \, \phi \, dx$ when the integration exists and omit the tilde notation distinguishing g and $\widetilde{g}$ as the representation should be clear from context.

Consider the following two general topological results regarding distributions. See [Hörmander, 1983, Chapter 2] for the proofs of each.

Theorem 2.1.1. *[Hörmander, 1983] Suppose (f_n) is a sequence in $\mathscr{D}^*(\Omega)$ such that $\lim\langle f_n, \phi\rangle$ exists for all $\phi \in \mathscr{D}(\Omega)$, then there exists a unique $f \in \mathscr{D}^*(\Omega)$ such that*

$$\langle f, \phi\rangle = \lim_{n\to\infty} \langle f_n, \phi\rangle. \tag{2.2}$$

The existence of the linear functional f can be readily established by using the completeness of the associated field (either $\mathbb{R}$ or $\mathbb{C}$), and the main difficulty of establishing the result is showing that the resulting linear functional is continuous with respect to the topology of $\mathscr{D}(\Omega)$. It essentially follows from a general statement of the Banach-Steinhaus theorem for Fréchet spaces, and a complete development from first principles can be found in [Rudin, 1991].

The next result, sometimes referred to as localization, establishes a dense embedding of the $\mathscr{D}(\Omega)$ into $\mathscr{D}^*(\Omega)$.

Theorem 2.1.2. *[Hörmander, 1983] Given $f \in \mathscr{D}^*(\Omega)$, there exists a sequence $(\phi_n) \subset \mathscr{D}(\Omega)$ such that*

$$\langle f, \psi\rangle = \lim_{n\to\infty} \langle \phi_n, \psi\rangle. \tag{2.3}$$

These results allow for operators defined on $\mathscr{D}(\Omega)$ to be extended in a continuous way to $\mathscr{D}^*(\Omega)$ so long as one can define an adjoint operation with respect to the sesquilinear form $\langle\cdot,\cdot\rangle$. Moreover, if the operation is well-defined on test functions, then uniqueness of the extension follows from Theorem 2.1.2. The classical example of this process is extending the differential operator $\frac{\partial}{\partial x_i} : \mathscr{D}^*(\Omega) \to \mathscr{D}^*(\Omega)$. First, for test functions observe that integrating by parts and using the compactness of the support of ψ yields

$$\left\langle \frac{\partial}{\partial x_i}\phi, \psi \right\rangle = \int_\Omega \frac{\partial}{\partial x_i}\phi(x)\psi(x)dx = -\int_\Omega \phi(x)\frac{\partial}{\partial x_i}\psi(x)dx = -\left\langle \phi, \frac{\partial}{\partial x_i}\psi \right\rangle. \tag{2.4}$$

This motivates defining $\frac{\partial}{\partial x_i} f \in \mathscr{D}^*(\Omega)$ by

$$\left\langle \frac{\partial}{\partial x_i} f, \phi \right\rangle \stackrel{\text{def}}{=} - \left\langle f, \frac{\partial}{\partial x_i} \psi \right\rangle \tag{2.5}$$

from which we can extend the definition of ∂^α

$$\langle \partial^\alpha f, \phi \rangle \stackrel{\text{def}}{=} (-1)^{|\alpha|} \langle f, \partial^\alpha \psi \rangle . \tag{2.6}$$

Note that the resulting functional is continuous, by the smoothness of $\psi \in \mathscr{D}(\Omega)$. Since the operator $\partial^\alpha : \mathscr{D}^*(\Omega) \to \mathscr{D}^*(\Omega)$ is expressed as an adjoint on test functions with respect to evaluation, its continuity follows directly from the weak*-topology induced from $\mathscr{D}(\Omega)$, i.e., suppose $f_n \to f$ in $\mathscr{D}^*(\Omega)$, then

$$\lim_{n\to\infty} \langle \partial^\alpha f_n, \phi \rangle = \lim_{n\to\infty} (-1)^{|\alpha|} \langle f_n, \partial^\alpha \phi \rangle = (-1)^{|\alpha|} \langle f, \partial^\alpha \phi \rangle = \langle \partial^\alpha f, \phi \rangle. \tag{2.7}$$

The idea of expressing an operation adjointly on test functions, serves as the model for extending radial symmetry to distributions. That is, the radial change of variables introduced in Chapter 1 can be 'moved' to test functions in an analgous way to express function composition on distributions. Before this argument is undertaken, we first show how L^2 functions can be thought of as distributions, and how regularity can be imposed through generalized derivatives with Sobolev spaces.

2.1.2 L^2 as a subspace of distributions

Hilbert spaces provide the rich structure of a complete inner product and guarantee completeness in the sense that Cauchy sequences converge to elements in the space. Moreover, formulating inverse problems on a Hilbert space is, in many cases, a prerequisite for employing variational or infinite dimensional probabilistic Bayesian methods. This development will

show how the space of square integrable functions L^2 and their generalized derivatives can be defined as distributions, which is necessary for defining the Hilbert space for radial profiles and also provides valuable insights in how they will be constructed. The development follows [Richtmyer, 1978; Griffel, 2002], for which we need only the notion the L^2 inner product, as opposed to the more common Fourier based approaches which can be found in [Rudin, 1991; Hörmander, 1983; Strichartz, 2000]. Our development also provides several details that are omitted in [Richtmyer, 1978].

We define the L^2 inner-product for test functions as the sesquilinear form $(\cdot,\cdot)_{L^2(\Omega)} : \mathscr{D}(\Omega) \times \mathscr{D}(\Omega) \to \mathbb{C}$ by the Riemann integral

$$(\phi, \psi)_{L^2(\Omega)} \stackrel{\text{def}}{=} \int_\Omega \phi(x)\overline{\psi(x)}\, dx, \tag{2.8}$$

with the induced norm

$$\|\phi\|^2_{L^2(\Omega)} \stackrel{\text{def}}{=} (\phi, \phi)_{L^2(\Omega)}. \tag{2.9}$$

The linearity of the first component (and anti-linearity of the second) in (2.8) are given by the linearity of integration, and positivity follows from the positivity of $\phi(x)\overline{\phi(x)} = |\phi(x)|^2$. For definiteness, note that if $\phi = 0$ then $\|\phi\|_{L^2(\Omega)} = 0$, and only if $\phi = 0$, otherwise, continuity of ϕ implies that there exists a neighborhood where $|\phi(x)| > 0$, which gives $\|\phi\|_{L^2(\Omega)} > 0$. Since (2.8) defines an inner-product, the triangle inequality of the norm follows from the Cauchy-Schwarz-Bunyakovsky inner product inequality $|(\phi, \psi)| \leq \|\phi\|\|\psi\|$; i.e.,

$$\begin{aligned}
\|\phi + \psi\|^2_{L^2(\Omega)} &= \|\phi\|^2_{L^2(\Omega)} + 2\mathrm{Re}(\phi, \psi)_{L^2(\Omega)} + \|\psi\|^2_{L^2(\Omega)} \\
&\leq \|\phi\|^2_{L^2(\Omega)} + 2|(\phi, \psi)_{L^2(\Omega)}| + \|\psi\|^2_{L^2(\Omega)} \\
&\leq \|\phi\|^2_{L^2(\Omega)} + 2\|\phi\|_{L^2(\Omega)}\|\psi\|_{L^2(\Omega)} + \|\psi\|^2_{L^2(\Omega)} \\
&= \left(\|\phi\|^2_{L^2(\Omega)} + \|\psi\|^2_{L^2(\Omega)}\right)^2.
\end{aligned} \tag{2.10}$$

A sequence $(\phi_n) \subset \mathscr{D}(\Omega)$ is Cauchy with respect to $L^2(\Omega)$ if

$$\lim_{n\to\infty} \sup_{k\geq n} \|\phi_n - \phi_k\|_{L^2(\Omega)} = 0. \tag{2.11}$$

Observe that $(\phi, \psi)_{L^2(\Omega)} = \langle \phi, \overline{\psi} \rangle$ when ϕ is viewed as an element of $\mathscr{D}^*(\Omega)$. Hence, if (ϕ_n) is Cauchy with respect to $L^2(\Omega)$, then the sequence of complex numbers $\big\{\langle \phi_n, \psi \rangle\big\}$ is Cauchy, i.e., using Cauchy-Schwarz-Bunyakovsky

$$|\langle \phi_n, \psi \rangle - \langle \phi_k, \psi \rangle| = |(\phi_n - \phi_k, \overline{\psi})_{L^2(\Omega)}| \leq \|\phi_n - \phi_k\|_{L^2(\Omega)} \|\psi\|_{L^2(\Omega)}. \tag{2.12}$$

Hence, $\lim \langle \phi_n, \psi \rangle$ exists for all $\psi \in \mathscr{D}(\Omega)$ (by completeness of $\mathbb{C}$), and Theorem 2.1.1 provides uniquely an $f \in \mathscr{D}^*(\Omega)$ such that

$$\lim_{n\to\infty} \langle \phi_n, \psi \rangle = \langle f, \psi \rangle. \tag{2.13}$$

All such f are elements of the space $L^2(\Omega)$.

Sequences (ϕ_n) and (ϕ'_n) are *equivalent* $L^2(\Omega)$ Cauchy sequences if

$$\|\phi_n - \phi'_n\|_{L^2(\Omega)} \to 0. \tag{2.14}$$

These distributions are well defined in the following sense:

Proposition 2.1.3. *Sequences (ϕ_n) and (ϕ'_n) determine the same distribution if and only if they are equivalent.*

The proof can be found in [Richtmyer, 1978], and follows from several straight-forward applications of the Cauchy-Schwarz-Bunyakovsky inequality.

It can be readily shown that (2.14) defines an equivalence relation on Cauchy sequences for which vector addition and scalar multiplication are well-defined, hence, the equivalence

classes of Cauchy sequences correspond uniquely to a linear subspace of $\mathscr{D}^*(\Omega)$.

We can now extend the inner product to elements in $L^2(\Omega)$ in the following way,

$$(f,g)_{L^2(\Omega)} \stackrel{\text{def}}{=} \lim_{n\to\infty} (\phi_n, \psi_n)_{L^2(\Omega)} \tag{2.15}$$

where (ϕ_n) and (ψ_n) are Cauchy sequences corresponding to f and g respectively. The following proposition is not proved in [Richtmyer, 1978], so we provide a short proof.

Proposition 2.1.4. *The limit in* (2.15) *exists and is well-defined for equivalent Cauchy sequences. Moreover,* $\lim_{n\to\infty}\langle f, \overline{\psi_n}\rangle_{L^2(\Omega)} = (f,g)_{L^2(\Omega)}$.

Proof. Since (ϕ_n) and (ψ_n) are Cauchy, the inequality $|\|\phi_n\|_{L^2(\Omega)} - \|\phi_k\|_{L^2(\Omega)}| \leq \|\phi_n - \phi_k\|_{L^2(\Omega)}$ implies that $\|\phi_n\|_{L^2(\Omega)}$ and $\|\psi_n\|$ are both Cauchy sequences of positive numbers, and thus have finite limits.

Now observe,

$$\begin{aligned}
|(\phi_n,\psi_n)_{L^2(\Omega)} - (\phi_k,\psi_k)_{L^2(\Omega)}| &= \big|(\phi_n, \psi_n - \psi_k)_{L^2(\Omega)} - (\phi_n - \phi_k, \psi_k)_{L^2(\Omega)}\big| \\
&\leq \|\phi_n\|_{L^2(\Omega)}\|\psi_n - \psi_k\|_{L^2(\Omega)} + \|\phi_n - \phi_k\|_{L^2(\Omega)}\|\psi_k\|_{L^2(\Omega)} \\
&\leq \|\phi_n\|_{L^2(\Omega)}\|\psi_n - \psi_k\|_{L^2(\Omega)} \\
&\qquad + \|\phi_n - \phi_k\|_{L^2(\Omega)}\big(\|\psi_k - \psi_n\|_{L^2(\Omega)} + \|\psi_n\|_{L^2(\Omega)}\big).
\end{aligned} \tag{2.16}$$

Since the limit as $n \to \infty$ of the $\sup_{k\geq n}$ of the right hand side of the inequality results in 0 (the sequences $(\|\phi_n\|)$ and $(\|\psi_n\|)$ have finite limits) the convergence of the left hand side follows. Thus, (ϕ_n, ψ_n) is a Cauchy sequence in $\mathbb{C}$, and hence, has a finite limit.

Let $\varepsilon > 0$ be given. For all n, choose m sufficiently large so that

$$|\langle f, \overline{\psi_n}\rangle - (f,g)_{L^2(\Omega)}| \leq |\langle f - \phi_m, \overline{\psi_n}\rangle| + |\langle \phi_m - \phi_n, \overline{\psi_n}\rangle| + |\langle \phi_n, \overline{\psi_n}\rangle - (f,g)_{L^2(\Omega)}|$$

$$\leq |\langle f - \phi_m, \overline{\psi_n}\rangle| + \|\phi_m - \phi_n\|\|\psi_n\| + |\langle \phi_n, \overline{\psi_n}\rangle - (f,g)_{L^2(\Omega)}|$$
$$< \varepsilon + |\langle \phi_n, \overline{\psi_n}\rangle - (f,g)_{L^2(\Omega)}|. \tag{2.17}$$

Taking limits on both sides of the inequality

$$\lim_{n\to\infty} \langle f, \overline{\psi_n}\rangle = (f,g)_{L^2(\Omega)}, \tag{2.18}$$

since $\varepsilon > 0$ is arbitrary.

Finally, to show that the inner product is well-defined, suppose (ϕ_n') and (ψ_n') are equivalent Cauchy sequences to (ϕ_n) and (ψ_n) respectively. Then

$$\begin{aligned} |\lim_{n\to\infty} (\phi_n, \psi_n)_{L^2(\Omega)} - (\phi_n', \psi_n')_{L^2(\Omega)}| &= |\lim_{n\to\infty} \langle f, \overline{\psi_n}\rangle - \lim_{n\to\infty} \langle f', \overline{\psi_n}'\rangle| && \text{by (2.18)} \\ &= |\lim_{n\to\infty} \langle f, \overline{\psi_n} - \overline{\psi_n}'\rangle| && \text{by Proposition 2.1.3} \\ &= 0. \end{aligned}$$

Thus, the inner product is well-defined for equivalent Cauchy sequences. □

Showing that (2.15) is an inner product on $L^2(\Omega)$ is straight-forward, and the resulting inner product space is a Hilbert space, as stated in the following theorem.

Theorem 2.1.5. *The space $L^2(\Omega)$ is complete with respect to the inner product defined in* (2.15)*, hence, is a Hilbert space.*

See [Richtmyer, 1978] for the proof, which follows a standard diagonalization argument. We remark that there is a correspondence with the standard notion of $L^2(\Omega)$ with respect to Lebesgue measure and this development. The basis of the correspondence comes from the result that a simple function can be arbitrarily approximated with a sequence of test functions in the standard $L^2(\Omega)$ sense with Lebesgue measure. See [Hörmander, 1983] for a rigorous

development of this correspondence.

2.1.3 The Sobolev space $\mathscr{H}^n(\Omega)$

Sobolev spaces provide a framework for imposing regularity on distributions in terms of their derivatives. In this work, we will be concerned only with Sobolev subspaces of L^2, which are sometimes denoted $W^{k,2}$. In this subsection, we briefly overview the definition of these spaces, and state a version of the Sobolev embedding theorem sufficient for characterizing PSFs of interest in this work.

A Sobolev space of order n over an open set $\Omega \subseteq \mathbb{R}^k$ is $\mathscr{H}^n(\Omega) = \{f \in L^2(\Omega) : \partial^\alpha f \in L^2(\Omega) \text{ whenever } |\alpha| \leq n\}$. Each of these form a sequence of linear subspaces $\mathscr{H}^n(\Omega) \subset \mathscr{H}^{n-1}(\Omega) \subset \cdots \subset \mathscr{H}^1(\Omega) \subset L^2(\Omega)$, however, the inclusion is strict and they are not closed with respect to the L^2 norm [Richtmyer, 1978].

They are, however, complete with respect to a new inner product and norm given by

$$(f, g)_{\Omega,n} = \sum_{0 \leq |\alpha| \leq n} (\partial^\alpha f, \partial^\alpha g)_{L^2(\Omega)} . \tag{2.19}$$

We denote the corresponding norm $\|f\|_{\Omega,n}$. The proof is by induction on the order of $|\alpha|$, where the base case is the completeness of $L^2(\Omega)$. To see the inductive step, note

$$(f, g)_{\Omega,1} = (f, g)_{L^2(\Omega)} + \sum_{i=1}^{k} \left(\frac{\partial}{\partial x_i} f, \frac{\partial}{\partial x_i} g \right)_{L^2(\Omega)} . \tag{2.20}$$

Let $(f_n) \subseteq \mathscr{H}^1$ be Cauchy with respect to the corresponding norm. Since $\|f_n\|_{L^2(\Omega)} \leq \|f_n\|_{\Omega,1}$, there exists a function $f \in L^2(\Omega)$ such that (f_n) converges to f in $L^2(\Omega)$ by completeness there. Moreover, each $\frac{\partial}{\partial x_i} f_n$ is similarly also a Cauchy sequence in $L^2(\Omega)$, so let $f^{(i)}$ be the corresponding limit in $L^2(\Omega)$. Since L^2 convergence is stronger than distributional convergence,

we also have that the distributional derivatives of f converge to $f^{(i)}$; explicitly,

$$\left\langle \frac{\partial}{\partial x_i} f, \psi \right\rangle = - \left\langle f, \frac{\partial}{\partial x_i} \psi \right\rangle = \lim_{n\to\infty} - \left\langle f_n, \frac{\partial}{\partial x_i} \psi \right\rangle = \lim_{n\to\infty} \left\langle \frac{\partial}{\partial x_i} f_n, \psi \right\rangle = \langle f^{(i)}, \psi \rangle. \quad (2.21)$$

Hence, each distributional derivative $\frac{\partial}{\partial x_i} f \in L^2(\Omega)$, and thus $f \in \mathscr{H}^1(\Omega)$. Finally,

$$\|f_n - f\|_{\Omega,1} = \|f_n - f\|_{L^2(\Omega)} + \sum_{i=1}^{n} \left\| \frac{\partial}{\partial x_i} f_n - f^{(i)} \right\|_{L^2(\Omega)} \to 0 \quad (2.22)$$

since each term does by their construction, hence $f_n \to f$ in $\mathscr{H}^1$.

In this argument, note that each Sobolev space inherits completeness from the lower order space, which will be similar to how the space of radial representations with an appropriate inner product inherit completeness.

We will occasionally use the convenient abbreviations from differential calculus

$$(\nabla^n f, \nabla^n g)_{L^2(\Omega)} \stackrel{\text{def}}{=} \sum_{|\alpha|=n} (\partial^\alpha f, \partial^\alpha g)_{L^2(\Omega)} . \quad (2.23)$$

When $n = 1$, the inner product and norm reduce to the familiar gradient forms,

$$(\nabla f, \nabla g)_{L^2(\Omega)} = \sum_{i=1}^{k} \left(\frac{\partial}{\partial x_i} f, \frac{\partial}{\partial x_i} g \right)_{L^2(\Omega)} \quad (2.24)$$

and

$$\|\nabla f\|^2_{L^2(\Omega)} = \sum_{i=1}^{k} \left\| \frac{\partial}{\partial x_i} f \right\|^2_{L^2(\Omega)} . \quad (2.25)$$

2.2 Radial symmetry

Symmetry is established by casting it as precomposition with a many-to-one smooth map T that is constant on circles of a fixed radius. If f is a function on $\mathbb{R}^2$ and there exists a

function p so that $f(x,y) = p(\sqrt{x^2+y^2})$, then observe that f has the common notion of radial symmetry with a radial profile p. This notion is easily adapted to distributions by developing a notion of composition with T and using a so-called linear pullback operator to $T^\sharp$. That is, $T^\sharp$ maps p to f by precomposition with $T(x,y)$ on sequences of test functions converging to p. A pullback is the analogous notion of composition for distributions, and the idea of the construction is to extend the range of T to a bijection similar to the polar coordinates map and use it to perform an integral change of variables to 'move' the composition to a test functions. For topological reasons, the bijection will only be with a dense subset of $\mathbb{R}^2$.

2.2.1 The pull-back operator

In this subsection, we will explicitly construct the pullback operator $T^\sharp$ for a slightly more general smooth map $T(x,y)$. It will turn out that by explicitly constructing the pullback operator for radial symmetry, we will obtain the additional property of injectivity.

Let $\Omega_1 \stackrel{\text{def}}{=} (0,\infty) \subset \mathbb{R}$ and $\Omega_2 \stackrel{\text{def}}{=} \mathbb{R}^2 \setminus \{x = 0 \text{ or } y = 0\}$. For $h : \Omega_1 \to \Omega_1$ with $h(t) \stackrel{\text{def}}{=} t^a$ and $0 < a < 1$, let $T : \Omega_2 \to \Omega_1$ by $T(x,y) = h(x^2+y^2)$. In order to establish the operator on distributions, we first observe its adjoint action on test functions.

Lemma 2.2.1. *There exists a map $T_\sharp : \mathscr{D}(\Omega_2) \to \mathscr{D}(\Omega_1)$ so that for any $\rho \in \mathscr{D}(\Omega_1)$*

$$\langle \rho \circ T, \phi \rangle_{\Omega_2} = \langle \rho, T_\sharp \phi \rangle_{\Omega_1}. \tag{2.26}$$

Proof. Let $Q_{ij} = \{(x,y) : (-1)^i x > 0, (-1)^j y > 0\}$ for $i,j \in \{0,1\}$ so that $\bigcup Q_{ij} = \Omega_2$. Define $T_{ij} : Q_{ij} \to R \subset \mathbb{R}^2$ by

$$T_{ij}(x,y) = \Big(T(x,y), (-1)^j y\Big). \tag{2.27}$$

Observe that each T_{ij} is a diffeomorphism onto $R = \left\{(r,t) : 0 < t < \sqrt{h^{-1}(r)}\right\} = \left\{(r,t) : 0 < t < r^{\frac{1}{2a}}\right\}$

with inverse

$$T_{ij}^{-1}(r,t) = \left((-1)^i\sqrt{h^{-1}(r)-t^2}, (-1)^j t\right) = \left((-1)^i\sqrt{r^{\frac{1}{a}}-t^2}, (-1)^j t\right), \tag{2.28}$$

and

$$\left|dT_{ij}^{-1}(r,t)\right| = \frac{1}{2}\frac{\partial}{\partial r}[h^{-1}(r)]\left(h^{-1}(r)-t^2\right)^{-1/2} = \frac{1}{2a}r^{\frac{1}{a}-1}\left(r^{\frac{1}{a}}-t^2\right)^{-1/2}, \tag{2.29}$$

which is positive and smooth for all $(r,t)\in\Omega_2$. Furthermore, note that

$$T\circ T_{ij}^{-1}(r,t) = r. \tag{2.30}$$

Now, given $\rho\in\mathscr{D}(\Omega_1)$, a change of variables results in

$$\begin{aligned}\langle\rho\circ T,\phi\rangle_{\Omega_2} &= \sum_{ij}\iint_{Q_{ij}}\rho\circ T(x,y)\cdot\phi(x,y)dxdy\\ &= \sum_{ij}\iint_R \rho(r)\cdot\phi\circ T_{ij}^{-1}(r,t)\left|dT_{ij}\right|drdt\\ &= \int_0^\infty \rho(r)\left(\int_0^{\sqrt{h^{-1}(r)}}\sum_{ij}\phi\circ T_{ij}^{-1}(r,t)\left|dT_{ij}\right|dt\right)dr.\end{aligned} \tag{2.31}$$

Let

$$[T_\sharp\phi](r) = \int_0^{\sqrt{h^{-1}(r)}}\sum_{ij}\phi\circ T_{ij}^{-1}(r,t)\left|dT_{ij}\right|dt \tag{2.32}$$

$$= \frac{r^{\frac{1}{a}-1}}{2a}\sum_{ij}\int_0^{r^{\frac{1}{2a}}}\phi\left((-1)^i\sqrt{r^{\frac{1}{a}}-t^2}, (-1)^j t\right)\left(r^{\frac{1}{a}}-t^2\right)^{-1/2}dt, \tag{2.33}$$

and we must show that $T_\sharp\phi\in\mathscr{D}(\Omega_1)$. Note that $\operatorname{supp}\left(\phi\circ T_{ij}^{-1}\right) = T_{ij}(\operatorname{supp}\phi)$ is compact in R as it is the continuous image of a compact set, and since T_{ij} is a diffeomorphism, $\phi\circ T_{ij}^{-1}\in\mathscr{D}(\Omega_2)$. Since $\phi\circ T_{ij}^{-1}(r,t)$ is smooth, a result in [Strichartz, 2000, pg. 433] guarantees that integrating out t, i.e., $\int\phi\circ T_{ij}^{-1}(r,t)\,dt$ results in a smooth function in r. The

support of this function is the projection of the support of $\phi \circ T_{ij}$ onto the second coordinate, and hence, is compact. Summing over i, j results in a compactly supported smooth function.

□

Using Lemma 2.2.1, we can now define the *pullback* by T on $\mathscr{D}^*(\Omega_1)$ by $T^\sharp : \mathscr{D}^*(\Omega_1) \to \mathscr{D}^*(\Omega_2)$ by

$$\langle T^\sharp p, \phi \rangle_{\Omega_2} \stackrel{\text{def}}{=} \langle p, T_\sharp \phi \rangle_{\Omega_1}. \tag{2.34}$$

As previously remarked, this operator generalizes smooth composition for distributions, and will be the basis for how we define radial symmetry. To see that $T^\sharp p \in \mathscr{D}^*(\Omega_2)$ (i.e. acts continuously on $\mathscr{D}(\Omega_2)$ as a linear functional), let $(\phi_n) \to 0$ in $\mathscr{D}(\Omega_2)$, so fixing $\alpha = 0$ in (2.1), we have $\sup_{\Omega_2} |\phi_n| \to 0$ as $n \to \infty$. Then, by (2.33), $\sup_{\Omega_1} |T_\sharp \phi_n| \to 0$ as $n \to \infty$, and thus $\langle p, T_\sharp \phi_n \rangle \to 0$ by the continuity of p.

The linearity and continuity of $T^\sharp$ follow directly from this definition. That is

$$\begin{aligned} \langle T^\sharp p_1 + \alpha T^\sharp p_2, \phi \rangle_{\Omega_2} &= \langle T^\sharp p_1, \phi \rangle_{\Omega_2} + \alpha \langle T^\sharp p_2, \phi \rangle_{\Omega_2} \\ &= \langle p_1, T_\sharp \phi \rangle_{\Omega_1} + \alpha \langle p_2, T_\sharp \phi \rangle_{\Omega_1} \\ &= \langle T^\sharp (p_1 + \alpha p_2), \phi \rangle_{\Omega_2} \end{aligned} \tag{2.35}$$

and if $\langle p_n, \psi \rangle \to 0$ for all $\psi \in \mathscr{D}(\Omega_1)$, then

$$\langle T^\sharp p_n, \phi \rangle_{\Omega_2} = \langle p_n, T_\sharp \phi \rangle_{\Omega_1} \to 0. \tag{2.36}$$

Loosely speaking, the pullback by T represents a change of variables from (x, y) to (r, v) by expanding the domain of T to an invertible $T_{ij}(x, y)$ with the choice of T_{ij} arbitrary. Observe that for a fixed $\widetilde{r} = r^{1/(2a)}$, the image of $T_{ij}^{-1}(r, v)$ is a quarter circle of radius $\widetilde{r}$ in the quadrant determined by i, j. Another valid choice of T_{ij}, which is similar to a polar-

coordinates transformation, would be $(T(x,y), \mathrm{Arg}(x,y))$, but the calculations in this case are somewhat more tedious. We will show that $T^\sharp$ is unique with respect to the choice of T_{ij}, and it will allow us to freely choose any other change of variables such that $T \circ T_{ij}(r,v) = r$ with the analysis on T remains valid. We remark that the existence and continuity of a more generally defined pullback operator can be carried for any smooth T and is outlined in [Hörmander, 1983]. However, in this case, because of the specific form of T under consideration, the induced pullback $T^\sharp$ is injective. This will be a consequence of the next lemma.

Lemma 2.2.2. *For all $p \in \mathscr{D}^*(\Omega_1)$ and $\omega \in \mathscr{D}(\Omega_1)$*

$$\langle T^\sharp p, \omega \circ T \rangle_{\Omega_2} = \langle p, S(\omega) \rangle_{\Omega_1}, \tag{2.37}$$

where $S(\omega)$ is the shift operator defined by $S(\omega) = \omega \cdot \pi h^{-1'}$, the derivative of the inverse of h in Theorem 2.2.4. When $h(r) = r^{1/2}$, then $h^{-1'}(r) = 2\pi r$.

Proof. First, note that both $\omega \circ T$ and $h^{-1'} \cdot \omega$ are elements of $\mathscr{D}(\Omega_1)$. From (2.29), we calculate the trigonometric integral

$$\begin{aligned} \int_0^{\sqrt{h^{-1}(r)}} |dT_{ij}(\,r,t)| dt &= \frac{h^{-1'}(r)}{2} \int_0^{\sqrt{h^{-1}(r)}} \left(h^{-1}(r) - t^2\right)^{-1/2} \\ &= \frac{\pi}{4} h^{-1'}(r). \end{aligned} \tag{2.38}$$

By invoking Theorem 2.1.2, let (ρ_n) be a sequence in $\mathscr{D}(\Omega_1)$ converging to p in $\mathscr{D}^*(\Omega_1)$, then substituting ρ_n for p and $\psi \circ T$ for ϕ in (2.31), we have

$$\begin{aligned} \langle T^\sharp \rho_n, \psi \circ T \rangle_{\Omega_2} &= 4 \int_0^\infty \rho_n(r)\psi(r) \left(\int_0^{\sqrt{h^{-1}(r)}} |dT_{ij}|\, dt \right) dr \\ &= \pi \int_0^\infty \rho_n(r)\psi(r)\, h^{-1'}(r) dr \\ &= \langle \rho_n, S(\psi) \rangle_{\Omega_1}. \end{aligned} \tag{2.39}$$

By continuity of $T^\sharp$, the desired equality is established. □

Recall that the action of $T^\sharp$ on test functions was the standard notion of composition with T, so the operator $T_\sharp T^\sharp : \mathscr{D}(\Omega_1) \to \mathscr{D}(\Omega_1)$ acts on ρ as a distribution by

$$\left\langle T_\sharp T^\sharp \rho, \omega \right\rangle_{\Omega_1} = \left\langle T^\sharp \rho, T^\sharp \omega \right\rangle_{\Omega_2} = \pi \int_0^\infty \rho_n(r) T^\sharp \omega(r)\, h^{-1\prime}(r) dr = \langle S(\rho), \omega \rangle_{\Omega_1}. \tag{2.40}$$

Since the shifting function is positive, we have that the operator $T_\sharp T^\sharp$ is invertible, and in particular we have the following corollary:

Corollary 2.2.3. *The operator $T_\sharp : \mathscr{D}(\Omega_2) \to \mathscr{D}(\Omega_1)$ is a surjection.*

Having established the previous lemmas, we can now show:

Theorem 2.2.4. *The pullback by T on $\mathscr{D}^*(\Omega_1)$ is a linear operator $T^\sharp : \mathscr{D}^*(\Omega_1) \to \mathscr{D}^*(\Omega_2)$ that is injective, continuous, and unique in the sense that $\langle T^\dagger \rho, \phi \rangle = \langle \rho \circ T, \phi \rangle$ for all $\phi \in \mathscr{D}(\Omega_2)$ and $\rho \in \mathscr{D}(\Omega_1)$ implies $T^\dagger = T^\sharp$ in the sense of evaluation on distributions.*

Proof. We have already established that $T^\sharp : \mathscr{D}^*(\Omega_1) \to \mathscr{D}^*(\Omega_2)$ is a continuous linear operator. Uniqueness is a consequence of Theorem 2.1.2. That is, suppose $T^\dagger : \mathscr{D}^*(\Omega_1) \to \mathscr{D}^*(\Omega_2)$ is a continuous linear functional such that $\langle T^\dagger p, \phi \rangle = \langle p \circ T, \phi \rangle$ for all $\phi \in \mathscr{D}(\Omega_2)$ whenever $p \in \mathscr{D}(\Omega_1)$. Then, for any $p \in \mathscr{D}^*(\Omega_1)$, let $\{p_n\} \subset \mathscr{D}(\Omega_1)$ converge to p (in the $\mathscr{D}^*(\Omega_1)$ sense), so

$$\left\langle (T^\sharp - T^\dagger) p, \phi \right\rangle_{\Omega_2} = \lim \langle T^\sharp p_n, \phi \rangle_{\Omega_2} - \lim \langle T^\dagger p_n, \phi \rangle_{\Omega_2} = 0. \tag{2.41}$$

Hence $T^\sharp = T^\dagger$. Thus, as previously remarked, any choice of T_{ij} to define $T^\sharp$ such that $T^\sharp \phi = \phi \circ T$ on test functions results in an equivalent distribution.

It remains to show that $T^\sharp$ is injective. This follows from the surjectivity of $T_\sharp$; explicitly, suppose $p \in \mathscr{D}^*(\Omega_1)$ is such that $T^\sharp p = 0$, then for ψ an arbitrary test function in $\mathscr{D}(\Omega_1)$

$$0 = \left\langle T^\sharp p, \psi \right\rangle_{\Omega_2} = \left\langle p, T_\sharp \psi \right\rangle_{\Omega_1}, \tag{2.42}$$

implies $p = 0$ as a distribution by Corollary 2.2.3. Hence, $T^\sharp$ has trivial kernel and, as it is a linear map, is injective. We have established all of the properties in Theorem 2.2.4. □

2.2.2 An extension theorem and a motivating example

Using the development in the last section, we can define what it means for a distribution to be radially symmetric. As before, let $\Omega_2 = \mathbb{R}^2 \setminus \{x = 0 \text{ or } y = 0\}, \Omega_1 = (0, \infty)$, and $T : \Omega_2 \to \Omega_1$, henceforth we will consider only $T(x, y) = \sqrt{x^2 + y^2}$ (that is, $h(r) = r^{1/2}$). A distribution $f \in \mathscr{D}^*(\Omega_2)$ is *radially symmetric* if $f = T^\sharp p$ for some distribution $p \in \mathscr{D}^*(\Omega_1)$. We call p the radial profile of f, and the results of the last section establish that the correspondence with p is well-defined. Care must be taken in order to develop radial symmetry that is compatible with the topology of L^2, as it is not necessarily the case that the corresponding radial profile $p \in L^2(\Omega_1)$ even if $f \in L^2(\Omega_2)$. To demonstrate this, we must first establish a result that allows us to explicitly construct Cauchy sequences of test functions from continuous square integrable functions. The construction of test functions is slightly different than the standard development used to prove Theorem 2.1.2 in [Hörmander, 1983; Richtmyer, 1978], in that it will allow us avoid convolving over a singularity.

The following result was established by [Seeley, 1964] for functions defined on $\mathbb{R}^k$, and we state a form on $\mathbb{R}$, whose proof can be found in [Strichartz, 2000].

Theorem 2.2.5. *Suppose $f : [a, b] \to \mathbb{R}$ is smooth in the sense that the one sided limits of the derivatives of f exist and are inductively given by*

$$f^{n+1}(a) = \lim_{h \to 0^+} \frac{f^{(n)}(a + h) - f^{(n-1)}(a)}{h} \tag{2.43}$$

and

$$f^{n+1}(b) = \lim_{h \to 0^-} \frac{f^{(n)}(b + h) - f^{(n-1)}(b)}{h}, \tag{2.44}$$

with $f^{(0)}(a) = f(a)$. Then, for any open $\Omega \subseteq \mathbb{R}$ such that $[a, b] \subset \Omega$, there exists a smooth

compactly supported $\widetilde{f} : \Omega \to \mathbb{R}$ *such that* $\widetilde{f}(x) = f(x)$ *for all* $x \in [a, b]$.

The idea of the proof is to construct continuous 'flaps' on $(a-\varepsilon, a]$ and $[b, b+\varepsilon)$ that converge to zero on the open side like $e^{-1/x}$ as $x \to 0$ and matching the one sided derivatives on the closed side by using an argument similar to one used by Borel to construct smooth functions with arbitrarily prescribed Taylor series.

With this powerful extension theorem, we can easily construct sequences of test functions that are given by restrictions of smooth functions to compact sets. We will use this to construct an important example that will guide the construction of the space of radially symmetric functions and their profiles.

Consider the sequence of functions $\rho_n : \Omega_1 \to \mathbb{R}$ defined by $\rho_n(r) = r^{-1/2}$ for $r \in [1/n, 1]$ and are smoothly extended by Theorem 2.2.5 such that $\rho_n \in \mathscr{D}(\Omega_1)$. Since ρ_n is compactly supported on $\Omega_1 = (0, \infty)$, there exists $\delta > 0$ and a sequence (k_n) such that $0 < k_n < 1/n$ with $\operatorname{supp} \rho_n \subseteq [k_n, 1 + \delta]$, hence, upon showing that it is Cauchy with respect to the L^2 norm, we will have shown that it corresponds to distribution in $L^2(\Omega_2)$. The limit of this sequence defines a distribution, since for a given $\omega \in \mathscr{D}(\Omega_1)$, let $N > 0$ such that $\frac{1}{N} < \inf\{r \in \Omega_1 : r \in \operatorname{supp}(\omega)\}$ and $M = \sup\{r > 0 : r^{-1/2}\omega(r)\}$. Then $\|\rho_n \omega\|_\infty$ is bounded by M for $n \geq N$, hence $\left| \lim_{n\to\infty} \langle \rho_n, \omega \rangle \right| < \infty$.

Now, observe that

$$\|\rho_n\|^2_{L^2(\Omega_1)} = \int_{k_n}^{1+\delta} \rho_n(r)^2 \, dr \geq \int_{1/n}^{1} r^{-1} \, dr = \ln n \tag{2.45}$$

so, ρ_n cannot define a Cauchy sequence in $L^2(\Omega_1)$. On the other hand, the Cauchy sequence $(T^\sharp \rho_n) \subset \mathscr{D}(\Omega_2)$ corresponds to a distribution in $\mathscr{D}^*(\Omega_2)$ by

$$\lim_{n\to\infty} \left\langle T^\sharp \rho_n, \psi \right\rangle_{\Omega_2} = \lim_{n\to\infty} \langle \rho_n, T_\sharp \psi \rangle_{\Omega_1} . \tag{2.46}$$

Moreover, since $T_\sharp\psi$ is compactly supported and the integrand in (2.46) is smooth, we can express the action of the distribution by integration with the point-wise limit $p(r) \stackrel{\text{def}}{=} r^{-1/2} I_{(0,1]} + \rho_1(r) I_{[1,\infty)}$ (since ρ_n changes only for $r < 1/n$),

$$\lim_{n\to\infty} \left\langle T^\sharp \rho_n, \psi \right\rangle_{\Omega_2} = \int_0^\infty p(r)[T_\sharp\psi](r)\, dr. \tag{2.47}$$

Since $p(r)$ is continuous on $(0,\infty)$, we can perform the same change of variables used to define $T_\sharp$ in Lemma 2.2.1, so

$$\lim_{n\to\infty} \left\langle T^\sharp \rho_n, \psi \right\rangle_{\Omega_2} = \iint_{\Omega_2} p\left(\sqrt{x^2+y^2}\right) \psi(x,y)\, dxdy, \tag{2.48}$$

hence, the distribution corresponding to the limit of $(T^\sharp\rho_n)$ is integration against the continuous function $p\left(\sqrt{x^2+y^2}\right)$. Observe that this function is square integrable since $p^2(r) = r^{-1} I_{(0,1]} + \rho_1(r)^2 I_{[1,\infty)}$, and changing variables yields

$$\iint_{\Omega_2} p\left(\sqrt{x^2+y^2}\right)^2 dxdy = \int_0^{2\pi}\int_0^1 dr + 2\pi \int_1^\infty \rho_1(r) r dr < \infty \tag{2.49}$$

since ρ_1 is smooth and compactly supported.

We have shown that $T^\sharp$ is not a closed map with respect to the L^2 topologies, or equivalently, that $T^{\sharp -1}$ is unbounded with respect to the L^2 topologies. This means that the closure of $T^\sharp\mathscr{D}(\Omega_1)$ with respect to the $L^2(\Omega_2)$ inner product is not equal to the image of the closure of $\mathscr{D}(\Omega_1)$ with respect to $L^2(\Omega_1)$ under $T^\sharp$. This lack of agreement is overcome in the same way that it was dealt with in the case of the Sobolev space construction. That is, we let the topology of radial profiles be inherited by enforcing $L^2(\Omega_2)$ convergence of the two-dimensional PSF through the linear map $T^\sharp$. This is carried out in detail in the next section.

2.2.3 Radial symmetry for $L^2(\Omega_2)$ and $\mathscr{H}^n(\mathbb{R}^2)$

As previously mentioned, the structure of the space of radial profiles will be induced by imposing structure on their images under $T^\sharp$. As before, we denote $\Omega_1 = (0, \infty)$ and $\Omega_2 = \mathbb{R}^2 \setminus \{x = 0 \text{ or } y = 0\}$. We also fix $T(x, y) = \sqrt{x^2 + y^2}$, and remark that the arguments can be modified in a straight forward was so that they apply to $T(x, y) = (x^2 + y^2)^a$. We define for a distribution $k \in L^2(\Omega_2)$ to be *radially symmetric* whenever there exists a sequence $(\rho_n) \subset \mathscr{D}(\Omega_1)$, so that $(T^\sharp \rho_n)$ is Cauchy with respect to $L^2(\Omega_2)$,and

$$\left\langle k, \phi \right\rangle_{\Omega_2} = \lim_{n\to\infty} \left\langle T^\sharp \rho_n, \phi \right\rangle_{\Omega_2} = \lim_{n\to\infty} \left\langle \rho_n, T_\sharp \phi \right\rangle_{\Omega_1}. \tag{2.50}$$

This collection forms a linear subspace of $L^2(\Omega_2)$ which we denote $\mathscr{K}^0 \subset \mathscr{D}^*(\Omega_2)$, and by construction is complete with respect to the L^2 norm. Moreover, the completion results in a radially symmetric distribution, as is stated in the following proposition.

Proposition 2.2.6. *The space $\mathscr{K}^0 \subseteq T^\sharp \mathscr{D}^*(\Omega_1)$.*

Proof. Let $\rho_n \in \mathscr{D}(\Omega_1)$ so that $(T^\sharp \rho_n)$ is Cauchy with respect to $L^2(\Omega_2)$. We must show that there exists $p \in \mathscr{D}^*(\Omega_1)$ such that

$$\lim_{n\to\infty} \left\langle T^\sharp \rho_n, \psi \right\rangle_{\Omega_2} = \left\langle T^\sharp p, \psi \right\rangle_{\Omega_2} \tag{2.51}$$

for all $\psi \in \mathscr{D}(\Omega_2)$. Since the limit

$$\lim_{n\to\infty} \left\langle T^\sharp \rho_n, \psi \right\rangle_{\Omega_2} = \lim_{n\to\infty} \left\langle \rho_n, T_\sharp \psi \right\rangle_{\Omega_1} \tag{2.52}$$

for all $\psi \in \mathscr{D}(\Omega_2)$, and by the surjectivity of $T_\sharp$, Corollary 2.2.3, and the localization theorem Theorem 2.1.2, there exists $p \in \mathscr{D}^*(\Omega_1)$ so that

$$\lim_{n\to\infty} \left\langle \rho_n, T_\sharp \psi \right\rangle_{\Omega_1} = \left\langle p, T_\sharp \psi \right\rangle_{\Omega_1} = \lim_{n\to\infty} \left\langle T^\sharp p, \psi \right\rangle_{\Omega_2}. \tag{2.53}$$

□

Now, the space of radial profiles corresponding to radially symmetric $L^2(\Omega_2)$ distributions is $\mathscr{P}^0 = \{p \in \mathscr{D}^*(\Omega_1) : T^\sharp p \in \mathscr{K}^0\}$. Observe that this set is precisely the preimage of $\mathscr{K}^0$ under $T^\sharp$, and similar to the computation in Lemma 2.2.2, imposes a regularity on the shift of p. That is, let $\langle S_{1/2}(p), \omega\rangle_{\Omega_1} \stackrel{\text{def}}{=} \left\langle p, \sqrt{2\pi r}\cdot\omega\right\rangle_{\Omega_1}$ and consider the following proposition:

Proposition 2.2.7. *If $p \in \mathscr{P}^0$ then $S_{1/2}(p) \in L^2(\Omega_1)$.*

Proof. We will construct a sequence of test functions $(\rho_n) \subset \mathscr{D}(\Omega_1)$ that converges in $L^2(\Omega_1)$ to p. By construction, $p \in \mathscr{P}^0$ implies that there is a sequence of test functions $(\rho_n) \subset \mathscr{D}(\Omega_1)$ such that $(T^\sharp\rho_n)$ is Cauchy with respect to $L^2(\Omega_2)$ and

$$\langle p, \omega\rangle_{\Omega_1} = \lim_{n\to\infty}\langle \rho_n, \omega\rangle_{\Omega_1} \tag{2.54}$$

for all $\omega \in \mathscr{D}(\Omega_1)$. The action of the shift on test functions is $[S_{1/2}(\rho_n)](r) = \sqrt{2\pi r}\rho_n(r)$ since ρ_n are smooth. Now, observe

$$\begin{aligned}
\|S_{1/2}(\rho_n) - S_{1/2}(\rho_k)\|^2_{L^2(\Omega_1)} &= \int_{\Omega_1}(\rho_n - \rho_k)\left(2\pi r\cdot\overline{(\rho_n-\rho_k)}\right)dr \\
&= \left\langle(\rho_n-\rho_k), T_\sharp T^\sharp\overline{(\rho_n-\rho_k)}\right\rangle_{\Omega_1} \\
&= \left\langle T^\sharp(\rho_n-\rho_k), \overline{T^\sharp(\rho_n-\rho_k)}\right\rangle_{\Omega_2} \\
&= \left\|T^\sharp(\rho_n-\rho_k)\right\|^2_{L^2(\Omega_2)}.
\end{aligned} \tag{2.55}$$

Hence $\big(S_{1/2}(\rho_n)\big)$ inherits the Cauchy criterion from $(T^\sharp\rho_n)$, and by completeness of $L^2(\Omega_1)$, there exists a distribution $g \in L^2(\Omega_1)$ corresponding to $\big(S_{1/2}(\rho_n)\big)$. That is,

$$\left\langle g, \omega\right\rangle_{\Omega_1} = \lim_{n\to\infty}\left\langle S_{1/2}(\rho_n), \omega\right\rangle_{\Omega_1} = \lim_{n\to\infty}\int_{\Omega_1}\rho_n(r), \sqrt{2\pi r}\cdot\omega(r)dr = \left\langle S_{1/2}(p), \omega\right\rangle_{\Omega_1}. \tag{2.56}$$

□

In fact, the computation in (2.55) gives precisely the method for translating the structure of $L^2(\Omega)$ onto $\mathscr{P}^0$.

Theorem 2.2.8. *The space $\mathscr{P}^0$ is the completion of $\mathscr{D}(\Omega_1)$ with inner product and norm given by $(\rho,\omega)_{T(\Omega_1)} = \big(S_{1/2}(\rho), S_{1/2}(\omega)\big)_{L^2(\Omega_1)}$. Moreover, the space $\mathscr{P}^0$ with this inner product is isometric with $\mathscr{K}^0 \subset L^2(\Omega_2)$.*

Proof. Note that $T^{\sharp^{-1}}\mathscr{K}^0 = \mathscr{P}^0$ by definition, hence restricting $T^\sharp : \mathscr{P}^0 \to \mathscr{K}^0$ is a surjection, and by Theorem 2.2.4 a bijection. Moreover, for any $p \in \mathscr{P}^0$, let (ρ_n) be a corresponding Cauchy sequence with respect to $L^2(\Omega_1)$, then the same calculation as (2.55) yields

$$\|p\|^2_{T(\Omega_1)} = \lim_{n\to\infty} \|(S_{1/2}(\rho_n)\|^2_{L^2(\Omega_1)} = \lim_{n\to\infty} \|T^\sharp\rho_n\|^2_{L^2(\Omega_2)}, \tag{2.57}$$

Thus, $T^\sharp$ restricted to $\mathscr{P}^0$ into $\mathscr{K}^0$ is an isometry, and since $\mathscr{K}^0$ is the completion of $T^\sharp\mathscr{D}(\Omega_1)$ with respect to $\|\cdot\|_{L^2(\Omega_2)}$, it must be that $\mathscr{P}^0$ is the completion of $\mathscr{D}(\Omega_1)$ with respect to $\|\cdot\|_{T(\Omega)}$. □

A straightforward modification of the exact same construction by replacing $L^2(\Omega_2)$ with $\mathscr{H}^n(\Omega_2)$ gives rise to radial profiles with regularity that is inherited from $\mathscr{H}^n(\Omega_2)$. That is, $k \in \mathscr{K}^1 \subset \mathscr{D}^*(\Omega_2)$ provided there exists a sequence $(\rho_n) \subset \mathscr{D}(\Omega_1)$, so that $(T^\sharp\rho_n)$ is Cauchy with respect to the Sobolev inner product $(\cdot,\cdot)_{\Omega_2,1}$ and corresponds with k as in (2.50). Note that $\mathscr{K}^1 \subseteq \mathscr{K}^0 \subseteq T^\sharp\mathscr{D}^*(\Omega_1)$. The following proposition results from the multi-dimensional chain rule and links the Sobolev inner products to the $L^2(\Omega)$ norms.

Proposition 2.2.9. *If $\rho,\omega \in \mathscr{D}(\Omega_1)$, then $(\nabla T^\sharp\rho, \nabla T^\sharp\omega)_{L^2(\Omega_2)} = (\partial\rho,\partial\omega)_{T(\Omega_1)}$.*

Proof. First, note that the chain rule for partial derivatives gives the shift $\partial_x T^\sharp\rho = \partial_x(\rho\circ T) = \rho'\circ T\cdot T_x$, where $T_x = x(x^2+y^2)^{-1/2} = x/T(x,y)$. So $T_x(x,y)^2 + T_y(x,y)^2 = T(x,y)$ implies $T_x^2 + T_y^2 = T^\sharp Id_\Omega$, where Id_Ω denotes the identity on Ω_1. Using these facts, consider the

following computation,

$$\begin{aligned}(\nabla T^\sharp\rho, \nabla T^\sharp\omega)_{L^2(\Omega_2)} &= (\partial_x T^\sharp\rho, \partial_x T^\sharp\omega)_{L^2(\Omega_2)} + (\partial_y T^\sharp\rho, \partial_y T^\sharp\omega)_{L^2(\Omega_2)} \\ &= \Big(T^\sharp(\partial\rho), T^\sharp(\partial\omega)\cdot T_x^2\Big)_{L^2(\Omega)} + \Big(T^\sharp(\partial\rho), T^\sharp(\partial\omega)\cdot T_x^2\Big)_{L^2(\Omega_2)} \\ &= \Big(T^\sharp(\partial\rho), T^\sharp(\partial\omega)\cdot T^\sharp Id_{\Omega_1}\Big)_{L^2(\Omega_2)} \\ &= \Big(T^\sharp(\partial\rho), T^\sharp(\partial\omega)\Big)_{L^2(\Omega_2)} \\ &= \Big(\partial\rho, \partial\omega\Big)_{T(\Omega_1)}. \end{aligned} \tag{2.58}$$

□

We remark that this computation can be modified slightly to accommodate the more general form of $T(x,y) = (x^2+y^2)^a$, and instead of a shift by $T^\sharp Id_{\Omega_1}$, the shift is $T^\sharp 4a^2(r^{2-1/a})$.

Now, let $\mathscr{P}^n = \{p \in \mathscr{D}^*(\Omega_1) : T^\sharp p \in \mathscr{K}^n\}$, and by inductively applying the preceding result in Proposition 2.2.9, and Proposition 2.2.7, the induced Sobolev semi-norm is

$$(p,q)_{T(\Omega),n} \stackrel{\text{def}}{=} \lim_{m\to\infty}\sum_{k=0}^{n}(\nabla^k T^\sharp\rho_m, \nabla^k T^\sharp\omega_n)_{L^2(\Omega_2)} = \lim_{m\to\infty}\sum_{k=0}^{n}(\partial^k\rho_m, \partial^k\omega_m)_{T(\Omega_1)}, \tag{2.59}$$

where (ρ_n) and (ω_n) are the corresponding Cauchy sequence to p and q respectively.

2.3 The PSF inverse problem

The ultimate goal of this theory will be to rigorously develop the inverse problem introduced at the end of Chapter 1. The development so far has defined the space of radial representations on distributions only on $\Omega_1 = (0,\infty)$. A complete description of the space will require that we have some notion of the distributions p on the boundary, i.e. if p has a representation as a locally integrable function, what is $p(0)$ and $\lim_{r\to\infty} p(r)$? The theory to rigorously develop this

is ongoing, however, we will state results that will provide insight on how to proceed.

2.3.1 The Sobolev embedding theorem and extending to $\mathscr{P}^n([0,\infty))$

We now state a version of the Sobolev embedding theorem that shows how higher order Sobolev spaces enforce regularity in terms of spaces of continuous functions, whose proof can be found in [Strichartz, 2003].

Theorem 2.3.1. *Suppose $\|\partial^\alpha f\|_{L^2(\mathbb{R}^N)}$ is finite whenever all α are such that $|\alpha| < n$ where n is equal to the smallest integer greater than $n/2$ (so $n = \frac{N+1}{2}$ if N is odd and $n = \frac{N+2}{2}$ if N is even). Then f corresponds to integration against a continuous and bounded function with*

$$\|f\|_\infty \leq c \sum_{0\leq|\alpha|\leq n} \|\partial^\alpha f\|_{L^2(\mathbb{R}^N)}. \tag{2.60}$$

More generally, if $\|\partial^\alpha f\|_{L^2(\mathbb{R}^N)}$ is finite for all α such that $|\alpha| < n + k$, then there is a continuous embedding of $\mathscr{H}^n(\mathbb{R}^N)$ into $C^k(\mathbb{R}^N)$.

In particular, we will be concerned primarily with the case when $N = 2$ and $n = 2$, from which (2.60) reduces to

$$\sum_{0\leq|\alpha|\leq 2} \|\partial^\alpha f\|_{L^2(\mathbb{R}^N)} = \|f\|_{L^2(\mathbb{R}^N)} + \|\nabla f\|_{L^2(\mathbb{R}^N)} + \|\nabla^2 f\|_{L^2(\mathbb{R}^N)}. \tag{2.61}$$

This norm corresponds to PSFs with second order negative Laplacian, or biharmonic, regularization, and Theorem 2.3.1 implies that these distributions are bounded continuous functions. This is formally stated in the following corollary

Corollary 2.3.2. *For $\mathbb{R}^N \subseteq \mathbb{R}^2$, suppose $\|\nabla^2 f\|_{L^2(\mathbb{R}^N)} < \infty$ then f is continuous and bounded with*

$$\|f\|_\infty \leq c\left(\|f\|_{L^2(\mathbb{R}^N)} + \|\nabla f\|_{L^2(\mathbb{R}^N)} + \|\nabla^2 f\|_{L^2(\mathbb{R}^N)}\right) \tag{2.62}$$

for some real c.

This theorem is not directly applicable since it is only stated for domains $\Omega = \mathbb{R}^k$. Yet, if we can appropriately extend $k \in \mathscr{K}^n \subset \mathscr{D}^*(\Omega_2)$ to a distribution on $\mathscr{D}(\mathbb{R}^2)$, then Corollary 2.3.2 implies that the extension corresponds to integration with a bounded continuous function. Providing this extension rigorously is a delicate, and the work is ongoing. The goal of this future work is to prove the following claim:

Claim 2.3.3. *If $k \in \mathscr{K}^2 \subseteq \mathscr{H}^2(\Omega_2)$, then k is equal in distribution to a bounded continuous function and k continuously extends to $\mathbb{R}^2$. Moreover, this implies that corresponding radial profile $p \in \mathscr{P}^2$ has $S_{1/2}(p) \in \mathscr{H}^2(\Omega_2)$, hence is equal in distribution to a bounded continuous function on Ω_1 and p can be continuously to $\overline{\Omega}_1 = [0, \infty)$.*

Chapter 3

Markov Chains and Modified Gibbs Sampling

Observation informed estimation within a stochastic model can be analytically intractable, especially when the model deviates from standard models for independent or systematically sampled data. This is especially the case in many Bayesian methods, where inference is typically drawn from a posterior distribution, usually known only up to a constant of normalization. *Monte Carlo* methods use pseudo-random simulation methods to construct a simulated sample in order to characterize and estimate statistics about the underlying intractable probability density. This chapter is devoted to introducing Monte Carlo methods that take advantage of a stochastic structure known as a Markov chain in order to perform inference. In particular, we investigate modifications to the widely used Gibbs sampler, and present modifications that improve its convergence. The main tool for inference is the ergodic theorem for Markov chains, which is stated along with its requisite hypotheses in Section 3.1. This development will establish an important necessary condition, the invariance of the Markov chain with respect to the target density, which is crucial to appropriately applying ergodic based inference on the Markov chain.

We derive the standard Gibbs sampling algorithm, outlined by [Geman and Geman, 1984], and present a modification using a technique called *partial collapse*, which is motivated by several recent theoretical and practical analyses [Van Dyk and Park, 2008; Agapiou et al., 2014; Fox and Norton, 2015]. This chapter gives a full development from first principles, and proves the assertions of invariance stated but not explicitly shown in [Van Dyk and Park, 2008]. We will also briefly review standard convergence diagnostics for comparing Markov chain based sampling algorithms, which will establish statistical benchmarks to show that the adapted algorithm is indeed an enhancement of standard Gibbs sampling when applied to PSF estimation. Chapter 4 will apply this general framework to the Bayesian PSF estimation problem.

3.1 Markov Chain Monte Carlo Simulation

In this section we give an overview of Markov chain Monte Carlo (MCMC) methods for analyzing a probability distribution known up to a scaling constant. Statistical analysis is based on the ergodic theorem for Markov chains, which is analogous to the central limit theorem for independently sampled data. The theory will be briefly overviewed in the next section. Complete treatments can be found in [Robert and Casella, 2013]. Our development will lead to an algorithm based on Gibbs sampling that uses a technique referred to as partial collapse. In partially collapsed Gibbs sampling, conditional densities are modified to remove problematic dependence within steps in the Gibbs sampler. Our use of partial collapse will be motivated by the marginal algorithm in [Agapiou et al., 2014], a similar infinite dimensional sampler.

The development undertaken in the general setting of Gibbs sampling with potential modifications to the hierarchical model in mind. In [Howard et al., 2016], they observed potential sensitivity to the uninformative hyper-prior parameters in a similar hierarchical Bayesian es-

timation problem. A possible extension that may alleviate the sensitivity would be to impose a prior on these parameters, forming an additional level of hierarchy to allow flexibility to sample each of these parameters. Additionally, the analysis of the partially collapsed Gibbs samplers presented in [Van Dyk and Jiao, 2015; Van Dyk and Park, 2008] does not provide an argument that the resulting Markov chains remain invariant, and the following discussion fills that gap in the literature.

3.1.1 Markov Chains

This subsection is devoted to developing the preliminary notions of Markov chains and the prerequisite theory for using Markov chains for Monte Carlo estimation. We assume a probability (measure) space $(\Omega, \mathcal{F}, \mathbb{P})$ where Ω is the set of outcomes, with $\mathcal{F}$ a sigma-algebra of events from Ω and $\mathbb{P}$ a measure on $\mathcal{F}$ into $[0, 1]$. We will often be concerned with sampling an m component (each component possibly multivariate) random variable

$$\boldsymbol{X} = (X_1, \ldots, X_m) : \Omega \to \mathbb{R}^M, \tag{3.1}$$

so that the measure $\mu_{\mathbf{x}}$ on $\mathbb{R}^M$ (known as its *law*), induced by $\boldsymbol{X}$ by taking pre-images of Borel sets in $\mathbb{R}^M$ is absolutely continuous with respect to Lebesgue measure. This means that events corresponding to zero Lebesgue measure sets in $\mathbb{R}^M$ have zero probability. Hence, the law corresponding to $\boldsymbol{X}$ has a Radon-Nykodym derivative with respect to Lebesgue measure on $\mathbb{R}^M$, which we refer to as its *density*, denoted by

$$\pi_{\mathbf{x}}(\boldsymbol{x}) = \pi_{\mathbf{x}}(x_1, \ldots, x_m), \tag{3.2}$$

where $x_i \in \mathbb{R}^{k_i}$ such that $\sum_{i=1}^{m} k_i = M$ characterize the range of each component of $\boldsymbol{X}$. Note that the support of the density contains all sets corresponding to non-zero probability events.

We will also occasionally consider random variables whose ranges are fixed, finite, and

discrete sets $\{\omega_1, \ldots, \omega_s\} \subset \mathbb{R}$ with corresponding probabilities $\{\alpha_1, \ldots, \alpha_s\}$ summing to 1, i.e.,

$$\mathbb{P}(X = \omega_i) = \alpha_i. \tag{3.3}$$

In this case, the associated law is no longer absolutely continuous with respect to Lebesgue measure since finite discrete sets have zero Lebesgue measure, so $\boldsymbol{X}$ does not have a real valued density function. However, we can use a finite linear combination of Dirac distributions translated to ω_i with coefficients α_i as a notion of a density where, for consistency of notation, we denote the action of the Dirac distributions with

$$\mathbb{P}(X \in A \subset \{\omega_1, \ldots, \omega_s\}) = \int_A (\alpha_1 \delta_{\omega_1} + \cdots + \alpha_s \delta_{\omega_s})\, dx. \tag{3.4}$$

When two or more of the variables $(X_1, \ldots, X_m)$ are considered together, referred to as blocking, the resulting variable is given in boldface, although each component may be itself multivariate. For a complete development of the measure-based probabilistic formulation of random variables see [Durrett, 2010; Billingsley, 2008]. When the density is clear from context, we will omit the subscript on $\pi(\boldsymbol{x})$. For any subset $\{j_i\}_{i=1}^k \subset \{1, \ldots, m\}$, let $\boldsymbol{x}_{\widehat{j_i}}$ denote the vector with each of the j_ith components removed, then the *marginal distribution* is

$$\pi(\boldsymbol{x}_{\widehat{j_i}}) \stackrel{\text{def}}{=} \int_{x_{j_1}} \ldots \int_{x_{j_k}} \pi(x_1, \ldots, x_m) dx_{j_k} \ldots dx_{j_1}, \tag{3.5}$$

and the *conditional distribution* of $(x_{j_1}, \ldots, x_{j_k})$ is

$$\pi(x_{j_1}, \ldots, x_{j_k} | \boldsymbol{x}_{\widehat{j_i}}) \stackrel{\text{def}}{=} \frac{\pi(\boldsymbol{x})}{\pi(\boldsymbol{x}_{\widehat{j_i}})}. \tag{3.6}$$

A family of probability densities $K(\boldsymbol{x}, \cdot)$ is a *transition kernel*, if for all $\boldsymbol{x} \in \mathbb{R}^M$, $K(\boldsymbol{x}, \cdot)$, it

defines a probability measure given by

$$\int_A K(\boldsymbol{x}, \boldsymbol{x}')d\boldsymbol{x}' = \mathbb{P}\left(\boldsymbol{X} \in A\right), \tag{3.7}$$

and $K(\cdot, \boldsymbol{x}')$ is absolutely integrable. For a transition kernel, the corresponding *transition operator* acts on an absolutely integrable π by

$$\mathcal{K}[\pi](\boldsymbol{x}') = \int K(\boldsymbol{x}, \boldsymbol{x}')\pi(\boldsymbol{x})d\boldsymbol{x}. \tag{3.8}$$

Note that $\mathcal{K} : L^1(\mathbb{R}^K) \to L^1(\mathbb{R}^K)$ is a linear operator such that $\|\mathcal{K}f\|_{L^1} \leq \|f\|_{L^1}$ since $K(\boldsymbol{x}, \cdot)$ is a probability measure, i.e., $\int |K(\boldsymbol{x}, \boldsymbol{x}')|d\boldsymbol{x}' = 1$.

A Markov chain is a stochastic process $\{\boldsymbol{X}^0, \boldsymbol{X}^1, \boldsymbol{X}^2, \dots, \}$ with $\boldsymbol{X}^k : \Omega \to \mathbb{R}^M$ defined on a common probability space such that for a given transition kernel K,

$$\mathbb{P}\left(\boldsymbol{X}^{k+1} \in A | \boldsymbol{X}^k = \boldsymbol{x}^k, \dots, \boldsymbol{X}^0 = \boldsymbol{x}^0\right) = \mathbb{P}\left(\boldsymbol{X}^{k+1} \in A | \boldsymbol{X}^k = \boldsymbol{x}^k\right) = \int_A K(\boldsymbol{x}^k, \boldsymbol{x}')d\boldsymbol{x}' \tag{3.9}$$

for all events A, i.e., the random variable $\boldsymbol{X}^{k+1}$ depends only on the previous realization $\boldsymbol{X}^k = \boldsymbol{x}^k$, and subsequent densities of the elements in the Markov chain are given by the action of the transition operator.

In order to establish intuition with many of the notions of Markov chains, we will often provide examples of them on finite discrete state Markov chains, from which many of these concepts were initially developed [Billingsley, 2008]. Suppose $X^k \in \{\omega_1, \dots, \omega_s\}$ for all k, then all probability densities are finite linear combinations of Dirac measures as in (3.4), where the coefficients are the probabilities of transitioning to that state, i.e.,

$$\int_A K(\omega_i, x)dx = \mathbb{P}(X^{k+1} \in A \subseteq \{\omega_1, \dots, \omega_n\} | X^k = \omega_i) = \sum_{i=1}^{n} k_{i,j} \int_A \delta_{\omega_i} dx, \tag{3.10}$$

where $K(\omega_i, \omega_j) = k_{i,j}$ and $\sum_{j=1}^{n} k_{i,j} = 1$. Taking the states $\{\delta_{\omega_1}, \dots, \delta_{\omega_n}\}$ as basis vectors,

the probability densities form a finite dimensional vector space, and transition operators correspond to multiplication by a *transition matrix* $\boldsymbol{K}$ with entries k_{ij}. To see this, note the action of $\mathcal{K}$ on a given density $\pi = \alpha_1\delta_{\omega_1} + \ldots \alpha_n\delta\omega_n$ is

$$\begin{aligned}
\mathcal{K}(\alpha_1\delta_{\omega_1} + \ldots \alpha_n\delta\omega_n) &= \alpha_1 K(\omega_1, x) + \ldots \alpha_n K(\omega_n, x) \\
&= [\delta_{\omega_1} \ldots \delta\omega_n] \begin{bmatrix} K(\omega_1,\omega_1) & K(\omega_1,\omega_2) & \ldots & K(\omega_1,\omega_n) \\ K(\omega_2,\omega_1) & \ddots & \ldots & \vdots \\ \vdots & \ldots & \ddots & \vdots \\ K(\omega_n,\omega_1) & K(\omega_n,\omega_2) & \ldots & K(\omega_n,\omega_n) \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix} \\
&= [\delta_{\omega_1} \ldots \delta\omega_n]\boldsymbol{K}\boldsymbol{\alpha}. \qquad (3.11)
\end{aligned}$$

Now, consider the joint density $\pi(\boldsymbol{x}^0, \ldots, \boldsymbol{x}^N)$ for the truncated chain $\{\boldsymbol{X}^0, \ldots, \boldsymbol{X}^N\}$ with $\pi_0(\boldsymbol{x})$ the density for $\boldsymbol{X}_0$, then the definition in (3.8) implies

$$\begin{aligned}
\pi(\boldsymbol{x}^1) &= \int_{\boldsymbol{x}^0} \pi(\boldsymbol{x}^1, \boldsymbol{x}^0) d\boldsymbol{x}^0 \\
&= \int_{\boldsymbol{x}^0} \pi(\boldsymbol{x}^1|\boldsymbol{x}^0)\pi(\boldsymbol{x}^0) d\boldsymbol{x}^0 \\
&= \int_{\boldsymbol{x}^0} K(\boldsymbol{x}^0, \boldsymbol{x}^1)\pi(\boldsymbol{x}^0) d\boldsymbol{x}^0 \\
&= \mathcal{K}[\pi_0](\boldsymbol{x}^1) \qquad (3.12) \\
\pi(\boldsymbol{x}^2) &= \int_{\boldsymbol{x}^1}\int_{\boldsymbol{x}^0} \pi(\boldsymbol{x}^2, \boldsymbol{x}^1, \boldsymbol{x}^0) d\boldsymbol{x}^0 d\boldsymbol{x}^1 \\
&= \int_{\boldsymbol{x}^1} \pi(\boldsymbol{x}^2|\boldsymbol{x}^1) \int_{\boldsymbol{x}^0} \pi(\boldsymbol{x}^1, \boldsymbol{x}^0) d\boldsymbol{x}^0 d\boldsymbol{x}^1 \\
&= \int_{\boldsymbol{x}^1} K(\boldsymbol{x}^1, \boldsymbol{x}^2) \int_{\boldsymbol{x}^0} \pi(\boldsymbol{x}^1, \boldsymbol{x}^0) d\boldsymbol{x}^0 d\boldsymbol{x}^1 \\
&= \mathcal{K}\Big(\mathcal{K}[\pi_0](\boldsymbol{x}^2)\Big) \qquad (3.13) \\
&\vdots \\
\pi(\boldsymbol{x}^N) &= \int_{\boldsymbol{x}^{N-1}} \ldots \int_{\boldsymbol{x}^0} \pi(\boldsymbol{x}^N, \boldsymbol{x}^{N-1}, \ldots, \boldsymbol{x}^0) d\boldsymbol{x}_0 \ldots, d\boldsymbol{x}_{N-1} \\
&= \int_{\boldsymbol{x}^{N-1}} \ldots \int_{\boldsymbol{x}^0} \pi(\boldsymbol{x}^N|\boldsymbol{x}^{N-1})\pi(\boldsymbol{x}^{N-1}, \ldots, \boldsymbol{x}^0) d\boldsymbol{x}_0 \ldots d\boldsymbol{x}_{N-1}
\end{aligned}$$

$$= \int_{\boldsymbol{x}^{N-1}} K(\boldsymbol{x}^{N-1}, \boldsymbol{x}^N) \dots \int_{\boldsymbol{x}^0} K(\boldsymbol{x}^0, \boldsymbol{x}^1)\pi(\boldsymbol{x}^0)d\boldsymbol{x}_0 \dots d\boldsymbol{x}_{N-1}$$
$$= \mathcal{K}^N[\pi_0](\boldsymbol{x}^N). \tag{3.14}$$

So, the Nth marginal density of the Markov chain is given by the Nth composition of the transition operator $\mathcal{K}$ on the initial density π_0. In some sense, all of the information of the Markov chain up to $\boldsymbol{X}^N$ is embedded in the transition operator $\mathcal{K}$, since each marginal density and all conditional probabilities are encoded into $K(\boldsymbol{x}, \boldsymbol{x}')$. Furthermore, we see that it is natural to think of a Markov chain evolving as N increases, with the evolution given by successively iterating $\mathcal{K}$. With this in mind, two natural questions arise, "How does the initial state effect the chain and what is its end behavior?" These notions are encapsulated by *irreducibility* and *stationarity*, respectively.

For a given measure λ, a Markov chain is λ-irreducible if for every event A with $\lambda(A) > 0$, there exists an N such that $\int_A \mathcal{K}^N(\boldsymbol{x}, \boldsymbol{x}')dx' > 0$ [Robert and Casella, 2013]. This means that every event that can be measured by λ has a positive probability of being reached by the Markov chain in a finite number of steps. When the number of steps is 1, the chain is called *strongly irreducible*, and this holds for transition kernels with full support, that is, the probability distribution associated with $K(\boldsymbol{x}, \cdot)$ has a range that is positive over the range of each $\boldsymbol{X}^k$. For the continuous densities of interest to us, which are derived from Gaussian and gamma distributions, this condition is satisfied [Liu, 2008].

A Markov chain with transition operator $\mathcal{K}$ is *stationary* with an *invariant* density π if

$$\mathcal{K}[\pi(\boldsymbol{x})](\boldsymbol{x}') = \pi(\boldsymbol{x}'). \tag{3.15}$$

Note that an invariant distribution π is an eigenvector for the transition operator $\mathcal{K}$ corresponding to the eigenvalue 1. Since transition operators consist of probability densities, then $\int |\pi(x)| \leq 1$ implies all eigenvalues are bounded in modulus by 1.

When the Markov chain has a finite and discretely supported transition kernel, there is an interesting connection with the concept of stationary and the power-iteration method for finding leading order eigenvalues and corresponding eigenvectors. In the power-iteration method, the sequence given by the recursive relation $\boldsymbol{\alpha}^k = \boldsymbol{K}\boldsymbol{\alpha}^{k-1}/\|\boldsymbol{K}\boldsymbol{\alpha}^{k-1}\|_{\mathbb{R}^n}$ can be shown to converge to the leading order (in modulus) eigenvector, and when $\boldsymbol{\alpha}^k$ correspond to finite discrete probability densities, the normalization is not required. Hence, in the discrete finite case, the invariant density has coefficients such that $\lim_{k\to\infty} \boldsymbol{\alpha}_k = \lim_{k\to\infty} \boldsymbol{K}^k\boldsymbol{\alpha}_0$. One of the main results of the ergodic theorem for Markov chains is to extend this notion to continuous probability densities.

There are two last technical conditions that must be defined in order to establish the hypotheses of the ergodic theorem for Markov chains, known as Harris recurrence and aperiodicity. To illustrate aperiodicity, we will define it first for finite discrete Markov chains. The *period* of a state ω_i is the greatest common divisor of the set $\{k \geq 1 : K^k(\omega_i, \omega_i) > 0\}$; that is, if ω_i is d-periodic, then returns to state ω occur in multiples of d. For example, the simple deterministic two state Markov chain associated with the transition matrix

$$\boldsymbol{K} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \tag{3.16}$$

jumps between two states with probability 1 and has period 2. A chain is *aperiodic* if each state has period 1. To extend this definition to continuously supported Markov chains that take values in $\mathbb{R}^M$, see the analogous notion in [Robert and Casella, 2013, Chapter 6.3]. Defining aperiodicity precisely for continuous state Markov chains requires probability theory that is beyond the scope of this work, but can be thought of informally as a Markov chain whose transition kernel has orbits (with respect to iteration) that do not get trapped into cycles and visit enough of the support of the invariant density, regardless of the initial density π_0, to asymptotically approximate π. Verifying rigorously the requirement of aperiodicity for continuous Markov chains is technical, and again we cite [Liu, 2008] who states that

transition kernels associated with Gibbs sampling with Gaussian and gamma conditionals and Metropolis-Hastings sampling with Gaussian proposals are aperiodic, and the algorithms presented in this work are compositions of such transitions. See [Robert and Casella, 2013] for the complete definition and details.

The other technical condition that must be addressed to state the ergodic theorem is Harris recurrence. This condition ensures that a Markov chain re-enters events often enough to 'fill-out' π. Formally, for a Borel set $A \subseteq \mathbb{R}^M$ and its indicator function I_A, the *average number of passages* of $(\boldsymbol{X}^k)$ in A is the random variable (possibly infinite valued)

$$\eta_A \stackrel{\text{def}}{=} \sum_{k=1}^{\infty} I_A(\boldsymbol{X}^k), \tag{3.17}$$

and a Markov chain is *Harris recurrent* if $\mathbb{P}(\eta_A = \infty | X_0 = x) = 1$ [Robert and Casella, 2013]. Again, verifying this condition is beyond the scope of this work, and we cite [Liu, 2008] who ensures that transitions from Gaussian and gamma densities associated with Gibbs sampling and Metropolis-Hastings sampling with Gaussian proposals are Harris recurrent.

We now state the main theorem that allows for the end behaviour Markov chains to be used as tools for estimating statistics of a given probability distribution:

Theorem 3.1.1. *[Tierney, 1994] Suppose $\mathcal{K}$ defines a stationary Markov chain with invariant density π. If the chain is π-irreducible and Harris recurrent, then π is unique for π-almost all $\boldsymbol{x}$. Moreover,*

(i) Almost surely with respect to π, for any integrable h

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} h(\boldsymbol{X}^n) = \int h(\boldsymbol{x})\pi(\boldsymbol{x})d\boldsymbol{x}. \tag{3.18}$$

(ii) If in addition, the chain is aperiodic, then

$$\lim_{N\to\infty} \|\mathcal{K}^N \pi_0 - \pi\|_{TV} = 0, \tag{3.19}$$

where $\|\pi\|_{TV}$ *denotes the supremum of* $\int_A \pi(x)dx$ *over all Borel sets* A*.*

Equation (3.18) of the ergodic theorem is analogous to the Law of Large Numbers for independent samples and allows us to use chain averages to estimate statistics about π. Equation (3.19) justifies using the 'late stages' of the chain as approximate samples of π.

The goal of MCMC methods is to simulate a Markov chain *designed* so that it has π as its invariant density. In the context of our Bayesian hierarchical model, this will be the discrete approximation to the posterior density. A widely used method, known as Gibbs sampling, can be easily implemented when sampling from full conditional distributions is available and is presented in the next section.

3.1.2 Gibbs sampling

The origin of the Gibbs sampler is relatively recent (despite its eponymous relation to the 19th century physicist Josiah Gibbs) and has its origins in computational imaging. In [Geman and Geman, 1984], they modeled the spatial structure of pixels in an image via the Gibbs distribution, which originally arose from modeling particles in a lattice system [Ising, 1925]. They developed a simulation algorithm for approximating the mode of the posterior of the Gibbs distribution. Because of its ease of implementation and ubiquitous application, the Gibbs sampler has become the workhorse of the MCMC world [Robert and Casella, 2013], and arguably, its fame has overtaken that of its namesake. When the Gibbs sampler is applied to hierarchical Bayesian posteriors, it is sometimes referred to as the hierarchical Gibbs sampler, as is the case in this work.

The following algorithm outlines Gibbs sampling for simulating the transition of a general m-component Markov chain:

Algorithm 1 Gibbs sampler

Given $\boldsymbol{x}^{k-1} = (x_1^{k-1}, \ldots, x_m^{k-1})$, simulate

1. $X_1^k \sim \pi(x_1 | x_2^{k-1}, x_3^{k-1}, \ldots, x_m^{k-1})$
2. $X_2^k \sim \pi(x_2 | x_1^k, x_3^{k-1}, \ldots, x_m^{k-1})$

...

m. $X_m^k \sim \pi(x_m | x_1^k, x_2^k, \ldots, x_{m-1}^k)$

Algorithm 1 simulates outcomes from the transition kernel

$$K(\boldsymbol{x}, \boldsymbol{x}') = \pi(x'_m | x'_1, \ldots, x'_{m-1}) \ldots \pi(x'_2 | x'_1, x_3, \ldots, x_m) \pi(x'_1 | x_2, \ldots, x_m). \tag{3.20}$$

Note that we can view the action of the transition in iterated integrations since subsequent conditional densities factor out of antecedent integrations, i.e.

$$\begin{aligned}
\mathcal{K}[\pi_0](\boldsymbol{x}') &= \int K(\boldsymbol{x}, \boldsymbol{x}') \pi_0(\boldsymbol{x}) d\boldsymbol{x} \\
&= \int_{x_m} \ldots \int_{x_1} \pi(x'_m | x'_1, \ldots, x'_{m-1}) \ldots \pi(x'_2 | x'_1, x_3, \ldots, x_m) \pi(x'_1 | x_2, \ldots, x_m) \pi_0(\boldsymbol{x}) dx_1 \ldots dx_m \\
&= \int_{x_m} \pi(x'_m | \boldsymbol{x}'_{\widehat{m}}) \int_{x_{m-1}} \pi(x'_{m-1} | \boldsymbol{x}'_{\widehat{m,m-1}} x_m) \ldots \int_{x_1} \pi(x'_1 | \boldsymbol{x}_{\widehat{1}}) \pi_0(x_1, \ldots, x_m) dx_1 \ldots dx_m.
\end{aligned} \tag{3.21}$$

Each integration in (3.21) can be thought of as a composition of sub-transitions on $\pi_0(x_1, \ldots, x_m)$; that is, given $(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_m)$, let

$$\mathcal{K}_i[\pi_0(\boldsymbol{x})](\boldsymbol{x}') \stackrel{\text{def}}{=} \int_{x_i} \pi(x'_i | x'_1, \ldots, x'_{i-1}, x_{i+1}, \ldots, x_m) \pi_0(\boldsymbol{x}) dx_i, \tag{3.22}$$

then we can express $\mathcal{K} = \mathcal{K}_m \mathcal{K}_{m-1} \ldots \mathcal{K}_1$. Note that, functionally, each operator $\mathcal{K}_i$ depends on $(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_p)$ being given, and that only after successively integrating each sub-transition is the operator uniquely defined. For example, $\mathcal{K}_1$ depends on $(x_2, \ldots, x_p)$, $\mathcal{K}_2 \mathcal{K}_1$ depends on $(x_3, \ldots, x_m)$, etc., until the full composition in $\mathcal{K}$ does not depend on $\boldsymbol{x}$.

In this form, it will be easy to see that the Gibbs sampler is invariant with respect to π, and the technique used in the proof by characterizing sub-transitions (alluded to in [Robert and Casella, 2013], but not carried out in full detail) will be useful for designing and verifying the stationarity of algorithms that modify the Gibbs sampler in the following sections.

Proposition 3.1.2. *The transition kernel associated with Algorithm 1 produces a Markov chain that is invariant to the density π.*

Proof. Observe that given $(x_2, \ldots, x_m)$,

$$\begin{aligned}
\mathcal{K}_1[\pi(\boldsymbol{x})](\boldsymbol{x}') &= \int_{x_1} \pi(x_1'|x_2, \ldots, x_m)\pi(x_1, \ldots, x_m)dx_1 \\
&= \int_{x_1} \frac{\pi(x_1', x_2, \ldots, x_m)\pi(x_1, \ldots, x_m)}{\pi(x_2, \ldots, x_m)}dx_1 \\
&= \pi(x_1', x_2, \ldots, x_m). \qquad (3.23)
\end{aligned}$$

Moreover, for a fixed $(x_{i+1}, \ldots, x_m)$, the assumption that $\mathcal{K}_{i-1} \ldots \mathcal{K}_1 = \pi(x_1', \ldots, x_{i-1}', x_i, \ldots, x_p)$ implies

$$\begin{aligned}
\mathcal{K}_i \ldots \mathcal{K}_1[\pi(\boldsymbol{x})](\boldsymbol{x}') &= \int_{x_i} \pi(x_i'|x_1', \ldots, x_{i-1}', x_{i+1}, \ldots, x_m)\pi(x_1', \ldots, x_{i-1}', x_i, \ldots, x_m)dx_i \\
&= \int_{x_i} \frac{\pi(x_1', \ldots, x_i', x_{i+1}, \ldots, x_m)\pi(x_1', \ldots, x_{i-1}', x_i, \ldots, x_m)}{\pi(x_1', \ldots, x_{i-1}', x_{i+1}, \ldots, x_m)}dx_i \\
&= \pi(x_1', x_2', \ldots, x_i', x_{i+1}, \ldots, x_m). \qquad (3.24)
\end{aligned}$$

By induction, $\mathcal{K}[\pi(\boldsymbol{x})](\boldsymbol{x}') = \mathcal{K}_m \ldots \mathcal{K}_1[\pi(\boldsymbol{x})](\boldsymbol{x}') = \pi(x_1', x_2', \ldots, x_m')$. Hence π is invariant. □

In fact, the argument above proves more than invariance with respect to π. The partial composition $\mathcal{K}_i \ldots \mathcal{K}_1$ is invariant with respect to $\pi(x_1, \ldots, x_i|x_{i+1}, \ldots, x_m)$. To see this, when $(x_{i+1}, \ldots, x_m)$ are given, then $\pi(\boldsymbol{x})/\pi(x_{i+1}, \ldots, x_m) = \pi(x_1, \ldots, x_i|x_{i+1}, \ldots, x_m)$ and

since each integration does not depend on $(x_{i+1}, \ldots, x_m)$, by (3.24)

$$\mathcal{K}_i \ldots \mathcal{K}_1[\pi(x_1, \ldots, x_i | x_{i+1}, \ldots, x_m)](\boldsymbol{x}') = \frac{\pi(x_1', \ldots, x_i', x_{i+1}, \ldots, x_m)}{\pi(x_{i+1}, \ldots, x_m)} = \pi(x_1', \ldots, x_i' | x_{i+1}, \ldots, x_m). \tag{3.25}$$

Viewing Gibbs sampling as composed conditional sub-transitions allows for the flexibility to design and analyze algorithms that modify each sub-transition step. That is, if an intermediate step in the Gibbs sampler is modified, say with $\widetilde{K}_i$, then in order to prove invariance, we need only show that $\widetilde{\mathcal{K}}_i \mathcal{K}_{i-1} \ldots \mathcal{K}_1$ is invariant with respect to $\pi(x_1, \ldots, x_i | x_{i+1}, \ldots, x_m)$. We state this result formally:

Corollary 3.1.3. *Suppose $\mathcal{K} = \mathcal{K}_m \ldots \mathcal{K}_1$ is the transition operator for Algorithm 1, and $\widetilde{\mathcal{K}}_i$ given $\boldsymbol{x}_{\widehat{i}}$ is an operator such that $\widetilde{\mathcal{K}}_i\big[\pi(x_i | \boldsymbol{x}_{\widehat{i}})\big] = \pi(x_1' | \boldsymbol{x}_{\widehat{i}})$, then $\mathcal{K}_m \ldots \widetilde{\mathcal{K}}_i \mathcal{K}_{i-1} \ldots \mathcal{K}$ is invariant with respect to π.*

Proof. Using (3.24) twice, we have

$$\begin{aligned}
\left(\mathcal{K}_m \ldots \widetilde{\mathcal{K}}_i\right) \mathcal{K}_{i-1} \ldots \mathcal{K}_1 \pi &= \mathcal{K}_m \ldots \widetilde{\mathcal{K}}_i \pi(x_1', \ldots, x_{i-1}', x_i, \ldots, x_m) \\
&= \mathcal{K}_m \ldots \widetilde{\mathcal{K}}_i \pi(x_i | x_1', \ldots, x_{i-1}', x_{i+1}, \ldots, x_m) \pi(x_1', \ldots, x_{i-1}' x_{i+1}, \ldots, x_m) \\
&= \mathcal{K}_m \ldots \mathcal{K}_{i+1} \pi(x_i' | x_1', \ldots, x_{i-1}', x_{i+1}, \ldots, x_m) \pi(x_1', \ldots, x_{i-1}' x_{i+1}, \ldots, x_m) \\
&= \mathcal{K}_m \ldots \mathcal{K}_{i+1} \pi(x_1', \ldots, , x_i', x_{i+1}, \ldots, x_m) \\
&= \pi(x_1', \ldots, x_m').
\end{aligned} \tag{3.26}$$

□

The previous result will be important for showing that embedding alternative simulation techniques (such as a Metropolis-Hastings step) will maintain invariance with respect to π.

In Section 4.1, we will define a discretization, $\boldsymbol{p}$, for the PSF p and a corresponding three component discrete posterior density $\pi(\boldsymbol{p}, \lambda, \delta | \boldsymbol{b})$. As will be seen in Chapter 4, the Gibbs

sampler will produce good Monte Carlo estimates for $\boldsymbol{p}$ and λ, a parameter determining the measurement noise level. However, the δ component of the chain exhibits poor convergence, and hence, the asymptotic application of the ergodic theorem for Markov chains for the joint density $\pi(\boldsymbol{p}, \lambda, \delta|\boldsymbol{b})$ is not available. In fact, [Agapiou et al., 2014] give theory showing that for general linear inverse problems, the infinite dimensional hierarchical Gibbs sampler for linear inverse problems with λ known and a Gaussian prior whose precision operator is a power of the negative Laplacian will always exhibit degenerate convergence in δ when the discrete representation of the unknown approaches the infinite dimensional representation. They presented an algorithm that 'marginalizes' the dependence of the unknown with δ. This process, known as partial collapse, can be carried out in general and is presented in [Van Dyk and Park, 2008], but must be done with care. In their paper, they showed various examples of improperly partially collapsed Gibbs samplers that lead to Markov chains that are no longer invariant with respect to the target density π. They also presented theory that when partial collapse is possible, it improves convergence; however, they did not give an explicit argument that shows that partial collapse maintains the invariant density π. We outline this process for the Gibbs sampler presented above, and show explicitly that it maintains π as an invariant density in the next section.

3.1.3 The partially collapsed Gibbs sampler

The partially collapsed Gibbs (PCG) sampler we present in this section is based on the work of [Van Dyk and Park, 2008; Van Dyk and Jiao, 2015], where they outlined how the algorithm arises naturally from trying to improve the convergence of the standard Gibbs sampler. In both [Van Dyk and Park, 2008; Van Dyk and Jiao, 2015], they highlight that partial collapse must be done prudently, else the resulting Markov chain may no longer be invariant with respect to π, and thus statistics derived from the chain will not converge to those of the distribution of interest. They even give some examples in the literature where partial collapse was implemented improperly and resulted in incorrectly estimated parameters. They carefully

outline methods for ensuring that the pitfalls of improper sampling are avoided, although they did not formally prove the invariance of the resulting Markov chains. In this section, we give rigorous novel arguments that show the Markov chains associated with proper partial collapse are indeed invariant.

Consider modifying Algorithm 1 by simulating an additional component

$$\widetilde{\boldsymbol{X}}^k = (X_1^k, \ldots, X_{m-1}^k, \widetilde{X}_m^k, X_m^k) \tag{3.27}$$

by taking a 'redundant' sample of $\widetilde{X}_p$ from the joint conditional $\pi(x_{m-1}, x_m | x_1^k, x_2^k, \ldots, x_{m-2}^k)$ at step m-1. The resulting algorithm is described in Algorithm 2.

Algorithm 2 m-Conditioned Gibbs sampler

Given $\widetilde{\boldsymbol{x}}^{k-1} = (x_1^{k-1}, \ldots, \widetilde{x}_m^{k-1}, x_m^{k-1})$, simulate

1. $X_1^k \sim \pi(x_1 | x_2^{k-1}, x_3^{k-1}, \ldots, x_m^{k-1})$
2. $X_2^k \sim \pi(x_2 | x_1^k, x_3^{k-1}, \ldots, x_m^{k-1})$

$\vdots$

m-1. $(X_{m-1}^k, \widetilde{X}_m^k) \sim \pi(x_{m-1}, x_m | x_1^k, x_2^k, \ldots, x_{m-2}^k)$

m. $X_m^k \sim \pi(x_m | x_1^k, x_2^k, \ldots, x_{m-1}^k)$

The corresponding transition operator to Algorithm 2 is

$$\widetilde{\mathcal{K}}\pi_0 = \widetilde{\mathcal{K}}_m \widetilde{\mathcal{K}}_{m-1} \ldots \widetilde{\mathcal{K}}_2 \widetilde{\mathcal{K}}_1 \pi_0 \tag{3.28}$$

where $\widetilde{\mathcal{K}}_{m-1}$ is integration with respect to $(x_{m-1}, \widetilde{x}_m)$ against the transition kernel

$$\widetilde{K}_{m-1}(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{x}}') \stackrel{\text{def}}{=} \widetilde{K}_{m-1}(\widetilde{\boldsymbol{x}}') \stackrel{\text{def}}{=} \pi(x'_{m-1}, \widetilde{x}'_m | x'_1, x'_2, \ldots, x'_{m-2}), \tag{3.29}$$

and every other $\widetilde{K}_i$ acts the same as in Algorithm 1 on $m+1$ component densities. Algorithm 2 produces a Markov chain with $m+1$ components by drawing $(X_{m-1}^k, \widetilde{X}_m^k)$ jointly at step m-1.

Note that the transition to the next state does not depend on previous values of $\widetilde{X}_m$. This lack of dependence is crucial for partially collapsing components out of the sampler, else the resulting transition kernel will *not* produce a Markov chain invariant with respect to π.

Proposition 3.1.4. *The Markov chain associated with the transition kernel corresponding to Algorithm 2 is invariant with respect to* $\widetilde{\pi}(\widetilde{\boldsymbol{x}}) \stackrel{\text{def}}{=} \pi(\boldsymbol{x})\pi(\widetilde{x_m}|\boldsymbol{x}_{\widehat{m}})$.

Proof. Denote the transition operator associated to Algorithm 2 as $\widetilde{\mathcal{K}}$, then

$$\begin{aligned}
&\widetilde{\mathcal{K}}\Big[\pi(\boldsymbol{x})\pi(\widetilde{x}_m|\boldsymbol{x})\Big](\widetilde{\boldsymbol{x}}') \\
&\quad= \mathcal{K}_m\widetilde{\mathcal{K}}_{m-1}\mathcal{K}_{m-2}\ldots\mathcal{K}_1\big[\pi(\boldsymbol{x})\pi(\widetilde{x}_m|\boldsymbol{x})\big](\widetilde{\boldsymbol{x}}') \\
&\quad= \int_{x_m} \pi(x'_m|\boldsymbol{x}'_{\widehat{m}}) \iint_{\widetilde{x}_m,x_{m-1}} \widetilde{K}_{m-1}(\widetilde{\boldsymbol{x}}') \int_{x_{m-2},\ldots,x_1} \cdots \int \ldots \pi(\boldsymbol{x})\pi(\widetilde{x}_m|\boldsymbol{x})dx_1\ldots d\widetilde{x}_m dx_m \\
&\quad= \int_{x_m} \pi(x'_m|\boldsymbol{x}'_{\widehat{m}}) \int_{x_{m-1}} \widetilde{K}_{m-1}(\widetilde{\boldsymbol{x}}') \int_{x_{m-2},\ldots,x_1} \cdots \int \ldots \pi(\boldsymbol{x})dx_1\ldots dx_m \qquad (3.30)
\end{aligned}$$

where we used Fubini's theorem to integrate first in $\widetilde{x}_m$ for which each kernel K_i does not depend. Since $\int \pi(\widetilde{x}_m|\boldsymbol{x})d\widetilde{x}_m = 1$, and each of the inner $m-2$ integrations express the action of the first $m-2$ steps of the standard Gibbs sampler, continuing from (3.30) and using Corollary 3.1.3, we have

$$\begin{aligned}
\widetilde{\mathcal{K}}\Big[\pi(\boldsymbol{x})\pi(\widetilde{x}_m|\boldsymbol{x})\Big](\widetilde{\boldsymbol{x}}') &= \int_{x_m} \pi(x'_m|\boldsymbol{x}'_{\widehat{m}}) \int_{x_{m-1}} \widetilde{K}_{m-1}(\widetilde{\boldsymbol{x}}') \cdot \big(\mathcal{K}_{m-2},\ldots,\mathcal{K}_1[\pi(\boldsymbol{x})](\boldsymbol{x}')\big)\, dx_{m-1}dx_m \\
&= \int_{x_m} \pi(x'_m|\boldsymbol{x}'_{\widehat{m}}) \int_{x_{m-1}} \pi(x'_{m-1},\widetilde{x}'_m|x'_1,x'_2,\ldots,x'_{m-2})\cdot \\
&\qquad \ldots\pi(x'_1,\ldots,x'_{m-2},x_{m-1},x_m)dx_{m-1}dx_m \qquad (3.31) \\
&= \pi(x'_m|\boldsymbol{x}'_{\widehat{m}})\pi(x'_{m-1},\widetilde{x}'_m|x'_1,x'_2,\ldots,x'_{m-2})\cdot\pi(x'_1,\ldots,x'_{m-2}) \\
&= \frac{\pi(\boldsymbol{x}')\pi(x'_1,x'_2,\ldots,x'_{m-1},\widetilde{x}'_m)}{\pi(\boldsymbol{x}'_{\widehat{m}})} \\
&= \pi(\boldsymbol{x}')\pi(\widetilde{x}'_m|\boldsymbol{x}'_{\widehat{m}}). \qquad (3.32)
\end{aligned}$$

□

Note that it is essential that each sub-transition does not depend on $\widetilde{x}_m$, else the initial integration in $\widetilde{x}_m$ would involve products of kernels depending on $\widetilde{x}_m$ with $\pi(\widetilde{x}_m|\boldsymbol{x}_{\widehat{m}})$. In the language of [Van Dyk and Park, 2008], this would correspond to subsequently sampling the auxiliary variable, which they've shown can result in the loss of invariance.

Also, the placement of the conditioned variable at the last step is crucial for the argument to work. It can be shown that for a kernel with a different placement of the conditioned variable, a density of the form $\pi(\boldsymbol{x})q(\widetilde{\boldsymbol{x}})$ with $\int_{\widetilde{x}_i} q(\widetilde{\boldsymbol{x}})d\widetilde{\boldsymbol{x}} = 1$ will *not* be invariant. In practice, this has no practical effect on an implementation that cyclically permutes the steps in Algorithm 2, since it can be viewed as a Markov chain with the same transition kernel, only that it has a different initial distribution, and that at the last step, the transition kernel has partially completed.

In some sense, this algorithm is artificial, as we do not need to sample the auxiliary variable $\widetilde{X}_m$. Moreover, if we integrate the invariance condition $\widetilde{\mathcal{K}}\widetilde{\pi} = \widetilde{\pi}$ in $\widetilde{x}'_m$, then

$$\int_{\widetilde{x}'_m} \widetilde{\mathcal{K}}[\widetilde{\pi}(\widetilde{\boldsymbol{x}})](\widetilde{\boldsymbol{x}}') = \int_{\widetilde{x}'_m} \widetilde{\mathcal{K}}[\pi(\boldsymbol{x})\pi(\widetilde{x}_m|\boldsymbol{x}_{\widehat{m}})](\widetilde{\boldsymbol{x}})d\widetilde{x}'_m = \pi(\boldsymbol{x}') \int_{\widetilde{x}'_m} \pi(x'_m|\boldsymbol{x}'_{\widehat{m}})dx'_m = \pi(\boldsymbol{x}'). \tag{3.33}$$

This results in a transition kernel whose action is the same as the standard Gibbs transition kernel, except at step $m-1$, the kernel for the transition is

$$\overline{K}_{m-1}(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{x}}') \stackrel{\text{def}}{=} \int_{x'_m} \pi(x'_{m-1}, \widetilde{x}'_m|x'_1, x'_2, \ldots, x'_{m-2}) = \pi(x'_{m-1}|x'_1, x'_2, \ldots, x'_{m-2}). \tag{3.34}$$

Note that this is exactly the action given in (3.31), and by (3.33), the corresponding Markov chain is invariant with respect to π, and Algorithm 3 simulates this chain.

The effect of this process is that we have removed conditioning of $X_m^{k-1} = x_m^{k-1}$ from the

Algorithm 3 m-Partially Collapsed Gibbs sampler

Given $\widetilde{\boldsymbol{x}}^{k-1} = (x_1^{k-1}, \ldots, \widetilde{x}_m^{k-1}, x_m^{k-1})$, simulate

1. $X_1^k \sim \pi(x_1 | x_2^{k-1}, x_3^{k-1}, \ldots, x_m^{k-1})$
2. $X_2^k \sim \pi(x_2 | x_1^k, x_3^{k-1}, \ldots, x_m^{k-1})$

$\vdots$

m-1. $X_{m-1}^k \sim \pi(x_{m-1} | x_1^k, x_2^k, \ldots, x_{m-2}^k)$

m. $X_m^k \sim \pi(x_m | x_1^k, x_2^k, \ldots, x_{m-1}^k)$

simulation of X_{m-1}^k. Note that the first $m-2$ steps of the algorithm can be permuted with the appropriate re-labeling with respect to k without changing the transition kernel. We can generalize the partial collapse process by removing the conditioning on either X_{m-1} or X_m on X_{m-2}. Without loss of generality, X_{m-2} can be chosen from $X_1, \ldots, X_{m-2}$ by permuting and relabeling. Hence, X_m can be partially collapsed out of any number of proceeding variables, and subsequently, X_{m-1}, etc.

In practice, one starts with the standard Gibbs sampler, and observes convergence of each component. If a component exhibits poor convergence (see Section 3.2), see if any conditioned variables can be partially collapsed. This choice is likely not obvious, unless guided by the specific situation (as is the case for the hierarchical Gibbs sampler for sampling δ). If it is possible to sample the density with one of the conditioned variables collapsed out, re-order the sampler so that the collapsed component is last and each of the poorly converging variables directly precedes it. The theory presented in [Van Dyk and Park, 2008] guarantees that the convergence of $(\boldsymbol{X}^k)$ will be improved. If some components still exhibit poor convergence, continue by removing the conditioning of one of the previous $m-1$ variables. See [Van Dyk and Park, 2008] for examples and a further discussion of the general process of partially collapsing variables.

There is one last modification to the transition kernel that will be required. In many cases, as will be the case for PSF reconstruction, a simulation from $\pi(x_{m-1} | x_1, \ldots, x_{m-2})$ may not

be directly available. In the standard Gibbs case, when a full conditional density is difficult to simulate, a compromise suggested first by [Müller, 1992] and outlined in [Robert and Casella, 2013] is the so-called 'Metropolis-within-Gibbs' method. The idea is to replace a direct sample of the conditional density with a Metropolis-Hastings transition. In the next section, we give a brief overview of the random walk Metropolis-Hastings method, and show that directly substituting a Metropolis-Hastings transition into the m-partially collapsed Gibbs sampler remains invariant with respect to π.

3.1.4 Metropolis-Hastings within partially collapsed Gibbs

The Metropolis-Hastings algorithm [Metropolis et al., 1953] has been studied extensively as an MCMC method, and over the last half-century, has been generalized and adapted to encompass a large class of MCMC algorithms for simulating samples for a large class of problems. In fact, Gibbs sampling can be viewed as successive Metropolis-Hastings transitions [Robert and Casella, 2013]. We will focus on Metropolis-Hastings algorithms with symmetric proposals and how they can be incorporated into the partially collapsed Gibbs sampler. Again, see one of the books [Calvetti and Somersalo, 2007; Liu, 2008; Robert and Casella, 2013] and references there for a complete description of the Metropolis-Hastings algorithm.

Consider the following algorithm for simulating a transition for a univariate Markov chain $(X^1, X^2 \dots)$:

Algorithm 4 Reversible Metropolis-Hastings

Given $X^k = x^k$, and proposal density such that $\rho(y|x) = \rho(x|y)$.

1. Simulate $Y^k \sim \rho(y|x^k)$

2. Set

$$X^{k+1} = \begin{cases} Y^k & \text{with probability } \alpha(x^k, Y^k) \\ x^k & \text{with probability } 1 - \alpha(x^k, Y^k) \end{cases}$$

where $\alpha(x, y) = \min\left\{1, \dfrac{\pi(y)}{\pi(x)}\right\}$.

The simulation $Y^k \sim \rho(y|x^k)$ is called the *proposal* transition. The idea of the Metropolis-Hastings method is: first generate a 'proposal' from a given transition operator, $\rho(y|x)$, that describes the probability of transitioning to y, given your current state is x. Then, if your guess improves how likely the transition is from the desired distribution π, then move there, otherwise stay put [Calvetti and Somersalo, 2007]. At first glance, this algorithm might not seem useful since it requires a computation involving π, which may not be completely known, but that it appears as a ratio is what makes the method useful – we need only know π up to a constant of proportionality since it cancels in the ratio.

To see formally that Algorithm 4 defines an invariant Markov chain for π, we will need a general result from Markov chain theory known as *detailed balance*.

Theorem 3.1.5. *Suppose that a Markov chain with a transition kernel K satisfies the* detailed balance condition

$$K(x, x')\pi(x) = K(x', x)\pi(x'). \tag{3.35}$$

Then, the corresponding Markov chain is invariant with respect to π.

Proof. The corresponding transition operator has

$$\mathcal{K}[\pi](x') = \int K(x, x')\pi(x)dx = \int K(x', x)\pi(x')dx = \pi(x') \tag{3.36}$$

since $K(x', \cdot)$ is a probability density. □

The Metropolis-Hastings kernel is designed to satisfy detailed balance and [Calvetti and Somersalo, 2007] present the development of the Metropolis-Hastings algorithm with that perspective. We summarize that discussion to give an explicit description of the transition kernel corresponding to Algorithm 4 and show that it satisfies the detailed balance condition.

Proposition 3.1.6. *The Markov chain generated by Algorithm 4 has a transition kernel that satisfies the detailed balance condition for π; hence it is invariant with respect to π.*

Proof. Let $X_k = x_k$ be given and U be a binomial random variable such that $U = 1$ if the proposal is accepted and $U = 0$ otherwise. Then, for any event A

$$\begin{aligned}
\mathbb{P}\left(X^{k+1} \in A | X^k = x^k\right) &= \mathbb{P}\left(X^{k+1} \in A \text{ and } U = 1 | X^k = x^k\right) \\
&\quad + \mathbb{P}\left(X^{k+1} \in A \text{ and } U = 0 | X^k = x^k\right) \\
&= \mathbb{P}\left(Y^k \in A \text{ and } U = 1 | X^k = x^k\right) \\
&\quad + \mathbb{P}\left(Y^k \in A \text{ and } U = 0 | X^k = x^k\right).
\end{aligned} \tag{3.37}$$

The mixed continuous/discrete density for $(Y^k, U | X^k = x^k)$ satisfies $\pi(y, u|x^k) = \pi(u|y, x^k)\rho(y|x^k)$ by the definition of conditional density. Moreover, $\pi(u = 1|y, x^k) = \alpha(x^k, y)$ and

$$\pi(u = 0|x^k) = \int \pi(u = 0, y'|x^k)dy' = \int \pi(u = 0|y', x^k)\pi(y'|x^k)dy' = \int (1-\alpha(x^k, y'))\rho(|y'-x^k|)dy'.$$

Continuing from (3.37),

$$\mathbb{P}\left(X^{k+1} \in A | X^k = x^k\right) = \int_A \alpha(x^k, y)\rho(y|x^k)dy + I_A(x^k)\int (1 - \alpha(x^k, y'))\rho(y|x^k))dy' \quad (3.38)$$

where I_A denotes the indicator function for the set A. Note that $I_A(x) = \int_A \delta_x(y)dy$, where δ_x is the Dirac probability density, so the transition kernel for Algorithm 4 is

$$K(x, y) = \alpha(x, y)\rho(y|x) + \delta_x(y)\left(1 - \int \alpha(x, y')\rho(y'|x)dy'\right). \quad (3.39)$$

In order to show that $K(x, y)$ satisfies the detailed balance equation, it suffices to show it for each term in (3.39). If $\pi(y) \geq \pi(x)$ then $\alpha(x, y) = 1$ and $\alpha(y, x) = \pi(x)/\pi(y)$ implies

$$\alpha(x, y)\rho(y|x)\pi(x) = \rho(x|y)\pi(x) = \frac{\pi(x)}{\pi(y)}\rho(x|y)\pi(y) = \alpha(y, x)\rho(x|y)\pi(y). \quad (3.40)$$

Moreover, for any integrable function f, we have (in the distributional sense)

$$f(x)\int_A \delta_x(y)\pi(y)dy = f(x)I_A(x)\pi(x) = \pi(x)\int_A \delta_x(y)f(y)dy \quad (3.41)$$

for all events A. Thus taking $f(x) = 1 - \int \alpha(x, y')\rho(y'|x)dy'$ proves that $K(x, y)$ satisfies the detail balance condition, and hence the Markov chain for Algorithm 4 is invariant with respect to π. □

A Metropolis-Hastings sub-transition can be used to sample any component of the standard Gibbs sampler, since we were able to decompose it into conditionally invariant sub-transitions as in Corollary 3.1.3. For the collapsed sampler, we consider the last two steps together as a sub Metropolis-Hastings transition, where the acceptance depends only on the $m - 1$ component.

Algorithm 5 Metropolis Hastings within m-Partially Collapsed Gibbs sampler

Given $\widetilde{\boldsymbol{x}}^{k-1} = (x_1^{k-1}, \ldots, \widetilde{x}_m^{k-1}, x_m^{k-1})$, simulate

1. $X_1^k \sim \pi(x_1 | x_2^{k-1}, x_3^{k-1}, \ldots, x_m^{k-1})$
2. $X_2^k \sim \pi(x_2 | x_1^k, x_3^{k-1}, \ldots, x_m^{k-1})$

$\vdots$

m-1. Simulate X_{m-1}^k from Algorithm 4 for $\pi(x_{m-1} | x_1^k, x_2^k, \ldots, x_{m-2}^k)$

m. $X_m^k \sim \pi(x_m | x_1^k, x_2^k, \ldots, x_{m-1}^k)$

One implementation question remains concerning whether to iterate step m-1. to obtain 'better' simulations from $\pi(x_{m-1} | x_1^k, x_2^k, \ldots, x_{m-2}^k)$. That is, we can insert any number of Metropolis steps before step m., and the resulting sampler will still be invariant by Corollary 3.1.3. That is, by Proposition 3.1.6 each of n_{mh} Metropolis draws resulting from the sub-transition in Algorithm 4 at state m-1. is invariant with respect to $\pi(x'_{m-1} | x_1, x_2^k, \ldots, x_{m-2}^k)$, hence their composition is also invariant, thus, Corollary 3.1.3 implies the invariance of the full transition with respect to π of Algorithm 5. We state this formally in the following theorem.

Theorem 3.1.7. *The transition kernel associated with Algorithm 5 where at step $m-1$., n_{mh} iterations of Algorithm 4 are taken, generates a Markov chain that is invariant with respect to π.*

When implemented in standard Gibbs sampling, [Robert and Casella, 2013] recommend only one simulation, but in [Van Dyk and Jiao, 2015], they recommend that iterating the Metropolis step may improve the convergence rate. As is the case for PSF estimation, generating the proposal for Algorithm 4 may be computationally expensive, and the improvement in convergence may not be worth the computational expense since the less expensive but slower to converge scheme can be run for longer. These issues are problem dependent, and in

Section 3.2, we will develop tools to address them explicitly.

3.2 Evaluating Convergence

In these last sections, we briefly address estimators for evaluating the convergence of the MCMC algorithms. As has been mentioned, convergence can be addressed theoretically by direct analysis as in [Agapiou et al., 2014], or by analyzing the spectrum of an operator associated with the transition operator as is done in [Agapiou et al., 2014; Van Dyk and Park, 2008]. We take an empirical approach, that estimates convergence based on real and simulated data.

In this section, we give a brief overview of two statistical estimators that can be used to verify this convergence given a realization of an MCMC algorithm. Both estimators address issues that inform how long to run the MCMC algorithm in order to effectively analyze the chain as a robust sample for the PSF posterior.

The first issue is concerned with how close the Markov chain is to the target invariant density. The realizations from the initial density of the Markov chain may correspond to low probability events of the target distribution. This is acutely the case for the prior parameter δ in PSF posterior estimation, as its meaning in the model is quite subtle. Nevertheless, the ergodic theorem guarantees that a valid MCMC algorithm will produce realizations from densities that converge to the target. The MCMC simulations that occur from the beginning of the chain until empirical convergence is observed are called the *burn-in* portion of the Markov chain, and we will briefly overview a statistical test in the next section on how to estimate it.

The second practical convergence issue is related to the correlation of subsequent steps of the MCMC algorithm. We employ a method from time-series analysis that estimates

the correlation of a realized Markov chain from an MCMC algorithm. This will result in a parametric measurement for 'how far' the simulated samples are from an ideal independent sample and how long the chain must be run in order to obtain, in some sense, an equivalent estimator to one derived from independently sampled data.

3.2.1 Estimating the burn-in

One method for estimating the burn-in stage of the chain is to visually inspect the realizations of the MCMC algorithm and identify the portion of the chain that appears to settle over the support of the invariant density. This is somewhat subjective, and an alternative statistically motivated approach, initially suggested by [Geweke, 1991], uses the *convergence diagnostic test* to evaluate the test hypothesis that the joint mean value of the early portion of the Markov chain is equal to the joint mean value of the latter.

Formally, for a given partial Markov chain $\{X^1, \ldots, X^N\}$, let N_m denote the mth percentile of N, μ_m to be the mean of $\{X^1, \ldots, X^{N_m}\}$ and $\mu_{m'}$ the mean of $\{X^{N_{m'}+1}, \ldots, X^N\}$. Following [Geweke, 1991], we choose the 10th and 50th percentiles. Estimators for μ_{10} and $\mu_{50'}$ are

$$\overline{X}_{10} = \frac{1}{N_{10}} \sum_{k=1}^{N_{10}} X^k, \quad \text{and} \quad \overline{X}_{50'} = \frac{1}{N - N_{50'}} \sum_{k=N_{50}+1}^{N} X^k. \tag{3.42}$$

For the test $H_0 : \mu_{10} = \mu_{50'}$, [Geweke, 1991] shows the corresponding convergence diagnostic test statistic satisfies

$$R_{\text{Geweke}} \stackrel{\text{def}}{=} \frac{\overline{X}_{10} - \overline{X}_{50'}}{\sqrt{\widehat{S}_{10}(0)/N_{10} + \widehat{S}_{50'}(0)/N_{50}}} \xrightarrow{d} \mathcal{N}(0,1), \quad \text{as} \quad N \to \infty, \tag{3.43}$$

where $\widehat{S}_{10}(0)$ and $\widehat{S}_{50'}(0)$ denote consistent spectral density estimates for the variances of $\{X^1, \ldots, X^{N_{10}}\}$ and $\{X^{N_{51}}, \ldots, X^N\}$, respectively. These can be estimated via a periodogram estimator, and in our results, we use a Danielle window of width $2\pi/(0.3p^{1/2})$ as recommended

by [Geweke, 1991].

3.2.2 Autocorrelation and essential sample size

When the burn-in has been identified, the later portion of the chain can be empirically assumed to be identically distributed by the invariant density via the ergodic theorem. However, the MCMC samples are not independent; hence standard sampling theory does not apply. The notion of autocorrelation from time-series analysis provides a tool for controlling for this correlation. The idea is to estimate how many steps are required in the Markov process to 'forget' the state where you came from; specifically, to be empirically uncorrelated. This is referred to as the *integrated autocorrelation time*. To develop this notion formally, we summarize the arguments in [Sokal, 1997]. Suppose $\{X^1, X^2, \dots\}$ is a correlated, identically distributed stochastic process with individual variance σ^2. Then for the estimator $\overline{X}_N = \frac{1}{N}\sum_{k=1}^{N} X^i$, the *Monte Carlo error* is

$$\begin{aligned}
\mathrm{Var}(\overline{X}_N) &= \frac{1}{N^2}\sum_{k=1}^{N}\mathrm{Var}(X^k) + \frac{1}{N^2}\sum_{k\neq l}^{N}\mathrm{Cov}(X^l, X^k) \\
&= \frac{1}{N^2}\left(N\sigma^2 + 2N\sum_{k=1}^{N-1}\left(1-\frac{k}{N}\right)\mathrm{Cov}(X^1, X^{1+k})\right) \\
&= \frac{\sigma^2}{N}\left(1 + 2\sum_{k=1}^{N-1}\left(1-\frac{k}{N}\right)\frac{\mathrm{Cov}(X^1, X^{1+k})}{\sigma^2}\right). \qquad (3.44)
\end{aligned}$$

Note that we've divided the Monte Carlo error into a contribution from the inherent variance of (X^k) and the contribution of its correlation at step l with all other steps in the chain. For the full stochastic sequence $\{X^1, X^2, \dots\}$, define for $k \in \mathbb{Z}$

$$\rho(k) \stackrel{\mathrm{def}}{=} \frac{\mathrm{Cov}(X^1, X^{|k|})}{\sigma^2}, \qquad (3.45)$$

which is referred to as the normalized autocorrelation function (ACF) in time-series analysis. Hence, the Monte Carlo error for $\{X^1, \ldots, X^N\}$ when N is large, is approximately

$$\mathrm{Var}(\overline{X}) \approx \frac{\sigma^2}{N} \sum_{k=-\infty}^{\infty} \rho(k). \tag{3.46}$$

When (X^k) are independent, σ^2/N is the variance of the estimator $\overline{X}_N$. The approximation in (3.46) shows that, asymptotically, the Monte Carlo error is scaled by the factor

$$\tau_{\mathrm{int}} \stackrel{\mathrm{def}}{=} \sum_{k=-\infty}^{\infty} \rho(k). \tag{3.47}$$

The parameter τ_{int} is referred to as the *integrated auto corresponding time*. So, for a given N, an independent sample with equivalent sample variance has

$$N_{\mathrm{ESS}} \stackrel{\mathrm{def}}{=} N/\tau_{\mathrm{int}} \tag{3.48}$$

samples. We refer to this quantity as the *essential sample size* (ESS), and we think of it as the number of effectively independent samples, and note

$$\frac{\sigma^2}{N}\tau_{\mathrm{int}} = \frac{\sigma^2}{N_{\mathrm{ESS}}}. \tag{3.49}$$

To estimate these parameters, [Sokal, 1997] gives the following unbiased estimator for the normalized autocorrelation function,

$$\widehat{\tau}_{\mathrm{int}} = \sum_{k=-\overline{N}}^{\overline{N}} \widehat{\rho}(k), \tag{3.50}$$

where $\overline{N} < N - 1$ is some window length, and $\widehat{\rho}(k)$ is the empirical normalized covariance

estimator over that interval. That is,

$$\widehat{\rho}(k) \stackrel{\text{def}}{=} \widehat{C}(k)/\widehat{C}(0), \quad \text{where} \quad \widehat{C}(k) = \frac{1}{N-k}\sum_{i=1}^{N-k}(X_i - \overline{X_N})(X_{i+k} - \overline{X_k}). \tag{3.51}$$

The choice suggested by Sokal [1997] for the window size is the smallest integer such that $\overline{N} \geq 3\widehat{\tau}_{\text{int}}$. Finally the ESS is estimated as

$$\widehat{N}_{\text{ESS}} = N/\widehat{\tau}_{\text{int}}. \tag{3.52}$$

Chapter 4

Discrete Bayesian Posterior PSF estimation

We return to the problem of PSF estimation, and apply the theory and methods outlined in the previous chapters for carrying it out on a computer. At the end of Chapter 2, we established the theory for defining the inverse problem in an infinite dimensional Hilbert space. In both the variational and the infinite dimensional Bayesian formulation, the development led to defining the functional $\Phi : \mathscr{P} \times \mathcal{G}(\mathscr{P}) \to \mathbb{R}$

$$\Phi(p; b, \lambda, \delta) = \frac{1}{2}\left(\lambda \|\mathcal{G}p - b\|^2_{L^2(\mathbb{R})} + \delta\,(p, \mathcal{L}^n p)_{\mathscr{P}}\right), \tag{4.1}$$

where $\mathcal{G} : \mathscr{P} \to L^2(\mathbb{R})$ is the operator that takes a radial profile, p, to a line-out of a blurred image of an edge, b, with $\mathcal{L}$ the induced negative radial Laplacian. We also showed that $\mathcal{L}^n$ was trace class for all $n \geq 2$ in $\mathscr{P}$. In **??**, we modeled the problem with the infinite dimensional Bayesian perspective which results in a posterior Gaussian random variable taking values in $\mathscr{P}$, determined indirectly by the functional in (4.1).

Of course, to carry out numerical estimation, the data and the estimate for the PSF must be represented by a finite set of numbers on a computer. In the framework of [Stuart, 2010], one would design an algorithm that samples the infinite dimensional posterior. This approach was undertaken in [Agapiou et al., 2014] for linear inverse problems with powers of a Laplacian precision operator and a hierarchical gamma prior for δ, with a fixed value for λ. They analyzed an infinite dimensional Gibbs sampling algorithm, and showed that it had deficiencies in sampling δ that exacerbate as the discretization converges. They then introduced two modifications to the algorithm that alleviate this issue. Although their analysis is not directly applicable to PSF reconstruction (our prior is the negative radial Laplacian), it is closely related and we take cues from their work to design the algorithm for exploring a discrete approximation to the infinite dimensional posterior density. Also, by discretizing at this stage, we will be able to develop an algorithm that allows for the noise precision λ to be estimated. We follow the general development outlined in [Bardsley, 2012], which has been adopted successfully in many other applications of linear inverse problems related to imaging [Howard et al., 2016; Bardsley and Luttman, 2016; Fowler et al., 2016; Bardsley and Luttman, 2015; Bardsley et al., 2013]. We then derive all of the necessary probability densities for carrying out the algorithms in Section 3.1. The discrete representations correspond to numerical discretizations of the linear operators defined in Chapter 2. Since each operator is linear, the corresponding numerical approximations will also be linear and can be effectively implemented with an appropriate matrix multiplication. We then develop the discrete probability spaces associated with the matrix-operators defined in Section 4.1, which will serve as our discrete approximation of the infinite dimensional posterior defined in **??**. In this development, we will add prior assumptions for the parameters λ and δ, forming a hierarchical Bayesian model. From there, the discrete posterior distribution can be expressed in terms of conditional distributions in such a way so that Markov Chain Monte Carlo sampling techniques (e.g., Gibbs sampling) can be applied to provide estimates and quantification of uncertainty.

4.1 From the continuum to the discrete

Transitioning from the model on the continuum to a discrete representation is a delicate process for which error is introduced at many levels. For example, we do not even have full access to all of $\mathbb{R}$, since a computer must represent a real number with a floating-point approximation corresponding to a binary integer from a finite set (although, this error is not addressed in this work). This approximation provides a good analogy for how we will use smooth functions as approximations for p and b. The formal notions of p and b are as functionals that act on compactly supported smooth functions, which have many levels of abstraction beyond a point-wise definition, and we will attempt to briefly address the approximation at each of these levels.

4.1.1 Discretization methods

Our primary tools for discretization will be finite-differencing for the regularizing differential operator $\mathcal{L}$ and numerical quadrature for the integral operator $\mathcal{G}$ and integral inner products associated with $\mathscr{P}$. Both methods ensure point-wise convergence to known evaluations on a discrete grid. The error analysis associated with these methods is based on Taylor expansions of a function that is at least twice differentiable at each point in the interior of their domain and that the second derivatives are uniformly bounded in order to obtain error $O(h)$, where h is the width between grid points. This analysis is not directly applicable for $p \in \mathscr{P}$ since point-wise convergence is not applicable to distributions. In fact, any element with discrete support in $\mathscr{P}$ is equivalent to 0. Yet, we justify our use of quadrature methods by recalling from Chapter 2 that smooth functions are dense in $\mathscr{P}$. Since the proof of that theorem is constructive, in theory, one could use it to construct a smooth approximation, then apply quadrature on the approximation in order to explicitly control the error. Such an analysis is beyond the scope of this work, and we discretize each operator assuming that they act on

smooth functions and that the data and computational grids are sufficiently fine so that second order methods introduce errors at a scale so that the aggregate error of these approximations is negligible.

We mention that there are other methods that are theoretically more appealing which use a truncation of orthonormal bases, sometimes referred to as Gelerkin methods. It can be shown that a class of Bessel functions are an orthonormal set of eigenvectors for the negative radial Laplacian where the eigenvalues are the first positive root of the corresponding Bessel function; hence elements of $\mathscr{P}$ can be easily represented in that basis. However, proceeding with this method requires estimation of roots, as well as evaluation of the forward operator on Bessel functions which are both analytically difficult.

We assume that the domain of b is scaled so that data are collected on equally spaced points in $x_i \in [-1, 1]$, with $x_i = \frac{i}{N}$ for $-N \leq i \leq N$. Let $\boldsymbol{b} \in \mathbb{R}^{2N+1}$ with entries $b_i \stackrel{\text{def}}{=} b(x_i)$ and $h \stackrel{\text{def}}{=} \frac{1}{N}1$ This enforces an assumption that the data has an odd number of points with $b(0)$ corresponding to the Nth element of $\boldsymbol{b}$. Note that the point-symmetry of the operator implies that for N point estimates of p, a full line-out of data will have $2N + 1$ points (the extra estimate is for $b(0)$). In what follows, we define $\operatorname{acos}(t)$ on all of $\mathbb{R}$ by taking the convention that $\operatorname{acos}(t) = 0$ if $|t| > 1$. Figure 1.3 is useful for visualizing the following arguments.

Recall in (1.22), the integral kernel for $\mathcal{G}$ was

$$g(x, r) = \begin{cases} 0 & x < -r \\ 2(\pi - \operatorname{acos}(x/r)) & |x| \leq r \\ 2\pi & x > r \end{cases} . \tag{4.2}$$

For a fixed $x_i < 0$, we have

$$[\mathcal{G}p](x_i) = \int_0^{-x_i} p(r) 2(\pi - \operatorname{acos}(x_i/r)) r \, dr. \tag{4.3}$$

As in [Bardsley, 2012], we discretize the integral using midpoint quadrature which guarantees a second-order integration method. Because the upper bound in (4.3) depends on x_i, $\{r_j\}$ are placed at midpoints of $\{|x_i|\}$, hence, $r_j \stackrel{\text{def}}{=} j - \frac{h}{2}$ for $j = 1, \dots N$.

When $x_i \leq 0$, then $i \leq 0$, and using $\text{acos}(t) = 0$ for $t < -1$,

$$\begin{aligned}[\mathcal{G}p](x_i) &\approx \sum_{j=1}^{|i|} p(r_j) 2(\pi - \text{acos}(x_i/r_j)) r_j h \\ &= \sum_{j=1}^{N} p(r_j) 2(\pi - \text{acos}(x_i/r_j)) r_j h. \end{aligned} \tag{4.4}$$

When $x_i > 0$, then $i \geq 0$ and using $\text{acos}(t) = 0$ for $t > 1$,

$$\begin{aligned}[\mathcal{G}p](x_i) &= 2\pi \int_0^{x_i} p(r)\, r dr + \int_{x_i}^{\infty} p(r) 2(\pi - \text{acos}(x_i/r))\, r dr \\ &= \int_0^{x_i} p(r) 2(\pi - \text{acos}(x_i/r))\, r dr + \int_{x_i}^{\infty} p(r) 2(\pi - \text{acos}(x_i/r))\, dr \\ &\approx \sum_{j=1}^{N} p(r_j) 2(\pi - \text{acos}(x_i/r_j)) r_j h. \end{aligned} \tag{4.5}$$

Now, let $\boldsymbol{p}$ be the $N \times 1$ column vector with entries $\boldsymbol{p}_j = p(r_j)$, then using (4.4) and (4.5), if we define the matrix $\boldsymbol{G}$ with entries $\boldsymbol{G}_{ij} \stackrel{\text{def}}{=} 2(\pi - \text{acos}(x_i/r_j)) r_j h$, the quadrature approximation can be expressed by the matrix-vector multiplication $\boldsymbol{Gp}$. Finally, we use $\|f\|_{L^2(\mathbb{R})} \approx h \|\boldsymbol{f}\|_{\mathbb{R}^{2N+1}}$ to approximate the L^2 norm of f, where elements of $\boldsymbol{f}$ are point-wise evaluations of the continuous approximation to f. Combining these approximations, we have for the first term in (4.1),

$$\lambda \|b - \mathcal{G}p\|_{L^2} \approx \lambda h \|\boldsymbol{b} - \boldsymbol{Gp}\|_{\mathbb{R}^{2N+1}}. \tag{4.6}$$

The negative radial Laplacian $\mathcal{L} : \mathscr{P} \to \mathscr{P}$ operates on continuous functions by

$$[\mathcal{L}p](r) = -r^{-1}\frac{d}{dr}\left(r\,\frac{d}{dr}p(r)\right). \tag{4.7}$$

Observe that

$$\begin{aligned}\langle p, \mathcal{L}^2 p\rangle_{\mathscr{H}_1} &= 2\pi\int_0^\infty p(r)[\mathcal{L}^2 p](r)\,r dr \\ &= 2\pi\int_0^\infty p(r)\left[\frac{d}{dr}\left(r\,\frac{d}{dr}\right)\right]^2 p(r)\,r^{-1}dr.\end{aligned} \tag{4.8}$$

The discretization of (4.8) will occur in two steps. We will use quadrature to estimate the integral in (4.8), and use finite differencing to estimate the differential operator $\frac{d}{dr}r\frac{d}{dr}$. We then square that estimate, and combine it with the quadrature estimate of the integral.

We use the differencing scheme outlined in [Morton and Mayers, 2005]. Let $r_{j\pm1/2} \stackrel{\text{def}}{=} r_j \pm \frac{h}{2}$, then

$$\left[\frac{d}{dr}r\frac{d}{dr}p\right]_{r_j} \approx \frac{1}{h}\left(\left[r\frac{d}{dr}p\right]_{r_{j-1/2}} + \left[r\frac{d}{dr}p\right]_{r_{j+1/2}}\right). \tag{4.9}$$

The center difference approximation of the first term is

$$\left[r\frac{d}{dr}p\right]_{r_{j-1/2}} \approx r_{j-1/2}\frac{p_{j-1}-p_j}{h} \tag{4.10}$$

and of the second term is

$$\left[r\frac{d}{dr}p\right]_{r_{j+1/2}} \approx r_{j+1/2}\frac{p_j-p_{j+1}}{h}. \tag{4.11}$$

Summing these gives for $1 < j < N$,

$$[\boldsymbol{R}\boldsymbol{p}]_j \stackrel{\text{def}}{=} \frac{1}{h^2}\Big(r_{j+1/2}(p_{j+1}-p_j) - r_{j-1/2}(p_j-p_{j-1})\Big). \tag{4.12}$$

The matrix stencil for the interior of $\boldsymbol{R}$ is thus

$$\frac{1}{h^2}\begin{bmatrix} -(r_{j-3/2}+r_{j-1/2}) & r_{j-1/2} & 0 \\ r_{j-1/2} & -(r_{j-1/2}+r_{j+1/2}) & r_{j+1/2} \\ 0 & r_{j+1/2} & -(r_{j+1/2}+r_{j+3/2}) \end{bmatrix}\begin{bmatrix} p_{j-1} \\ p_j \\ p_{j+1} \end{bmatrix}. \tag{4.13}$$

Recall the assumption that $rp(r) \to 0$ as $r \to \infty$ and that the scale of the domain of p is such that $p(1+\delta) \approx 0$ for all $\delta > 0$. Hence, the discretization of $\boldsymbol{R}$ has a zero right boundary condition, so $p_N = 0$ implies

$$[\boldsymbol{Rp}]_N = r_{N-1/2}\, p_{N-1}. \tag{4.14}$$

Since $p(r)$ is a radial profile, the implicit symmetry implies that $\frac{d}{dr}p(r) = 0$, i.e. a Neumann left-boundary condition. Since $p_1 = p(h/2)$, this implies $p_1 \approx p_0$ and

$$\begin{aligned}[\boldsymbol{Rp}]_1 &= r_{1/2}p_0 - (r_{1/2}+r_{3/2})p_1 + r_{3/2}p_2 \\ &= r_{3/2}p_1 + r_{3/2}p_2. \end{aligned} \tag{4.15}$$

Observe that $\boldsymbol{R}$ is a symmetric tridiagonal matrix.

We then take $\boldsymbol{L} \stackrel{\text{def}}{=} \left(\text{diag}(\boldsymbol{r}^{-1})\boldsymbol{R}\right)^2$ where $\text{diag}(\boldsymbol{r}^{-1})$ denotes the $N \times N$ diagonal matrix whose diagonal entries are (r_j^{-1}). Note that since $0 < r_j < 1$, the matrix $\text{diag}(\boldsymbol{r}^{-1/2})\boldsymbol{R}$ is strictly diagonally dominant, hence is positive definite [Golub and Van Loan, 2012, Theorem 3.4.3]. This is not surprising since it is a discretization of a positive definite operator. Finally, we approximate the integral in (4.8) with

$$\langle p, \mathcal{L}^2 p\rangle_{\mathscr{H}_1} \approx 2\pi h \langle \boldsymbol{p}, \boldsymbol{Lp}\rangle_{\mathbb{R}^N}. \tag{4.16}$$

So the complete approximation to (4.1) is

$$\Phi(p; b, \lambda, \delta) \approx \lambda h \|\boldsymbol{Gp} - \boldsymbol{b}\|^2_{\mathbb{R}^{2N+1}} + \delta 2\pi h \langle \boldsymbol{p}, \boldsymbol{Lp}\rangle_{\mathbb{R}^N}. \tag{4.17}$$

Since λ and δ will be stochastically modeled and estimated from the discrete hierarchical posterior, we absorb the constants h and $2\pi h$ into them, and define

$$F(\boldsymbol{p};\boldsymbol{b},\lambda,\delta) \stackrel{\text{def}}{=} \frac{1}{2}\left(\lambda\|\boldsymbol{G}\boldsymbol{p}-\boldsymbol{b}\|^2_{\mathbb{R}^{2N+1}} + \delta\langle \boldsymbol{p}, \boldsymbol{L}\boldsymbol{p}\rangle_{\mathbb{R}^N}\right). \tag{4.18}$$

4.1.2 The discrete hierarchical posterior distribution

As in [Bardsley, 2012], we employ a hierarchical model for λ and δ that employ independent prior distributions that form a natural conjugacy so that the resulting full conditional densities will be known up to a proportionality constant. In deriving this density, we will use a technique sometimes referred to as 'completing the square,' which in addition to showing that the posterior density for $\boldsymbol{p}$ is Gaussian, will allow us to marginalize the full conditional densities of the parameters λ and δ. This will be important for implementing the partially collapsed Gibbs sampler. In the following computations, $\langle\cdot,\cdot\rangle$ and $\|\cdot\|$ refer to the standard Euclidean inner product and norm on the appropriate finite dimensional subspace and should be clear in context. Moreover, we will not strictly adhere to the convention of capital letters corresponding to random variables since it conflicts with capital letters representing matrices, and again, this should be clear in context.

By the preceding discretization arguments, we have the following approximations for the prior, likelihood, and posterior densities,

$$\pi(\boldsymbol{p}|\delta) = (2\pi)^{-N/2}|\det \delta L|^{1/2}|\exp\left(-\frac{\delta}{2}\langle \boldsymbol{p}, \boldsymbol{L}\boldsymbol{p}\rangle\right), \tag{4.19}$$

$$\pi(\boldsymbol{b}|\boldsymbol{p},\lambda) = \left(\frac{\lambda}{2\pi}\right)^{-(2N+1)/2}\exp\left(-\frac{\lambda}{2}\|\boldsymbol{G}\boldsymbol{p}-\boldsymbol{b}\|^2\right), \tag{4.20}$$

and

$$\pi(\boldsymbol{p}|\boldsymbol{b},\lambda,\delta) \propto \exp\Big(-F(\boldsymbol{p};\boldsymbol{b},\lambda,\delta)\Big). \tag{4.21}$$

Taking the Bayesian perspective, the unknown parameters λ and δ are modelled as independent prior random quantities. We assume a hierarchical structure so that the prior $\boldsymbol{p}|\delta, \lambda$ is independent of the noise parameter λ (given δ) and that the measurement likelihood $\boldsymbol{b}|\boldsymbol{p}, \lambda, \delta$ is independent of the prior parameter δ (given $\boldsymbol{p}$ and λ) , i.e.

$$\pi(\boldsymbol{p}|\lambda, \delta) = \pi(\boldsymbol{p}|\delta) \tag{4.22}$$

$$\text{and } \pi(\boldsymbol{b}|\boldsymbol{p}, \lambda, \delta) = \pi(\boldsymbol{b}|\boldsymbol{p}, \lambda). \tag{4.23}$$

Note that both $\pi(\boldsymbol{b}|\delta)$ and $\pi(\boldsymbol{p}|\boldsymbol{b}, \lambda, \delta)$ are distributions in the exponential class of densities. As discussed in [Gelman et al., 2014], the exponential class forms a natural conjugacy, meaning roughly that for any prior and likelihood in the exponential class, there is a 'natural' well-defined posterior also in the exponential class. This convenience motivates the choice of prior distributions for λ and δ from within the exponential class, and in particular, the gamma distribution provides a flexible (albeit always right-skewed) family whose support is all positive real numbers. Assuming λ and δ are independent gamma-distributed random variables, they have probability density functions

$$\pi(\lambda) \propto \lambda^{\alpha-1} \exp(-\beta\lambda) \tag{4.24}$$

$$\text{and } \pi(\delta) \propto \delta^{\alpha-1} \exp(-\beta\delta), . \tag{4.25}$$

As recommended by [Higdon, 2006], we use parameter values $\alpha = 1$ and $\beta = 10^{-4}$ which provide a large prior variance (10^8) for λ and δ. Now, applying Bayes' theorem and the definition of conditional probability, the *joint posterior density* is

$$\begin{aligned}
\pi(\boldsymbol{p}, \lambda, \delta|\boldsymbol{b}) &= \frac{\pi(\boldsymbol{b}|\boldsymbol{p}, \lambda, \delta)\pi(\boldsymbol{p}, \lambda, \delta)}{\pi(\boldsymbol{b})} \\
&= \frac{\pi(\boldsymbol{b}|\boldsymbol{p}, \lambda, \delta)\pi(\boldsymbol{p}|\delta, \lambda)\pi(\lambda, \delta)}{\pi(\boldsymbol{b})} \\
&= \frac{\pi(\boldsymbol{b}|\boldsymbol{p}, \lambda)\pi(\boldsymbol{p}|\delta)\pi(\lambda)\pi(\delta)}{\pi(\boldsymbol{b})}
\end{aligned}$$

$$\propto \lambda^{(2N+1)/2+\alpha-1}\delta^{N/2+\alpha-1}\exp\left(-\frac{\lambda}{2}\|\boldsymbol{G}\boldsymbol{p}-\boldsymbol{b}\|^2-\frac{\delta}{2}\langle\boldsymbol{p},\boldsymbol{L}\boldsymbol{p}\rangle-\beta\lambda-\beta\delta\right). \quad (4.26)$$

Our primary goal for estimation and uncertainty quantification of $\boldsymbol{p}$ will be drawing inference from (4.26). As previously remarked, all priors are in the exponential family, hence there is a natural expression for each full conditional density that is also in the exponential family. We proceed by deriving full conditional densities for λ, δ and $\boldsymbol{p}$.

Observe first that

$$\begin{aligned}\pi(\lambda|\boldsymbol{b},\boldsymbol{p},\delta)=\frac{\pi(\boldsymbol{p},\lambda,\delta|\boldsymbol{b})}{\pi(\boldsymbol{p},\delta|\boldsymbol{b})}&\propto\lambda^{(2N+1)/2+\alpha-1}\exp\left(-\lambda\left(\frac{1}{2}\|\boldsymbol{G}\boldsymbol{x}-\boldsymbol{b}\|^2-\beta\right)\right),\\ \text{and }\pi(\delta|\boldsymbol{b},\boldsymbol{p},\lambda)=\frac{\pi(\boldsymbol{p},\lambda,\delta|\boldsymbol{b})}{\pi(\boldsymbol{p},\lambda|\boldsymbol{b})}&\propto\delta^{N/2+\alpha-1}\exp\left(-\delta\left(\frac{1}{2}\langle\boldsymbol{p},\boldsymbol{L}\boldsymbol{p}\rangle-\beta\right)\right),\end{aligned} \quad (4.27)$$

each of which are proportional to gamma distributions with shifted scale and rate parameters.

Deriving the density for $\boldsymbol{p}$ is more involved and uses a technique sometimes referred to as 'completing the square' [Stuart, 2010] which shows that the discrete posterior $\pi(\boldsymbol{p}|\boldsymbol{b},\lambda,\delta)$ is Gaussian. Since $\boldsymbol{G}$ is a discretization of an injective operator, the matrix $\boldsymbol{G}$ has linearly independent columns, so $\boldsymbol{G}^T\boldsymbol{G}$ is symmetric positive definite. Thus, the matrix

$$\boldsymbol{J}_{\lambda,\delta}\overset{\text{def}}{=}(\lambda\boldsymbol{G}^T\boldsymbol{G}+\delta\boldsymbol{L}) \quad (4.28)$$

is also symmetric positive definite, and hence, invertible. Define

$$\boldsymbol{m}_{\lambda,\delta}\overset{\text{def}}{=}\boldsymbol{J}_{\lambda,\delta}^{-1}\lambda\boldsymbol{G}^T\boldsymbol{b}, \quad (4.29)$$

then observe

$$\begin{aligned}2F(\boldsymbol{p};\boldsymbol{b},\lambda,\delta)&=\lambda\|\boldsymbol{G}\boldsymbol{x}-\boldsymbol{b}\|^2+\delta\,\langle\boldsymbol{p},\boldsymbol{L}\boldsymbol{p}\rangle\\ &=\lambda\langle\boldsymbol{G}\boldsymbol{p},\boldsymbol{G}\boldsymbol{p}\rangle-2\lambda\langle\boldsymbol{G}\boldsymbol{p},\boldsymbol{b}\rangle+\lambda\langle\boldsymbol{b},\boldsymbol{b}\rangle+\delta\langle\boldsymbol{p},\boldsymbol{L}\boldsymbol{p}\rangle\end{aligned}$$

$$
\begin{aligned}
&= \left\langle \boldsymbol{p}, (\lambda \boldsymbol{G}^T\boldsymbol{G} + \delta \boldsymbol{L})\boldsymbol{p} \right\rangle - 2\lambda \left\langle \boldsymbol{p}, \boldsymbol{G}^T\boldsymbol{b} \right\rangle + \lambda \|\boldsymbol{b}\|^2 \\
&= \langle \boldsymbol{p}, \boldsymbol{J}_{\lambda,\delta}\boldsymbol{p} \rangle - 2 \langle \boldsymbol{p}, \boldsymbol{J}_{\lambda,\delta}\boldsymbol{m}_{\lambda,\delta} \rangle + \lambda \|\boldsymbol{b}\|^2 \\
&= \langle \boldsymbol{p}, \boldsymbol{J}_{\lambda,\delta}(\boldsymbol{p} - \boldsymbol{m}_{\lambda,\delta}) \rangle - \langle \boldsymbol{p}, \boldsymbol{J}_{\lambda,\delta}\boldsymbol{m}_{\lambda,\delta} \rangle + \lambda \|\boldsymbol{b}\|^2 \\
&= \langle (\boldsymbol{p} - \boldsymbol{m}_{\lambda,\delta}), \boldsymbol{J}_{\lambda,\delta}(\boldsymbol{p} - \boldsymbol{m}_{\lambda,\delta}) \rangle + \langle \boldsymbol{m}_{\lambda,\delta}, \boldsymbol{J}_{\lambda,\delta}\boldsymbol{p} \rangle \\
&\qquad - \langle \boldsymbol{m}_{\lambda,\delta}, \boldsymbol{J}_{\lambda,\delta}\boldsymbol{m}_{\lambda,\delta} \rangle - \langle \boldsymbol{p}, \boldsymbol{J}_{\lambda,\delta}\boldsymbol{m}_{\lambda,\delta} \rangle + \lambda \|\boldsymbol{b}\|^2 \\
&= \langle (\boldsymbol{p} - \boldsymbol{m}_{\lambda,\delta}), \boldsymbol{J}_{\lambda,\delta}(\boldsymbol{p} - \boldsymbol{m}_{\lambda,\delta}) \rangle - \langle \boldsymbol{m}_{\lambda,\delta}, \boldsymbol{J}_{\lambda,\delta}\boldsymbol{m}_{\lambda,\delta} \rangle + \lambda \|\boldsymbol{b}\|^2, \qquad (4.30)
\end{aligned}
$$

where in the second to last equality, we used the symmetry of $\boldsymbol{J}_{\lambda,\delta}$. Hence

$$
\begin{aligned}
\pi(\boldsymbol{p}|\boldsymbol{b}, \lambda, \delta) &= \frac{\pi(\boldsymbol{p}, \lambda, \delta|\boldsymbol{b})}{\pi(\lambda, \delta|\boldsymbol{b})} \\
&\propto \exp\left(-\frac{1}{2} \langle (\boldsymbol{p} - \boldsymbol{m}_{\lambda,\delta}), \boldsymbol{J}_{\lambda,\delta}(\boldsymbol{p} - \boldsymbol{m}_{\lambda,\delta}) \rangle \right) \qquad (4.31)
\end{aligned}
$$

which is proportional to a multivariate Gaussian with mean $\boldsymbol{m}_{\lambda,\delta} = \boldsymbol{J}_{\lambda,\delta}^{-1}\lambda\boldsymbol{G}^T\boldsymbol{b}$ and covariance matrix $\boldsymbol{J}_{\lambda,\delta}^{-1}$.

With explicit expressions for the full conditional densities, we can now explicitly state the algorithms for posterior PSF estimation.

4.2 Sampling the PSF posterior

This section is devoted to applying the general algorithms presented in Section 3.1 to the specific PSF posterior estimation problem. The algorithms were presented generically, assuming at each step that the corresponding conditional random variable, or proposal in the case of Metropolis-Hastings, could be simulated on the computer. In the last section, we derived the full-conditional densities and wrote them in a form so that they can be easily sampled using standard algorithms for generating gamma and Gaussian random variables. These expressions will be sufficient for the simulations in the standard Gibbs sampler, which we present in Sec-

tion 4.2.1. The partially collapsed sampler for the problematic simulation of δ alluded to at the beginning of the chapter will require a derivation that shows that the collapsed conditional density is Gaussian. With this derivation, we can then present the detailed partially collapsed Gibbs sampler for PSF reconstruction.

The general idea for random variable generation involves transforming a uniform random variable on $[0, 1]$ into a random variable with the desired distribution. This is known as the *probability integral transformation*. In theory, this can always be done if an inverse of the cumulative distribution function $F(x) \stackrel{\text{def}}{=} \mathbb{P}(X \leq x)$ is computationally available. For random variables with continuous densities, F is strictly increasing onto $[0, 1]$ and is thus invertible. Hence, for $U \sim \mathrm{U}([0, 1])$, the variable $X = F^{-1}(U)$ has

$$\mathbb{P}(X \leq x) = \mathbb{P}(F^{-1}(U) \leq x) = \mathbb{P}(U \leq F(x)) = \int_0^{F(x)} ds = F(x). \tag{4.32}$$

So, to simulate X, one generates a pseudo-random number u from $[0, 1]$ (see [Knuth, 1982]), then $F^{-1}(u)$ serves as simulation for X. In practice, this method is usually analytically difficult, but the idea behind most algorithms is similar: generate a pseudo-random number then transform it in some way so that the resulting random variable has the desired density. For many common distributions these algorithms are implemented efficiently in many statistical and mathematical computing packages, and we assume for the following algorithms that they are available. In particular, we assume that simulated data from a uniform density $U([0, 1])$, a gamma distribution $\Gamma(\alpha, \beta)$ for given shape and rate parameters α and β, and a standard Gaussian $\mathcal{N}(0, 1)$ can each be computed.

4.2.1 Gibbs sampling the PSF posterior

Here, we describe how to explicitly obtain simulations from the full conditional densities $\pi(\lambda|\boldsymbol{b}, \boldsymbol{p}, \delta)$, $\pi(\delta|\boldsymbol{b}, \boldsymbol{p}, \lambda)$, and $\pi(\boldsymbol{p}|\boldsymbol{b}, \lambda, \delta)$. The equations in (4.27) imply that both $\pi(\lambda|\boldsymbol{b}, \boldsymbol{p}, \delta)$,

$\pi(\delta|\boldsymbol{b}, \boldsymbol{p}, \lambda)$ are gamma distributions

$$(\lambda|\boldsymbol{b}, \boldsymbol{p}, \delta) \sim \Gamma\left((2N+1)/2 + \alpha, \frac{1}{2}\|\boldsymbol{G}\boldsymbol{x} + \boldsymbol{b}\|^2 + \beta\right) \tag{4.33}$$

and

$$(\delta|\boldsymbol{b}, \boldsymbol{p}, \lambda) \sim \Gamma\left(N/2 + \alpha, \frac{1}{2}\langle \boldsymbol{p}, \boldsymbol{L}\boldsymbol{p}\rangle + \beta\right) \tag{4.34}$$

respectively, of which simulations are assumed to be available.

For simulating from $\pi(\boldsymbol{p}|\boldsymbol{b}, \lambda, \delta)$, let $\boldsymbol{z}$ be a vector whose entries are N independent realizations from $\mathcal{N}(0, 1)$. Hence, $\boldsymbol{z}$ is a realization of a multivariate Gaussian $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{N\times N})$. Recall that for $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, the linear transformation $\boldsymbol{w} = \boldsymbol{m} + \boldsymbol{B}\boldsymbol{z}$ results in $\boldsymbol{m} + \boldsymbol{B}\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{m}, \boldsymbol{B}\boldsymbol{B}^T)$. So in order to sample a Gaussian random variable with given precision, we will need to factor its inverse. An important feature of positive definite matrices, $\boldsymbol{A}$, is that they have an eigenvalue decomposition of the form $\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^*$ (here * denotes the conjugate transpose since columns of $\boldsymbol{U}$ may be complex valued), where the columns $\boldsymbol{U}^*$ are mutually orthonormal and that $\boldsymbol{\Lambda}$ is a diagonal matrix of positive eigenvalues. Therefore, there exists a matrix $\boldsymbol{M} = \boldsymbol{\Lambda}^{-1/2}\boldsymbol{U}^*$, such that $\boldsymbol{M}^*\boldsymbol{M} = \boldsymbol{A}^{-1}$, where the $-1/2$ power is computed on the diagonal entries of $\boldsymbol{\Lambda}$. Hence, the linear transformation $\boldsymbol{M}^*\boldsymbol{z} \sim \mathcal{N}(0, \boldsymbol{A}^{-1})$. In practice, computing the eigenvalue decomposition is overly expensive, but this argument establishes the existence of such a matrix.

An efficient method for computing such an $\boldsymbol{M}$ is the Cholesky factorization, which for a given symmetric positive definite matrix, gives a lower triangular matrix $\boldsymbol{R}$ with non-zero diagonals such that $\boldsymbol{A} = \boldsymbol{M}^T\boldsymbol{M}$ and can be computed in $O(N^3)$ floating-point operations (flops) [Golub and Van Loan, 2012]. For $\boldsymbol{J}_{\lambda,\delta}$, define the Cholesky factors

$$\boldsymbol{R}_{\lambda,\delta}^T\boldsymbol{R}_{\lambda,\delta} \stackrel{\text{def}}{=} \boldsymbol{J}_{\lambda,\delta}. \tag{4.35}$$

With $\boldsymbol{R}_{\lambda,\delta}$ in hand, it serves two purposes: first, we can solve $\boldsymbol{J}_{\lambda,\delta}\boldsymbol{m}_{\lambda,\delta} = \boldsymbol{R}_{\lambda,\delta}^T\boldsymbol{R}_{\lambda,\delta}\boldsymbol{m}_{\lambda,\delta} = \boldsymbol{G}^T\boldsymbol{b}$ efficiently by forward-substitution then by backward-substitution, both in $O(N^2)$ flops; second, the computation $\boldsymbol{m}_{\lambda,\delta}+\boldsymbol{R}_{\lambda,\delta}^{-1}\boldsymbol{z}$ by forward-substitution, transforms $\boldsymbol{z}$ into a realization from $\pi(\boldsymbol{p}|\boldsymbol{b},\lambda,\delta)$, since $(\boldsymbol{R}_{\lambda,\delta}^T\boldsymbol{R}_{\lambda,\delta})^{-1} = \boldsymbol{R}_{\lambda,\delta}^{-1}{\boldsymbol{R}_{\lambda,\delta}^T}^{-1}$.

Note that each time a realization from $\pi(\boldsymbol{p}|\boldsymbol{b},\lambda,\delta)$ is required, we must compute a factorization that depends on λ and δ, and this step will be the computational bottleneck for the Gibbs sampler. We remark that for the scale of our problem, Cholesky factorizations are feasible. In general, this may not always be the case, and [Bardsley, 2012] provides methods for sampling that rely only on linear solves which may be implemented efficiently via an algorithm like conjugate gradients.

With computational methods for each full-conditional density, Algorithm 6 describes Gibbs sampling the PSF posterior.

Algorithm 6 Hierarchical Gibbs sampler for PSF posterior estimation

Given λ^k, δ^k, and $\boldsymbol{p}^k$.

1. Simulate $\lambda^{k+1} \sim \Gamma\left((2N+1)/2+\alpha, \frac{1}{2}\|\boldsymbol{G}\boldsymbol{p}^k-\boldsymbol{b}\|^2+\beta\right)$.
2. Simulate $\delta^{k+1} \sim \Gamma\left(N/2+\alpha, \frac{1}{2}\left\langle \boldsymbol{p}^k, \boldsymbol{L}\boldsymbol{p}^k\right\rangle+\beta\right)$.
3. Compute $\boldsymbol{R}_{\lambda^{k+1},\delta^{k+1}}(4.35), \boldsymbol{m}_{\lambda^{k+1},\delta^{k+1}}(4.29)$,
   and set $\boldsymbol{p}^{k+1} = \boldsymbol{R}^{-1}_{\lambda^{k+1},\delta^{k+1}}\boldsymbol{z} + \boldsymbol{m}_{\lambda^{k+1},\delta^{k+1}}$ where $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{N\times N})$.

4.2.2 Partially collapsed Gibbs sampling for PSF reconstruction

As we will see, the (δ^k) component of the Markov chain in the Algorithm 6 exhibits poor convergence, hence asymptotic results from the ergodic theorem require longer runs of the Markov chain. Taking a cue from [Agapiou et al., 2014], we remove the conditioning of δ^{k+1} on $\boldsymbol{p}^k$ by implementing Algorithm 5 on the posterior PSF estimation problem.

As previously mentioned, this will require a simulation from the density $\pi(\delta|\boldsymbol{b}, \lambda)$. To express the kernel of this density, note (4.31) is the kernel of a Gaussian with mean $\boldsymbol{m}_{\lambda,\delta}$ and variance $\boldsymbol{J}_{\lambda,\delta}^{-1}$, thus the normalized density is

$$\begin{aligned}\frac{\pi(\boldsymbol{p}, \lambda, \delta|\boldsymbol{b})}{\pi(\lambda, \delta|\boldsymbol{b})} &= \pi(\boldsymbol{p}|\boldsymbol{b}, \lambda, \delta) \\ &= (2\pi)^{-N/2} |\det \boldsymbol{J}_{\lambda,\delta}|^{1/2} \exp\left(-\frac{1}{2}\langle \boldsymbol{p} - \boldsymbol{m}_{\lambda,\delta}, \boldsymbol{J}_{\lambda,\delta}(\boldsymbol{p} - \boldsymbol{m}_{\lambda,\delta})\rangle\right). \end{aligned} \tag{4.36}$$

Dividing (4.26) by (4.36), one obtains

$$\begin{aligned}\pi(\lambda, \delta|\boldsymbol{b}) &= \frac{\pi(\boldsymbol{p}, \lambda, \delta|\boldsymbol{b})}{\pi(\boldsymbol{p}|\boldsymbol{b}, \lambda, \delta)} \\ &\propto \lambda^{\frac{2N+1}{2}+\alpha-1} \delta^{\frac{N}{2}+\alpha-1} |\det \boldsymbol{J}_{\lambda,\delta}|^{-1/2} \\ &\quad \times \exp\left(\frac{1}{2}\langle \boldsymbol{p} - \boldsymbol{m}_{\lambda,\delta}, \boldsymbol{J}(\boldsymbol{p} - \boldsymbol{m}_{\lambda,\delta})\rangle - F(\boldsymbol{p}; \boldsymbol{b}, \lambda, \delta) - \beta\lambda - \beta\delta\right) \\ &= \propto \lambda^{\frac{2N+1}{2}+\alpha-1} \delta^{\frac{N}{2}+\alpha-1} |\det \boldsymbol{J}_{\lambda,\delta}|^{-1/2} \\ &\quad \times \exp\left(-\frac{1}{2}\left(\lambda\|\boldsymbol{b}\|^2 - \langle \boldsymbol{m}_{\lambda,\delta}, \boldsymbol{J}_{\lambda,\delta}\boldsymbol{m}_{\lambda,\delta}\rangle\right) - \beta\lambda - \beta\delta\right). \end{aligned} \tag{4.37}$$

Finally,

$$\begin{aligned}\pi(\delta|\boldsymbol{b}, \lambda) &= \frac{\pi(\lambda, \delta|\boldsymbol{b})}{\pi(\lambda|\boldsymbol{b})} \\ &\propto \delta^{\frac{N}{2}+\alpha} |\det \boldsymbol{J}_{\lambda,\delta}|^{-1/2} \exp\left(-\frac{1}{2}\left(\lambda\|\boldsymbol{b}\|^2 - \langle \boldsymbol{m}_{\lambda,\delta}, \boldsymbol{J}_{\lambda,\delta}\boldsymbol{m}_{\lambda,\delta}\rangle\right) - \beta\delta\right). \end{aligned} \tag{4.38}$$

The two terms $|\det \boldsymbol{J}_{\lambda,\delta}|$ and $\langle \boldsymbol{m}_{\lambda,\delta}, \boldsymbol{J}_{\lambda,\delta}\boldsymbol{m}_{\lambda,\delta}\rangle$ in (4.38) make the density depend in a complicated way on δ, so a direct simulation is not available. Additionally, they are computationally expensive in that they involve determinants and linear solves. Fortunately, the Cholesky factorization $\boldsymbol{R}_{\lambda,\delta}$ will allow both evaluations to be computed efficiently and the Metropolis-Hastings step described in Algorithm 5 can be used to sample from (4.38). Since $|\det \boldsymbol{J}_{\lambda,\delta}|$ involves N products and $\langle \boldsymbol{m}_{\lambda,\delta}, \boldsymbol{J}_{\lambda,\delta}\boldsymbol{m}_{\lambda,\delta}\rangle$ occurs in the argument of an exponential, we perform calculations on a logarithmic scale.

To simplify some arguments, the following calculations divide (4.38) into terms that depend only on expensive quantities $\boldsymbol{R}_{\lambda,\delta}$ and $\boldsymbol{m}_{\lambda,\delta}$. That is,

$$\begin{aligned}
-\frac{1}{2}\left(\lambda\|\boldsymbol{b}\|^2 - \langle \boldsymbol{m}_{\lambda,\delta}, \boldsymbol{J}_{\lambda,\delta}\boldsymbol{m}_{\lambda,\delta}\rangle\right) &= -\frac{1}{2}\left(\lambda\langle \boldsymbol{b},\boldsymbol{b}\rangle - \langle \boldsymbol{m}_{\lambda,\delta}, \lambda\boldsymbol{G}^T\boldsymbol{b}\rangle\right) \\
&= -\frac{\lambda}{2}\langle \boldsymbol{b} - \boldsymbol{G}\boldsymbol{m}_{\lambda,\delta}, \boldsymbol{b}\rangle \\
&\stackrel{\text{def}}{=} -\frac{\lambda}{2}a(\boldsymbol{m}_{\lambda,\delta}),
\end{aligned} \tag{4.39}$$

and

$$\begin{aligned}
\ln(|\det \boldsymbol{J}_{\lambda,\delta}|^{-1/2}) &= -\frac{1}{2}\ln(|\det \boldsymbol{R}_{\lambda,\delta}\boldsymbol{R}_{\lambda,\delta}^T|) \\
&= -\frac{1}{2}\ln(|\det \boldsymbol{R}_{\lambda,\delta}|^2) \\
&= -\frac{1}{2}\ln\left(\prod_{i=1}^{N}|\boldsymbol{R}_{\lambda,\delta_{ii}}|^2\right) \\
&= -\sum_{i=1}^{N}\ln|\boldsymbol{R}_{\lambda,\delta_{ii}}| \\
&\stackrel{\text{def}}{=} -b(\boldsymbol{R}_{\lambda,\delta}),
\end{aligned} \tag{4.40}$$

where we used the fact that $\boldsymbol{R}_{\lambda,\delta}$ is lower triangular to compute the determinant. Substituting these expressions into (4.38)

$$\begin{aligned}
\pi(\delta|\boldsymbol{b},\lambda) &\propto \delta^{\frac{N}{2}+\alpha}|\det \boldsymbol{J}_{\lambda,\delta}|^{-1/2}\exp\left(-\frac{\lambda}{2}\langle \boldsymbol{b}-\boldsymbol{G}\boldsymbol{m}_{\lambda,\delta},\boldsymbol{b}\rangle - \beta\delta.\right) \\
&= \exp\left(\left(\frac{N}{2}+\alpha-1\right)\ln\delta - b(\boldsymbol{R}_{\lambda,\delta}) - \frac{\lambda}{2}a(\boldsymbol{m}_{\lambda,\delta}) - \beta\delta\right) \\
&\stackrel{\text{def}}{=} \exp\Big(c(\boldsymbol{R}_{\lambda,\delta},\boldsymbol{m}_{\lambda,\delta},\delta)\Big).
\end{aligned} \tag{4.41}$$

We also use a logarithmic scale for the proposal, that is, a random walk on the logarithm of δ. This means that the proposal density is $\rho(\delta'|\delta) \stackrel{\text{def}}{=} \phi_\gamma(|\ln\delta' - \ln\delta|) = \rho(\delta|\delta')$, where ϕ_γ is the density of a mean-zero normal random variable with standard deviation γ. To simulate

the proposal, draw $w \sim \mathcal{N}(0,1)$ and set $\delta' \stackrel{\text{def}}{=} \exp(\gamma w + \ln \delta)$, then $\ln \delta' - \ln \delta \sim \rho(\delta'|\delta)$. This has the added benefit of producing proposal simulations such that $\delta' > 0$.

We compute the acceptance ratio on a logarithmic scale as follows: observe that accepting with probability

$$\alpha(\delta, \delta') = \min\left\{1, \frac{\pi(\delta'|\boldsymbol{b}, \lambda)}{\pi(\delta|\boldsymbol{b}, \lambda)}\right\} \tag{4.42}$$

is equivalent to accepting with log uniform probability

$$\ln \alpha(\delta, \delta') = \min\left\{0, c(\boldsymbol{R}_{\lambda,\delta'}, \boldsymbol{m}_{\lambda,\delta'}, \delta') - c(\boldsymbol{R}_{\lambda,\delta}, \boldsymbol{m}_{\lambda,\delta}, \delta)\right\} \tag{4.43}$$

since ln is increasing from $(0,1)$ onto $(-\infty, 0)$. To implement this, generate a uniform simulation u from $[0,1]$, then accept if $\ln u > \ln \alpha(\delta, \delta')$ and reject otherwise. All of the computational pieces are in place to explicitly describe Metropolis-Hastings within PCG for PSF reconstruction in Algorithm 5. The full implementation is described in Algorithm 7. Note that we are able to re-use the factorization $\boldsymbol{R}_{\lambda,\delta}$ and $\boldsymbol{m}_{\lambda,\delta}$ to sample $\boldsymbol{p}^{k+1}$, so there are $n_{mh} + 1$ Cholesky factorizations per Markov iteration.

Algorithm 7 Metropolis-Hastings within PCG sampler for PSF posterior estimation

Given $\gamma, \lambda^k, \delta^k$, and $\boldsymbol{p}^k$

1. Simulate $\lambda^{k+1} \sim \Gamma\left((2N+1)/2 + \alpha, \frac{1}{2}\|\boldsymbol{G}\boldsymbol{p}^k - \boldsymbol{b}\|^2 + \beta\right)$.
2. Set $\lambda = \lambda^{k+1}, \delta = \delta^k$ and compute $\boldsymbol{R}_{\lambda,\delta}(4.35), \boldsymbol{m}_{\lambda,\delta}(4.29)$, then $c(\boldsymbol{R}_{\lambda,\delta}, \boldsymbol{m}_{\lambda,\delta}, \delta)(4.41)$.
   For $j = 1 \ldots n_{mh}$
   i. Simulate $w \sim \mathcal{N}(0,1)$ and set $\delta' = \exp(\gamma w + \delta)$
   ii. Compute $\boldsymbol{R}_{\lambda,\delta'}, \boldsymbol{m}_{\lambda,\delta'}$, then $c(\boldsymbol{R}_{\lambda,\delta'}, \boldsymbol{m}_{\lambda,\delta'}, \delta')$.
   iii. Simulate $u \sim U([0,1])$ and
   if $\ln u > \min\left\{0, c(\boldsymbol{R}_{\lambda,\delta'}, \boldsymbol{m}_{\lambda,\delta'}, \delta') - c(\boldsymbol{R}_{\lambda,\delta}, \boldsymbol{m}_{\lambda,\delta}, \delta)\right\}$
   set $\delta = \delta', \boldsymbol{R}_{\lambda,\delta} = \boldsymbol{R}_{\lambda,\delta'}, \boldsymbol{m}_{\lambda,\delta} = \boldsymbol{m}_{\lambda,\delta}$, and $c(\boldsymbol{R}_{\lambda,\delta}, \boldsymbol{m}_{\lambda,\delta}, \delta) = c(\boldsymbol{R}_{\lambda,\delta'}, \boldsymbol{m}_{\lambda,\delta'}, \delta')$
   Set $\delta^{k+1} = \delta$
3. Simulate $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{N\times N})$ and set $\boldsymbol{p}^{k+1} = \boldsymbol{R}_{\lambda,\delta}^{-1}\boldsymbol{z} + \boldsymbol{m}_{\lambda,\delta}$.

We have not addressed how to choose the proposal variance γ^2. It is common to tune this parameter so that the long-run proportion of acceptances is about 0.4 [Calvetti and Somersalo, 2007]. An alternative is to use previous values to inform γ. The resulting stochastic process is no longer a Markov chain, but [Haario et al., 2005] have shown that the stochastic process $\{X^1, X^2, \dots\}$ resulting from a Metropolis-Hastings algorithm using the empirical covariance estimate of the previous k realizations as the proposal variance at step k enjoys a similar ergodic result as Theorem 3.1.1. The theory is not directly applicable, since we sample jointly $\{(\lambda^k, \delta^k, \boldsymbol{p}^k)\}$, and obtaining covariance estimate of the joint variable is computationally unfeasible. A feasible computation for γ^k is the marginal variance for δ would be to add to Algorithm 7

$$4.\ \text{Set } \gamma^2 = \frac{1}{k}\sum_{i=1}^{k+1}(\delta^i - \overline{\boldsymbol{\delta}_k})^2 \tag{4.44}$$

where $\overline{\boldsymbol{\delta}_k}$ is the sample mean for $\{\delta^1, \dots, \delta^k\}$. Although we do not directly have an ergodic theorem for this stochastic process, it exhibits similar convergence statistics in the numerical examples presented in Section 4.3 with much less 'tuning-effort' as the algorithm with a tuned γ^2. From a practical standpoint, one could use the adaptive estimate of γ^k, then when the chain has stabilized, fix γ to appeal to Theorem 3.1.1 for statistic estimation.

4.2.3 Blocking the sampler and a connection to marginal then conditional sampling

We now explore one more modification of the algorithm and illustrate a connection to the work of [Fox and Norton, 2015]. Note that the joint density in (4.26) factors in λ and δ, and since $\pi(\lambda, \delta|\boldsymbol{b}, \boldsymbol{p}) \propto \pi(\lambda, \delta, \boldsymbol{p}|\boldsymbol{b})$, the conditional variables $(\lambda|\boldsymbol{b}), (\boldsymbol{p}, \delta)$ and $(\delta|\boldsymbol{b}, \boldsymbol{p}, \lambda)$ are independent. Hence, steps 1. and 2. in Algorithm 6 can be thought of as a joint sample from $\pi(\lambda, \delta|\boldsymbol{b}, \boldsymbol{p}^k)$. This procedure is sometimes referred to as *blocking* [Liu, 2008]. To accomplish the Metropolis-Hastings step on the logarithmic scale, we derive the analogous c in (4.41), by

using (4.25) in

$$\begin{aligned}\pi(\lambda,\delta|\boldsymbol{b}) &= \pi(\delta|\boldsymbol{b},\lambda)\pi(\lambda|\boldsymbol{b})\\ &\propto \pi(\delta|\boldsymbol{b},\lambda)\pi(\lambda)\\ &= \exp\left(\left(\frac{2N+1}{2}+\alpha-1\right)\ln\lambda+\left(\frac{N}{2}+\alpha-1\right)\ln\delta-b(\boldsymbol{R}_{\lambda,\delta})-\frac{\lambda}{2}a(\boldsymbol{m}_{\lambda,\delta})-\beta\lambda-\beta\delta\right)\\ &\stackrel{\text{def}}{=} \exp\Big(c(\boldsymbol{R}_{\lambda,\delta},\boldsymbol{m}_{\lambda,\delta},\lambda,\delta)\Big). \qquad (4.45)\end{aligned}$$

Applying partial collapse to the blocked Gibbs sampler results in Algorithm 8

Algorithm 8 Metropolis-Hastings within blocked PCG sampling for PSF posterior estimation

Given $\boldsymbol{C}, [\lambda^k, \delta^k]$, and $\boldsymbol{p}^k$

1. Set $[\lambda,\delta] = [\lambda^k,\delta^k]$ and compute $\boldsymbol{R}_{\lambda,\delta}(4.35), \boldsymbol{m}_{\lambda,\delta}(4.29)$, then $c(\boldsymbol{R}_{\lambda,\delta},\boldsymbol{m}_{\lambda,\delta},\lambda,\delta)(4.45)$.
   For $j = 1 \dots M$
   i. Simulate $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{2\times 2})$ and set $[\lambda',\delta'] = \exp(\boldsymbol{C}\boldsymbol{w} + [\lambda,\delta]^T)$
   ii. Compute $\boldsymbol{R}_{\lambda',\delta'}, \boldsymbol{m}_{\lambda',\delta'}$, then $c(\boldsymbol{R}_{\lambda',\delta'},\boldsymbol{m}_{\lambda',\delta'},\lambda',\delta')$.
   iii. Simulate $u \sim U([0,1])$ and
      if $\ln u > \min\left\{0, c(\boldsymbol{R}_{\lambda',\delta'},\boldsymbol{m}_{\lambda',\delta'},\delta') - c(\boldsymbol{R}_{\lambda,\delta},\boldsymbol{m}_{\lambda,\delta},\delta)\right\}$
         set $[\lambda,\delta] = [\lambda',\delta'], \boldsymbol{R}_{\lambda,\delta} = \boldsymbol{R}_{\lambda',\delta'}, \boldsymbol{m}_{\lambda',\delta} = \boldsymbol{m}_{\lambda',\delta}$, and $c(\boldsymbol{R}_{\lambda',\delta},\boldsymbol{m}_{\lambda,\delta},\delta) = c(\boldsymbol{R}_{\lambda',\delta'},\boldsymbol{m}_{\lambda',\delta'},\lambda',\delta')$
   Set $\delta^{k+1} = \delta$
2. Simulate $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{N\times N})$ and set $\boldsymbol{p}^{k+1} = \boldsymbol{R}_{\lambda,\delta}^{-1}\boldsymbol{z} + \boldsymbol{m}_{\lambda,\delta}$.

By design, simulations of (λ^k,δ^k) are conditionally independent of $\boldsymbol{p}^k$. In the language of [Van Dyk and Park, 2008], $\boldsymbol{p}^k$ has been completely collapsed, and (λ^k,δ^k) provides an independent Markov chain invariant with respect to $\pi(\lambda,\delta|\boldsymbol{b})$. Markov chains that satisfy this property are said to satisfy the *Duality Principle* [Robert and Casella, 2013, Section 9.2.3] and are related to hidden Markov models. Of course, $\boldsymbol{p}^k$ is the primary quantity of interest for estimation and uncertainty quantification, and estimating λ and δ are auxiliary to that goal. Despite this apparent mismatch, it does suggest a strategy that can reduce the number of required Cholesky factorizations. Consider only iterating step 1. in Algorithm 8 to obtain a Markov chain $(\lambda^k,\delta^k|\boldsymbol{b})$ invariant with respect to $\pi(\lambda,\delta|\boldsymbol{b})$. After this chain has

sufficiently converged, in say N steps, we can produce a 'thinned' chain $\{\boldsymbol{p}^{a_k}|\lambda^{a_k}, \delta^{a_k}, \boldsymbol{b}\}$ that computes estimates for some sequence $\{a_k\} \subseteq \{1 \dots N\}$. How we choose this sequence is based on the integrated autocorrelation time of $\{(\lambda^k, \delta^k)\}$ which is defined in Section 3.2. This is precisely the MCMC algorithm presented in [Fox and Norton, 2015] for image deblurring, except in their case, the forward operator corresponds to a convolution, for which the discrete Fourier transform can be applied, rather than a Cholesky factorization. Moreover, they present several other methods for speeding up the algorithm so that the costly computations involving determinants and linear solves can be done offline.

To see this in our situation, first note that the computation of $a(\boldsymbol{m}_{\lambda,\delta})$ can be simplified as follows; continuing from (4.39) and using (4.29) then (4.28)

$$\begin{aligned} a(\boldsymbol{m}_{\lambda,\delta}) &= \langle \boldsymbol{b} - \boldsymbol{G}\boldsymbol{m}_{\lambda,\delta}, \boldsymbol{b} \rangle \\ &= \left\langle \boldsymbol{b}, \left(I - \boldsymbol{G}\boldsymbol{J}_{\lambda,\delta}^{-1} \lambda \boldsymbol{G}^T \right) \boldsymbol{b} \right\rangle \\ &= \left\langle \boldsymbol{b}, \left(I - \boldsymbol{G} \left(\boldsymbol{G}^T\boldsymbol{G} + \frac{\delta}{\lambda}\boldsymbol{L} \right)^{-1} \boldsymbol{G}^T \right) \boldsymbol{b} \right\rangle. \end{aligned} \tag{4.46}$$

The Woodbury matrix identity [Woodbury, 1950] states

$$(\boldsymbol{A} + \boldsymbol{U}\boldsymbol{C}\boldsymbol{V})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{U} \left(\boldsymbol{V}\boldsymbol{A}^{-1}\boldsymbol{U} + \boldsymbol{C}^{-1} \right)^{-1} \boldsymbol{V}\boldsymbol{A}^{-1}, \tag{4.47}$$

so taking $\boldsymbol{A} = \boldsymbol{I}, \boldsymbol{U} = \boldsymbol{G}, \boldsymbol{V} = \boldsymbol{G}^T$, and $\boldsymbol{C} = \left(\frac{\delta}{\lambda}\boldsymbol{L}\right)^{-1}$ gives

$$\left(I - \boldsymbol{G} \left(\boldsymbol{G}^T\boldsymbol{G} - \frac{\delta}{\lambda}\boldsymbol{L} \right)^{-1} \boldsymbol{G}^T \right) = \left(I + \frac{\lambda}{\delta}\boldsymbol{G}\boldsymbol{L}^{-1}\boldsymbol{G}^T \right)^{-1}. \tag{4.48}$$

So the term $a(\boldsymbol{m}_{\lambda,\delta})$ depends only on the ratio λ/δ and can be computed efficiently via a linear solve. In [Fox and Norton, 2015], they perform a similar calculation and compute a offline on a grid of λ/δ using fast Fourier transforms to avoid the costly linear solves in each step of the Markov chain.

Similarly,

$$\delta^{-1}b(\lambda,\delta) = \ln\left(\left|\det\left(\frac{\lambda}{\delta}\boldsymbol{G}^T\boldsymbol{G}+\boldsymbol{L}\right)\right|\right) \tag{4.49}$$

can also be computed offline.

4.3 Numerical Results

Finally, we implement the preceding development on synthetically derived data as well as on calibration radiographs from a high energy X-ray imaging system. The synthetic data are generated by adding simulated independent and identically distributed Gaussian measurement noise to the forward image of a PSF with a known analytic form. The radiographic data come from a large-scale diagnostic imaging system at the U.S. Department of Energy's Nevada National Security Site. In both cases, the theoretically predicted deficiency in the δ-chain is demonstrated using the statistical diagnostics introduced at the end of Chapter 3 in Section 3.2. The standard Gibbs sampler is compared to both versions of the partially collapsed Gibbs sampler derived in Chapter 4, and we investigate the trade-off between chain convergence and computational complexity in terms of number of expensive matrix factorizations. We demonstrate that even with taking into account the reduced computational complexity of the standard Gibbs sampler, partially collapsing $\boldsymbol{p}$ in the δ component improves convergence of the joint Markov chain.

4.3.1 Synthetic PSF Reconstruction

To simulate synthetic data, we reconstruct a radial profile of a two-dimensional Gaussian kernel

$$p(r) = (2\pi\sigma^2)^{-1}e^{\frac{-r^2}{2\sigma^2}}, \tag{4.50}$$

where $\sigma = \frac{1}{15}$ is chosen so that the effective width of the kernel is about 20% of the image width when scaled to $[-1, 1]$. Observe that in the case of a two-dimensional Gaussian, the action of the forward operator in (1.21) is the scaled error function

$$b(s) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{s} e^{-\frac{s'^2}{2\sigma^2}} \, ds'. \tag{4.51}$$

We synthetically add measurement error with noise strength that is 2% of the strength of the signal. For the PCG algorithms, the inner Metropolis-Hastings step was computed with $n_{mh} = 1$ and $n_{mh} = 5$, and the initial values of $\lambda_0 = 1$ and $\delta_0 = 1$ were used for each implementation.

The computed mean of the posterior density is shown in Figure 4.1 (right) in 2D, along with quantiles of the radial representations (left). The true solution is also shown on the left, showing the accuracy of the reconstruction. Note that the most uncertain region of the reconstruction are the initial discretization points corresponding to the height of the PSF, however, the true solution falls within the 90% quantiles at each point.

The MCMC diagnostics are summarized in Table 4.1. With the exception of the δ chain generated by the Gibbs sampler, each method yields Geweke p-values sufficiently large to indicate strong statistical evidence that burn-in has completed. The lack of convergence for the Gibbs sampler δ chain is the likely cause for the mean estimate of δ being marginally larger than the other three. The large autocorrelation in the δ chain generated by the Gibbs sampler results in an ESS that is significantly lower than the other three algorithms. Using #Chol/ESS as an efficiency measure, we see that PC Gibbs with $n_{mh} = 5$ is the most efficient of the four methods.

The chain autocorrelation plots in Figure 4.2 give additional insights. First, for the Gibbs sampler, the autocorrelation of the δ chain is quite large, and the λ chain is approximately uncorrelated, as values in Table 4.1 suggested would be the case. Note also that sampling

Algorithm	$\widehat{\lambda}_{\text{MCMC}}$ ($\times 10^4$)	$\widehat{\delta}_{\text{MCMC}}$ ($\times 10^{-8}$)	λ-p_{Geweke}	δ-p_{Geweke}	IACT	ESS	#Chol/ESS
Gibbs	1.102	6.132	0.998	0.850	36.2	138.0	72.4
PC Gibbs $n_{mh} = 1$	1.102	5.611	0.992	0.943	7.9	633.0	31.6
PC Gibbs $n_{mh} = 5$	1.102	5.515	0.999	0.985	1.3	3799.6	15.8
MTC	1.099	5.419	0.998	0.934	11.5	473.2	21.1

Table 4.1: Statistical diagnostics for the λ and δ chains associated with the synthetic PSF reconstruction problem. The total chain length is $M = 10^4$, with a burn-in of $k_{\text{burnin}} = 5\times 10^3$. The first column are the post-burn-in chain means of λ and δ. The maximum IACT of λ and δ are used to calculate IACT and ESS. For MTC algorithm, $\lceil (M - k_{\text{burnin}})/\tau_{\text{int}} \rceil$ is added to #Chol to evaluate the efficiency.

jointly in the MTC algorithm degrades the efficiency of λ. Finally, note that we also include autocorrelation plots for the first component, x_1, of $\boldsymbol{p}$, in order to show that the $\boldsymbol{p}$ chain is essentially uncorrelated, and hence that the correlation in the MCMC chains generated by Gibbs and PC-Gibbs samplers is driven by the δ chain.

4.3.2 PSF reconstruction from X-ray Radiographs

Next we reconstruct the point spread function of a high energy X-ray imaging system at the U.S. Department of Energy's Nevada National Security Site. The real edge data is shown in Figure 4.3 (upper left) along with a horizontal cross-section across the edge (upper right). The mean MCMC reconstruction is shown in Figure 4.3 (lower left), along with the 10%, 25%, 50%, 70%, and 90% quantiles of the chain $\boldsymbol{x}^k$. We estimated the PSF at grid points using the chain-wise mean after burn-in, $\widehat{\boldsymbol{p}} = \frac{2}{M}\sum_{k=M/2+1}^{M} \boldsymbol{p}^k$. Since the true PSF is unknown, we evaluate the accuracy of the estimation by its discrepancy; i.e. we compare forward mapping of the estimate $\mathbf{A}\widehat{\boldsymbol{p}}$ with the given data $\boldsymbol{b}$. This is shown in both linear and logarithmic scales in Figure 4.3 (lower right). In both cases the discrepancy is quite low, except at very low intensities where the data is dominated by the noise, which can be seen in the logarithmic scale.

As with the synthetic data, the PC-Gibbs algorithm with $n_{mh} = 5$ results in the largest ESS and the most efficient chain (as measured by #Chol/ESS).

Algorithm	$\widehat{\lambda}_{\text{MCMC}}$ $(\times 10^4)$	$\widehat{\delta}_{\text{MCMC}}$ $(\times 10^{-10})$	λ-p_{Geweke}	δ-p_{Geweke}	IACT	ESS	#Chol/ESS
Gibbs	9.146	1.245	0.995	0.964	14.0	357.6	28.0
PC Gibbs $n_{mh} = 1$	9.167	1.191	0.995	0.998	8.5	587.3	34.1
PC Gibbs $n_{mh} = 5$	9.178	1.189	0.994	0.980	1.5	3278.5	18.3
MTC	9.090	1.200	0.996	0.969	12.5	432.2	23.1

Table 4.2: Statistical diagnostics for the λ and δ chains associated with the measured data PSF reconstruction problem. The total chain length is $M = 10^4$, with a burn-in of $k_{\text{burnin}} = 5 \times 10^3$. The first column are the post-burn-in chain means of λ and δ. The maximum IACT of λ and δ are used to calculate IACT and ESS. For MTC algorithm, $\lceil (M - k_{\text{burnin}})/\tau_{\text{int}} \rceil$ is added to #Chol to evaluate the efficiency.

4.3.3 Conclusions

The synthetic and measured data both exhibit the theoretical degeneracy derived in Agapiou et al. [2014] for standard Gibbs sampler. We've also demonstrated that the issue is alleviated by applying the partially collapsed Gibbs framework to the Gibbs sampler, yielding two related MCMC methods, neither of which have have the δ chain correlation issues. This work provides the first, to our knowledge, successful non-parametric radial PSF reconstruction in X-ray imaging. Moreover, the theory for the modeling and algorithms has been rigorously derived from first principles, and may serve as a template for other symmetry-based prior regularization schemes. The results illustrate the effectiveness of the partially collapsed Gibbs approach, and show how a sample-based approach can be used for uncertainty quantification.

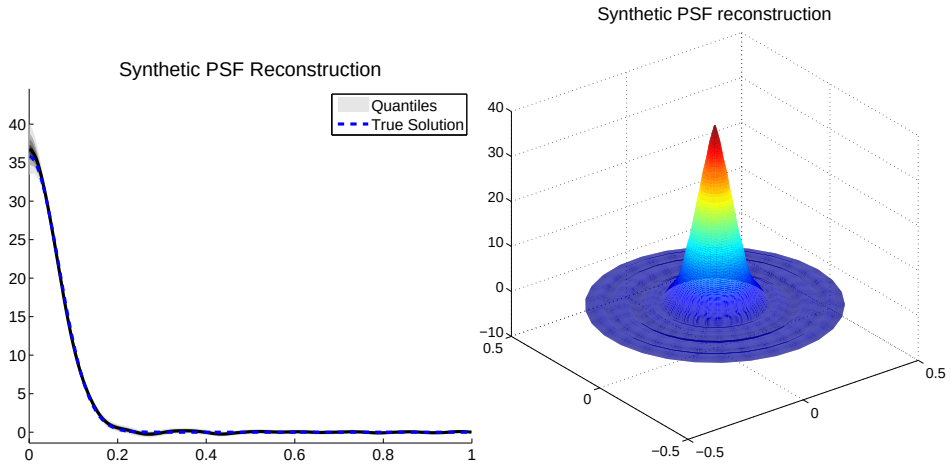

Figure 4.1: The 10%, 25%, 50%, 70%, and 90% quantiles of the reconstructed 1D radial representations of the synthetic Gaussian PSF (left) along with the mean 2D reconstruction (right).

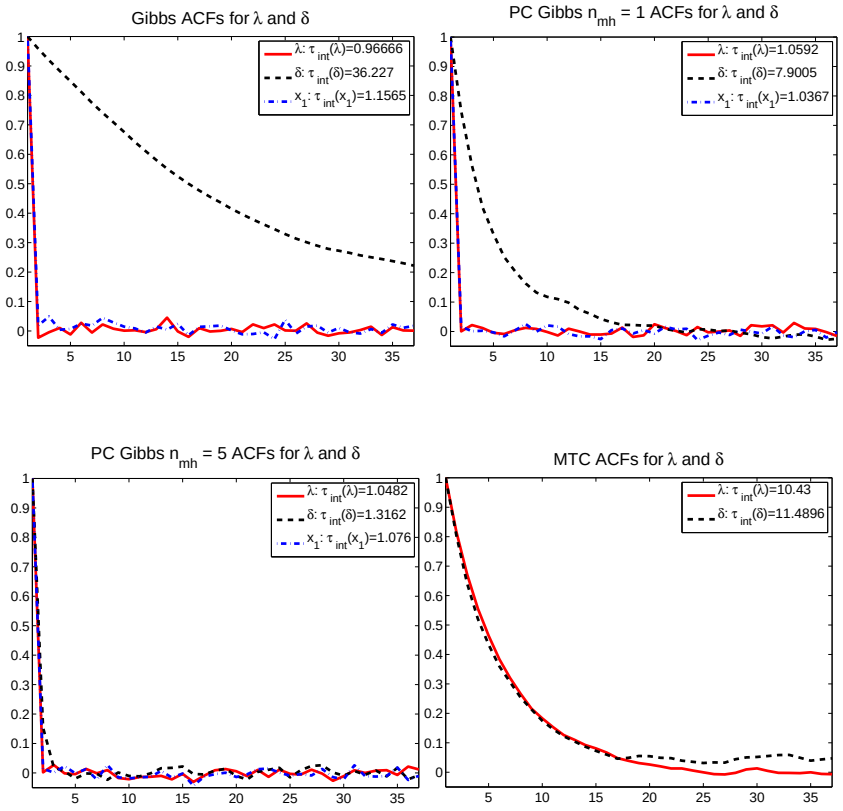

Figure 4.2: Autocorrelation plots for PSF reconstruction for synthetic data of the chains λ, δ and the central discretization point of $\boldsymbol{p}$: in the upper-left are the ACF for MCMC chains of λ, δ and central pixel of the radial profile for the Gibbs sampler; on the upper-right are the plots for the PC Gibbs sampler with 1 inner MH step; on the lower-left are plots for the PC Gibbs sampler with 5 inner MH steps; and in the lower-right are plots for the MTC sampler.

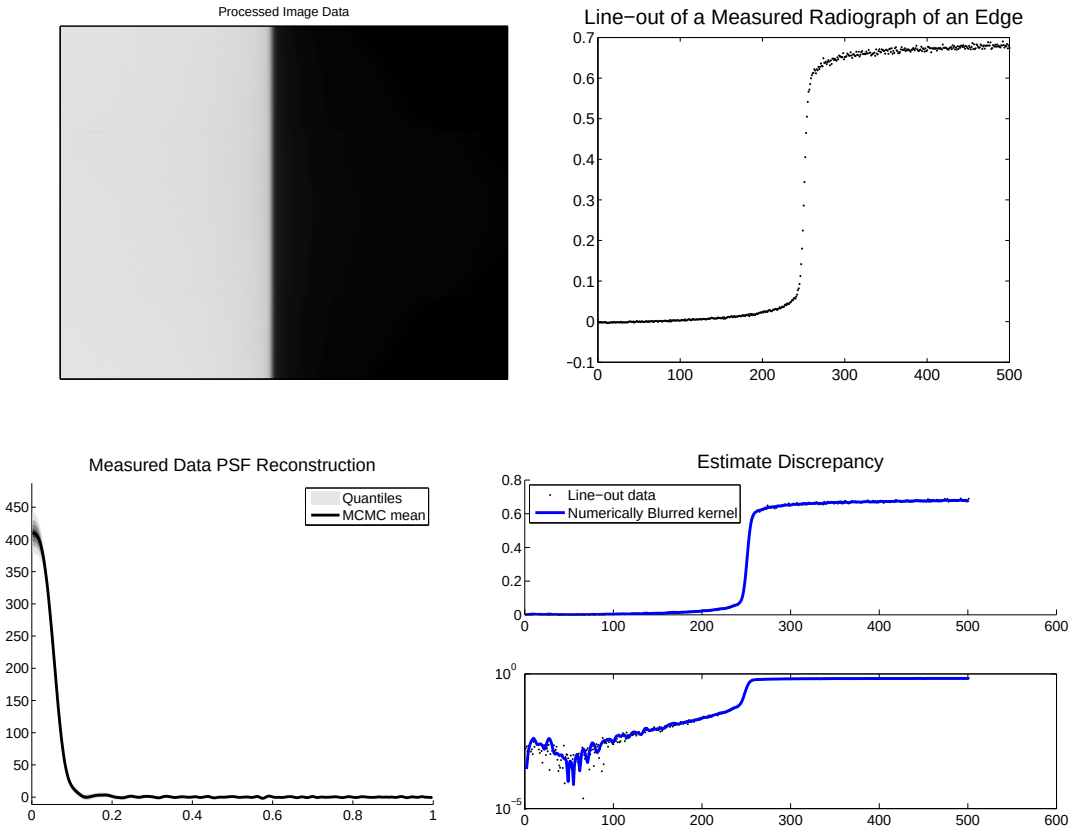

Figure 4.3: PSF reconstructions for radiographic data: in the upper left corner are the radiographic image data; in the upper right corner is a line-out taken from the image data; in the lower left corner are the central 10%, 25%, 50%, 70%, and 90% quantiles of the posterior reconstruction of $\boldsymbol{x}$ for each pixel; in the lower right corner are plots of the forward mapped discrepancy of the post burn-in chain mean.

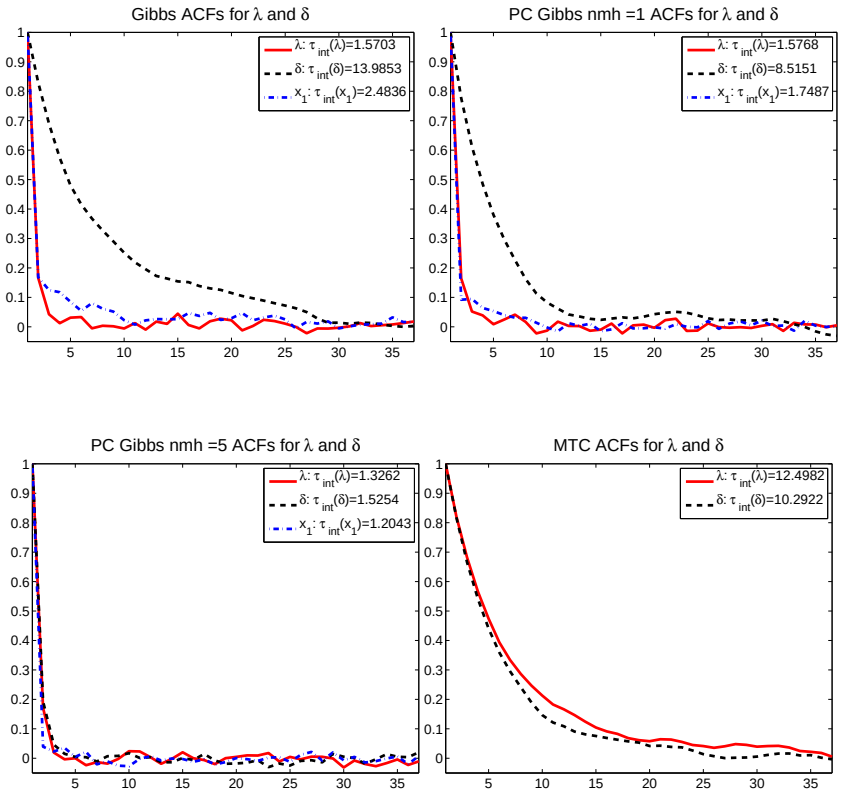

Figure 4.4: Autocorrelation plots for PSF reconstruction for the measured data of λ and δ chains: in the upper-left are the ACF for MCMC chains of λ, δ and central pixel of the radial profile for the Gibbs sampler; on the upper-right are the plots for the PC Gibbs sampler with 1 inner MH step; on the lower-left are plots for the PC Gibbs sampler with 5 inner MH steps; and in the lower-right are plots for the MTC sampler.

Bibliography

Sergios Agapiou, Johnathan M Bardsley, Omiros Papaspiliopoulos, and Andrew M Stuart. Analysis of the Gibbs sampler for hierarchical inverse problems. *SIAM/ASA Journal on Uncertainty Quantification*, 2(1):511–544, 2014.

George Bachman and Lawrence Narici. *Functional Analysis*. Academic Press, 1966.

Johnathan M Bardsley. MCMC-based image reconstruction with uncertainty quantification. *SIAM Journal on Scientific Computing*, 34(3):A1316–A1332, 2012.

Johnathan M Bardsley and Aaron Luttman. Dealing with boundary artifacts in MCMC-based deconvolution. *Linear Algebra and its Applications*, 473:339–358, 2015.

Johnathan M Bardsley and Aaron Luttman. A Metropolis-Hastings method for linear inverse problems with Poisson likelihood and Gaussian prior. *International Journal of Uncertainty Quantification*, 2016.

Johnathan M Bardsley, Marylesa Howard, and James G Nagy. Efficient MCMC-based image deblurring with Neumann boundary conditions. *Electronic Transactions on Numerical Analysis*, 40:476–488, 2013.

Patrick Billingsley. *Probability and Measure*. John Wiley & Sons, 2008.

R. Bracewell. *The Fourier Transform and Its Applications*. McGraw-Hill, 1965.

Daniela Calvetti and Erkki Somersalo. *An Introduction to Bayesian Scientific Computing: Ten Lectures on Subjective Computing*, volume 2. Springer Science & Business Media, 2007.

Noel AC Cressie. *Statistics for Spatial Data*. Wiley New York, 1993.

Edward R. Doering, Joseph Gray, and John P. Basart. Point spread function estimation of image intensifier tubes. *Review of Progress in Quantitative Nondestructive Evaluation*, 11: 323–329, 1992.

Richard Durrett. *Probability: Theory and Examples*. Cambridge university press, 2010.

F.W. Dyson, A.S. Eddington, and Davidson C. A determination of the deflection of light by the sun's gravitational field, from observations made at the total eclipse of 29 May 1919. *Philosophical Transactions of the Royal Society*, 220A:291–333, 1920.

Charles L. Epstein. *Introduction to the Mathematics of Medical Imaging*. Society for Industrial and Applied Mathematics, 2008. ISBN 9780898716429.

Michael J Fowler, Marylesa Howard, Aaron Luttman, Stephen E Mitchell, and Timothy J Webb. A stochastic approach to quantifying the blur with uncertainty estimation for high-energy X-ray imaging systems. *Inverse Problems in Science and Engineering*, 24(3):353–371, 2016.

Colin Fox and Richard A Norton. Fast sampling in a linear-Gaussian inverse problem. *arXiv preprint arXiv:1507.01614*, 2015.

Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian Data Analysis*, volume 2. Taylor & Francis, 2014.

Stuart Geman and Donald Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (6):721–741, 1984.

John Geweke. *Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments*, volume 196. Federal Reserve Bank of Minneapolis, Research Department Minneapolis, MN, USA, 1991.

Gene H Golub and Charles F Van Loan. *Matrix Computations*, volume 3. JHU Press, 2012.

L. Grafakos. *Classical Fourier Analysis*. Graduate Texts in Mathematics. Springer New York, 2014. ISBN 9781493911943.

David H Griffel. *Applied Functional Analysis*. Courier Corporation, 2002.

Heikki Haario, Eero Saksman, and J. Tamminen. An adaptive Metropolis algorithm. *Bernoulli*, 7:223–224, 2005.

Jacques Hadamard. Sur les problmes aux drives partielles et leur signification physique. *Princeton University Bulletin*, pages 49–52, 1902.

Martin Hanke and Otmar Scherzer. Inverse problems light: Numerical differentiation. *The American Mathematical Monthly*, 108(6):512–521, 2001.

P. C. Hansen. *Discrete Inverse Problems: Insight and Algorithms*. SIAM, Philadelphia, 2010.

Dave Higdon. A primer on space-time modeling from a Bayesian perspective. *Monographs on Statistics and Applied Probability*, 107:217, 2006.

Lars Hörmander. *The Analysis of Linear Partial Differential Operators I*. Springer-Verlag, 1983.

Marylesa Howard, Michael Fowler, Aaron Luttman, Stephen Mitchell, and Margaret C. Hock. Bayesian Abel inversion in qualitative X-ray radiography. *SIAM Journal on Scientific Computing*, 2016.

Ernst Ising. Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik A Hadrons and Nuclei*, 31(1):253–258, 1925.

A.K. Jain. *Fundamentals of Digital Image Processing*. Prentice-Hall information and system sciences series. Prentice Hall, 1989.

J. Kaipio and E. Somersalo. *Statistical and Computational Methods for Inverse Problems*. Springer, 2005.

O. Knill, R. Dgani, and M. Vogel. A new approach to Abel's integral operator. *Astronomy and Astrophysics*, 274:1002–1008, 1993.

Donald E Knuth. *The Art of Computer Programming. Volume 2: Seminumerical Algorithms*. Addison-Wesley, Reading, MA, 2 edition, 1982.

D. Kundur and D. Hatzinakos. Blind image deconvolution. *IEEE Signal Processing Magazine*, 13(3):43–64, 1996.

Jun S Liu. *Monte Carlo Strategies in Scientific Computing*. Springer Science & Business Media, 2008.

L.N. Magner. *History of the Life Sciences*. New York: Marcel Dekker, Inc, 2002.

Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

Vladimir Aleseevich Morozov and Michael Stessin. *Regularization Methods for Ill-Posed Problems*. CRC Press, 1993.

Keith W Morton and David Francis Mayers. *Numerical Solutions of Partial Differential Equations: An Introduction*. Cambridge university press, 2005.

Peter Müller. Alternatives to the Gibbs sampling scheme. 1992.

Robert D. Richtmyer. *Principles of Advanced Mathematical Physics: Volume I*. Springer-Verlag, 1978.

Christian Robert and George Casella. *Monte Carlo Statistical Methods*. Springer Science & Business Media, 2013.

Walter Rudin. *Functional analysis*. McGraw-Hill, Inc., New York, 1991.

Havard Rue and Leonhard Held. *Gaussian Markov Random Fields: Theory and Applications*. CRC Press, 2005.

Robert T Seeley. Extension of C^∞ functions defined in a half space. *Proceedings of the American Mathematical Society*, 15(4):625–626, 1964.

MR Showalter, MK Gordon, and D Olson. VG1/VG2 Saturn ISS processed images V1.0. NASA Planetary Data System, 2006.

A Sokal. *Monte Carlo Methods in Statistical Mechanics: Foundations and New Algorithms*. Springer, 1997.

Robert S Strichartz. *The Way of Analysis*. Jones & Bartlett Learning, 2000.

Robert S Strichartz. *A Guide to Distribution Theory and Fourier Transforms*. World Scientific, 2003.

A. M. Stuart. Inverse problems: A Bayesian perspective. *Acta Numerica*, 19:451–559, 2010.

Luke Tierney. Markov chains for exploring posterior distributions. *the Annals of Statistics*, pages 1701–1728, 1994.

A.N. Tikhonov. Regularization of incorrectly posed problems. *Soviet Mathematics Doklady*, 4:1624–1627, 1963.

Austin B Tomaney and Arlin PS Crotts. Expanding the realm of microlensing surveys with difference image photometry. *arXiv preprint astro-ph/9610066*, 1996.

David A Van Dyk and Xiyun Jiao. Metropolis-Hastings within partially collapsed Gibbs samplers. *Journal of Computational and Graphical Statistics*, 24(2):301–327, 2015.

David A Van Dyk and Taeyoung Park. Partially collapsed Gibbs samplers: Theory and methods. *Journal of the American Statistical Association*, 103(482):790–796, 2008.

Curtis R. Vogel. *Computational Methods of Inverse Problems*. Society for Industrual and Applied Mathematics, 2002.

Scott A. Watson. Real-time spot size measurement for pulsed high-energy radiographic machines. *Proceedings of the 1993 Particle Accelerator Converence*, 4:2447–2449, 1993.

Max A Woodbury. Inverting modified matrices. *Memorandum report*, 42:106, 1950.