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 Golobstov
Department of Physics
Lomonosov Moscow State University

Dr. Leonid Kalechev
Mathematical Sciences

Joyce, Kevin, Doctorate of Philosophy, May 2016 Mathematics

Point Spread Function Estimation and Uncertainty Quantification

Committee Chair: Johnathan Bardsley, Ph.D.

Acknowledgments

Notations

Image Model

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

Sets

$\mathbb{R}$ ..... Real numbers
$\mathbb{R}^k$ ..... k-dimensional vector space over real numbers
$\boldsymbol{x}_{\widehat{i}}$ ..... The vector $(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_k)$

Functional Analysis

$\mathscr{H}$ ..... Hilbert space
$\|\cdot\|$ ..... Norm
$\langle \cdot, \cdot \rangle$ ..... Inner product
$\mathscr{X}^*$ ..... Dual space of $\mathscr{X}$
$\mathscr{D}$ ..... Compactly supported smooth test functions
$\mathscr{D}^*$ ..... Space of distributions
$\mathscr{K}$ ..... Space of radially symmetric functions
$\mathscr{P}$ ..... 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 norm

Probability

$\mathbb{P}$ ..... Probabilty measure
$\mathbb{E}$ ..... Probabilty expectation
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 data from projections of light) are an important source of scientific information. Even the word "observation" generally connotes the 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 data to discover and explain natural phenomena; for example, visual observations such as the color and shape of various plant organs in Gregor Mendel's famous experiments on hybridized peas formed the primary source of data for developing his famous model for genetic inheritance [Magner, 2002]. Arthur Eddington's famous 1919 image of the gravitationally lensed path of a comet during a solar eclipse [Dyson et al., 1920] provided the first experimental evidence supporting Albert Einstein's general theory of relativity. Yet, 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

became possible, allowing the potential to *quantitatively* analyze visual information. The ever-progressing technology in optical science and engineering are rapidly increasing the amount of data that can be measured in an image; yet, the extreme quantity and spatial organization that comes with high resolution data makes a rigorous quantitative analysis quite challenging.

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 domain of interest for imaging systems can vary greatly in both spatial dimension and the spectrum of light captured. The energy response of light is spectrum dependent, and typically is analyzed at fixed bands of frequency with very high spatial resolution. 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 attenuation of electromagnetic radiation are used to reconstruct a cross-section of a scanned object [Epstein, 2008]. The primary focus in this work are pulsed X-ray measurements, referred to as radiographs, that are used as an experimental diagnostic of high-energy physics experiments. In this case, X-rays are pulsed through an experiment, 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 measured on a high resolution array (on the order of 1000×1000 or more) of carge coupled devices (CCD) calibrated to count photons at a specified spectral band.

Having established the form of image data, i.e. *what* the data is, *how* does one analyze it? Although images often provide a large volume of data, their spatial nature make measurements at each pixel highly dependent on measurments of adjacent pixels. 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.

From a statistical point of view, the field of spatial statistics considers broadly the analysis of data that is structured and spatially correlated in this way. This field has had 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 have considered image data, or more generally spatially correlated data, in one way or another; fields such as astronomy, astrophysics, biology, medicine, geology, computer science, and nuclear physics to name a few. The book [Cressie, 1993] provides an excellent overview of the history and current methods for statistical methods for spatial data. Although much work has been done, it is still a very active research area and is far from 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 a small component of image analysis related specifically to the *system for capturing images*. Understanding this component is an important preliminary step to developing methods for quantitatively analyzing the images themselves.

1.1.1 Organization

The following sections outline the modeling of image blur via the *point spread function*, the primary quantity of interest in this work. In particular, we will introduce a model for the blurring of an opaque edge. The model can be cast in various mathematical forms without changing the physical assumptions. We will exploit these different mathematical formulations (e.g. by changing variables) to show that a blurred edge is sufficient for estimation and to show that the problem is "ill-posed".

In the next chapter, we will derive the necessary theory and technical definitions so that the problem is well-defined on a Hilbert space. Chapter 2 will be mainly theoretical, but the explicit forms for the discrete model of the forward operator and prior information are motivated and derived there. Chapter 3 will give details on how to carry out the estimation

on a computer. There, we will deal with how to discretely represent each of the necessary components in the estimation problem. We will also motivate and present the design of a detailed algorithm for carrying out statistical estimation. Finally, in Chapter 4, we present the results of an implementation on synthetically derived data and on measured data 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 the ideal image f to 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. Informally, we can view the effect of blur point-wise by observing the system response of a "point-source" at $(\overline{x},\overline{y})$ (formally, take $f = \delta_{\overline{x},\overline{y}}$, Dirac's delta translated to $(\overline{x},\overline{y})$), then $b(x,y) = k(x,y;\overline{x},\overline{y})$ represents the "spread" 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 f by $(\overline{x},\overline{y})$

results in b translated by $(\overline{x}, \overline{y})$, we say that it is *spatially invariant*. In (1.1), 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

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

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

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

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

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

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, a general result about arbitrary linear filters defined on L^p spaces states that the action of any linear filter can be expressed through convolution with some generalized function k [Grafakos, 2014]. In any case, when blur is assumed to be spatial 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.

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 complex imaging systems such as the one described for X-ray radiography, this is not realistic. Instead, if 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 beam provides a good point-source estimate. 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 at these frequencies. Instead, the system response of a uniformly opaque calibration object with a simple geometry can be measured. Under the assumption that the object is sufficiently thick so that X-rays are completely attenuated on the profile of the object, then f 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]. An additional assumption of *radial symmetry* on k, will allow for a very simple calibration object – an edge. That is, 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 $f(s,t) \stackrel{\text{def}}{=} f_E(s) = 1$ if $s \geq 0$ and 0 if $s < 0$; see Figures 1.1 and 1.2 for a schematic of the calibration object and measurement system and 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 $b(x,0) = b(x)$, (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. So given any solution k, $k(s,t) + te^{-t^2-s^2}$ is also a solution but is *not* radially symmetric. We will see in the next section that the assumption of radial symmetry on k is sufficient for a well-defined analytic solution to (1.9).

In summary, we've assumed that the effect of blur is modelled by a spatially invariant linear filter with a radially symmetric kernel, and (1.9) describes the blurring of a calibration object whose profile is an edge. Solving the integral equation in (1.9) will be the primary focus of this work.

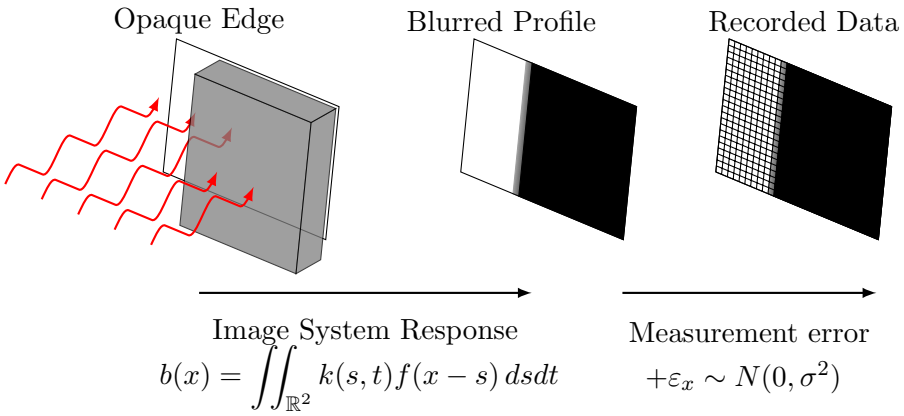

Figure 1.1: A schematic of the measurement model for an X-Ray image of an edge. An opaque block aligned with the imaging plane blocks light on the half plane to produce a blurred edge.

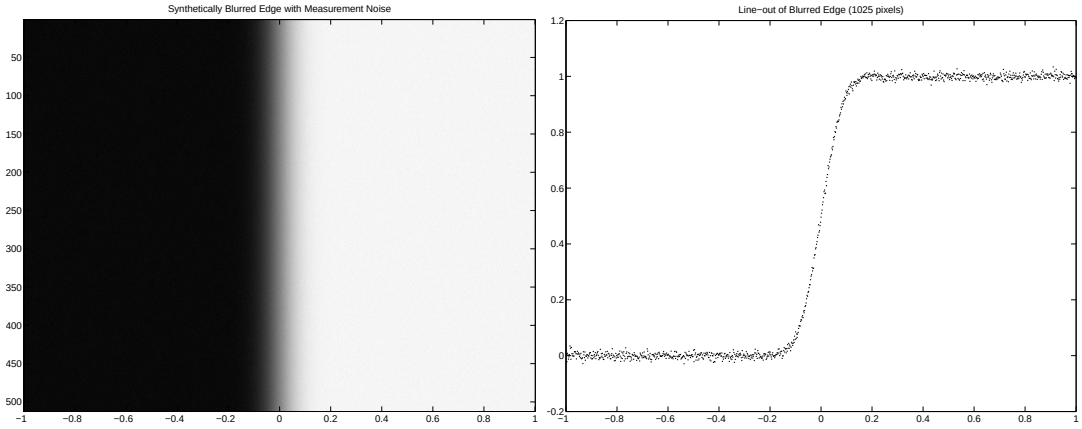

Figure 1.2: A synthetically blurred edge with simulated measurement error and a line-out (horizontal cross-section) from the data.

1.3 The Abel transform and a deterministic solution

In this section, we will show that radial symmetry is sufficient to guarantee a unique analytic answer. However, the suggested analytic method will be problematic as will be seen in the following section.

A large part of designing a system for imaging is to minimize the effect of blur. 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, engineering effort can still optimize accuracy, or equivalently, minimize bias. Hence, an optimally designed imaging system should exhibit isotropic blur (an absence of direction bias), so that, in the convolution model, the PSF is radial 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 radially symmetric, then it has a unique one-dimensional representation; that is, there exists a function p defined on $[0, \infty)$ 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. Viewing (1.9) as iterated integration first in t allows $f_E(s-x)$ to be factored out of the inner integral. Then substituting p for k, the inner integral has the form

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

The function $\ell(s)$ is the integration along a line perpendicular to the edge E, and its form is 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*, and for its study in imaging science, see [Bracewell, 1965, Epstein, 2008, Knill et al., 1993].

Substituting (1.11) for k into (1.9) and changing the bounds of integration in s according

to $f_E(x-s)$, results in

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

Observe that b exhibits point symmetry about $b(0)$. To see this analytically, note that when $x > 0$

$$\begin{aligned} b(x) - \int_{-\infty}^{0} \ell(s)ds &= \int_{0}^{x} \ell(s)ds \\ &= \int_{0}^{-x} \ell(s')ds' \\ &= b(-x) - b(0). \end{aligned} \tag{1.13}$$

Since the function $\widetilde{b}(x) = b(x) - b(0)$ is odd, $b(x)$ has reflection symmetry about $(0, b(0))$. Hence, data defined on either $x \in (-\infty, 0]$ or $x \in [0, \infty)$ should be sufficient for estimating p. 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 $\lim_{r\to\infty} rp(r) = 0$ and $\ell(s)$ in (1.11) is pointwise 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

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

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

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

Note that the resulting integral is independent of both r and t, and hence, is constant. To evaluate it, recall the Gamma function identity

$$\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)} = \int_0^1 \tau^{-\alpha}(1 - \tau)^{\alpha - 1}\, d\tau, \tag{1.19}$$

from which the expression in (1.18) reduces to $\Gamma(1/2)^2 = \pi$. 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

$$
\begin{aligned}
-\frac{\pi}{r}\frac{d}{dr}\left(\int_r^\infty \frac{\ell(s)sds}{(s^2-r^2)^{1/2}}\right) &= -\frac{\pi}{r}\frac{d}{dr}\int_r^\infty p(t)t\pi dt \\
&= \frac{1}{r}\left(p(r)r - \lim_{T\to\infty} p(T)T\right) \\
&= p(r), \qquad (1.20)
\end{aligned}
$$

which proves the identity in (1.14). □

With this result, p can be analytically 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 $p(r)$. Hence, the assumption of radial symmetry 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 noise-free model that has not been addressed – random effects due to measurement error – for which a direct application of 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 analytic solution outlined in the last section will not be sufficient when measurement errors are introduced. One way to see that the solution method outlined will be insufficient 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 modelling the stochastic effect of measurement error is system dependent and can be quite complicated. In the X-ray radiography example, uncertainty can enter into the system at the luminescing crystal response, at the counts of 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 in the early 20th century in [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 these *ill-posed* problems. In particular, most cases of interest exhibit a model where the inverse of the forward operator is discontinuous.

The discussion thus far for PSF reconstruction has been somewhat informal, as we have not defined a space for the PSF or its radial representation, so we have not formally defined the forward operator of the model. Defining these spaces in detail is technical and will be addressed in in Chapter 2; however, assuming these spaces have been defined, we can illustrate that the problem of reconstructing the radial profile is ill-posed.

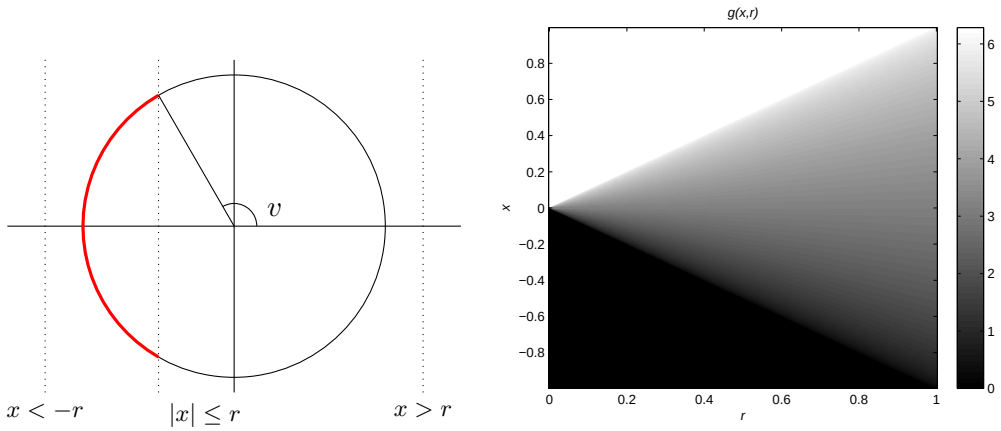

Figure 1.3: The PSF forward integral operator kernel $g(x, r)$ represented as the arc measure of v in $(-\pi, \pi)$ where $x \geq r \cos v$.

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

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

where

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

To see that g has this form, note that integrating $f_E(x - r\cos v)$ is the radian measure of the set $\{v \in (-\pi, \pi) : r \cos v \leq x\}$; 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 $b(0)$. So, $b(x)$ defined on either $(-\infty, 0]$ or $[0, \infty)$ completely determines p.

With these observations, we can define $G : \mathcal{H}_1 \to \mathcal{H}_2$ is an operator between closed isometric subspaces of $L^2([0, \infty))$ that acts by the integral equation in (1.21), i.e.,

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

Hence, G is a compact Hilbert-Schmidt operator since g is continuous and $\mathcal{H}_1$ and $\mathcal{H}_2$ are separable Hilbert spaces. Moreover, G is an injective operator since we showed that $b = Gp$ has an analytic solution. The spectral theorem for such operators implies that G has a countable spectrum which has zero as a limit point, and hence, its inverse G^{-1} is unbounded. See one of many texts on functional analysis [Bachman and Narici, 1966, Rudin, 1991] and [Tikhonov, 1963, Vogel, 2002, Morozov and Stessin, 1993] for the analytic treatment of ill-posed problems.

Solving ill-posed inverse problems requires *regularization* of 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. In particular, we will take a Bayesian approach to the inverse problem so that, in addition to estimating k, uncertainties in the estimate 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 spaces $\mathcal{H}_1$ and $\mathcal{H}_2$ have been defined, which we address in the next chapter.

Chapter 2

Reconstruction on the Continuum

The goal of this section is to develop the necessary mathematical tools to encapsulate prior notions of smoothness and radial symmetry within the structure of appropriately developed Hilbert spaces. This is done within the framework of distributions from linear PDE theory Hörmander [1983], Rudin [1991], Griffel [2002]. We will first establish the relevant background from that theory. Then, we define symmetry in terms of pre-composition with a smooth map that is many-to-one on circles. This defines a pullback mapping that will ultimately establish the space of radially symmetric functions, the space of their radial representations, and an isometry between them.

2.0.1 Distributional Hilbert Spaces

We first establish notation and preliminaries from the theory of distributions. For a complete development, see one of various classical texts on the subject, such as Hörmander [1983], Rudin [1991].

Let $\mathscr{D}(\Omega)$ denote compactly supported smooth test functions defined on an open set Ω. We

adopt the notation $\langle f, \phi \rangle$ for the action of a linear functional f on $\phi \in \mathscr{D}(\Omega)$. The space of continuous linear functionals, or distributions, on $\mathscr{D}(\Omega)$ are denoted with $\mathscr{D}^*(\Omega)$. Continuity is in the sense that for every compact set $K \subset \Omega$, there exists C and k so that

$$|\langle u, \phi \rangle| \leq C \sum_{|\alpha| \leq k} \sup |\partial^\alpha \phi|, \tag{2.1}$$

for any multi-index α. A functional f is continuous if and only if for every sequence $\{\phi_j\}$ in $\mathscr{D}(\Omega)$ converging to zero in the sense that for every fixed α, $\sup |\partial^\alpha \phi_j| \to 0$ with some fixed compact $K \subset \Omega$, such that $\operatorname{supp} \phi_j \subseteq K$ for all j, then $\langle f, \phi_j \rangle \to 0$ Hörmander [1983]. We 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.

The space of square Lebesgue integrable functions on Ω, $L^2(\Omega)$, embeds naturally into $\mathscr{D}^*(\Omega)$. This is done by taking sequences of test functions $\{\phi_n\}$ such that $\int \phi_n \overline{\phi_n} \, dx$ defines a Cauchy sequence in $\mathbb{R}^+$, and defining $\langle \phi, \psi \rangle := \lim_{n \to \infty} \langle \phi_n, \psi \rangle$. It can be shown that $\phi \in D^*(\Omega)$ and that sequences of test functions are sufficient for representing all elements of $L^2(\Omega)$. Moreover, it can be shown that

$$(\phi, \psi)_{L_2(\Omega)} := \lim_{n \to \infty} \int \phi_n, \psi_n \, dx \tag{2.2}$$

defines an inner product agrees with the usual L^2 inner product.

In a similar way, the Sobolev space $\mathscr{H}^m$ embeds in $\mathscr{D}^*(\Omega)$ by considering Cauchy functions with respect to inner product $\sum_{|\alpha| \leq m} (\partial^\alpha \cdot, \partial^\alpha \cdot)_{L^2}$ where the sum is over all multi-indices α. These spaces form a sequence of closed subspaces $L^2(\Omega) = \mathscr{H}^0(\Omega) \supset \mathscr{H}^1(\Omega) \supset \cdots \supset \mathscr{H}^m(\Omega) \supset \cdots \supset \mathscr{D}(\Omega)$ Hutson and Pym [1980]. In our development, we will be primarily concerned with open subsets of $\mathbb{R}^2$ and $\mathbb{R}$, and we denote such subsets as Ω_2 and Ω_1 respectively.

A standard result from distribution theory [Hörmander, 1983, Theorem 4.1.5] states

Theorem 2.0.1. *For Ω open in $\mathbb{R}^n$, if $f \in \mathscr{D}^*(\Omega)$, then there is a sequence of test functions $\{f_n\} \subset \mathscr{D}(\Omega)$ such that $\langle f_n, \phi\rangle \to \langle f, \phi\rangle$ for all $\phi \in \mathscr{D}(\Omega)$.*

Theorem 2.0.1 is used in Hörmander [1983] to establish many of the computation rules involving derivatives and, in particular, the various chain rules associated with differential forms are established on $\mathscr{D}(\Omega)$ and then extended continuously to $\mathscr{D}^*(\Omega)$ by a computation that expresses the action on test functions, such as integration by parts. This is the general strategy we will use to establish many of the results in this section.

2.0.2 Radial Symmetry via The Pullback Operator

Symmetry is established by casting it in terms of a many-to-one smooth map, T, that is constant on circles of a fixed radius. When f is a function, it has radial symmetry if there exists a function ρ, so that $f = \rho \circ T$. This notion is easily adapted to distributions by developing the corresponding linear pullback operator to T that maps ρ to f. In this subsection, we will explicitly construct T and show that it establishes a one-to-one correspondence. Using this correspondence, we will define the space of radially symmetric distributions and the space of their radial representations.

Theorem 2.0.2. *Let $\Omega_1 := (0, \infty) \subset \mathbb{R}$ and $\Omega_2 := \mathbb{R}^2 \setminus \{x = 0 \text{ or } y = 0\}$. For $h : \Omega_1 \to \Omega_1$ an increasing diffeomorphism, define $T : \Omega_2 \to \Omega_1$ by $T(x, y) = h(x^2 + y^2)$, then there exists a unique, injective, continuous, linear operator $T^\sharp : \mathscr{D}^*(\Omega_1) \to \mathscr{D}^*(\Omega_2)$ so that $\langle T^\sharp \rho, \phi\rangle = \langle \rho \circ T, \phi\rangle$ for all $\phi \in \mathscr{D}(\Omega_2)$ and $\rho \in \mathscr{D}(\Omega_1)$.*

To prove this result, we first establish a straight-forward, yet technical, computation in the following lemma.

Lemma 2.0.3. *For each $\phi \in \mathscr{D}(\Omega_2)$ there exists $\psi_\phi \in \mathscr{D}(\Omega_1)$ so that for any $\rho \in \mathscr{D}(\Omega_1)$*

$$\langle \rho \circ T, \phi\rangle_{\Omega_2} = \langle \rho, \psi_\phi\rangle_{\Omega_1}. \tag{2.3}$$

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

$$T_{ij}(x,y) = \Big(T(x,y), (-1)^j y\Big). \tag{2.4}$$

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

$$T_{ij}^{-1}(r,t) = \Big((-1)^i \sqrt{h^{-1}(r) - t^2}, (-1)^j t\Big), \tag{2.5}$$

and

$$\left|dT_{ij}^{-1}(r,t)\right| = \frac{1}{2} h^{-1\prime}(r) \left(h^{-1}(r) - t^2\right)^{-1/2}. \tag{2.6}$$

Furthermore, note that $T \circ T_{ij}^{-1}(r,t) = r$ by the definition of T_{ij}. Now, given $\rho \in \mathscr{D}(\Omega_1)$, a change of variables results in

$$\begin{aligned}
\langle \rho \circ T, \phi \rangle_{\Omega_2} &= \sum_{ij} \iint_{Q_{ij}} \rho \circ T(x,y) \cdot \phi(x,y) dx dy \\
&= \sum_{ij} \iint_R \rho(r) \cdot \phi \circ T_{ij}^{-1}(r,t) \left|dT_{ij}\right| dr dt \\
&= \int_0^\infty \rho(r) \left(\int_0^{\sqrt{h^{-1}(r)}} \sum_{ij} \phi \circ T_{ij}^{-1}(r,t) \left|dT_{ij}\right| dt \right) dr.
\end{aligned} \tag{2.7}$$

Let

$$\psi_\phi(r) = \int_0^{\sqrt{h^{-1}(r)}} \sum_{ij} \phi \circ T_{ij}^{-1}(r,t) \left|dT_{ij}\right| dt, \tag{2.8}$$

and we must show that $\psi_\phi \in \mathscr{D}(\Omega_1)$. Note that $\operatorname{supp}\left(\phi \circ T_{ij}^{-1}\right) = T_{ij}(\operatorname{supp}\phi) \subseteq R$, and since T_{ij} is a diffeomorphism, $\phi \circ T_{ij}^{-1} \in \mathscr{D}(\Omega_2)$. Since $\sum \phi \circ T_{ij}^{-1}(r,t)$ is smooth and compactly supported, a standard result from analysis [Strichartz, 2000, pg. 433] implies that $\psi_\phi(r)$ resulting from integrating in t is a smooth function whose support is the projection of the support of the integrand. □

We can now proceed to prove Theorem 2.0.2.

Proof. Using Lemma 2.0.3, define

$$\langle T^\sharp \rho, \phi \rangle_{\Omega_2} := \langle \rho, \psi_\phi \rangle_{\Omega_1}. \tag{2.9}$$

To see that $T^\sharp \rho \in \mathscr{D}^*(\Omega_2)$, let $\{\phi_n\} \to 0$ in $\mathscr{D}(\Omega_2)$, so in particular (for $\alpha = 0$), $\sup |\psi_n| \to 0$. Then, by (2.8), $\sup |\phi_{\psi_n}| \to 0$, and thus $\langle \rho, \psi_{\phi_n} \rangle \to 0$. The linearity and continuity of $T^\sharp$ follow directly from the definition. That is

$$\langle T^\sharp \rho_1 + T^\sharp \rho_2, \phi \rangle_{\Omega_2} = \langle \rho_1, \psi_\phi \rangle_{\Omega_1} + \langle \rho_2, \psi_\phi \rangle_{\Omega_1} = \langle T^\sharp(\rho_1 + \rho_2), \psi_\phi \rangle_{\Omega_2}, \tag{2.10}$$

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

$$\langle T^\sharp \rho_n, \phi \rangle_{\Omega_2} = \langle \rho_n, \psi_\phi \rangle_{\Omega_1} \to 0. \tag{2.11}$$

Uniqueness is a consequence of Theorem 2.0.1. That is, suppose $T^\dagger : \mathscr{D}^*(\Omega_1) \to \mathscr{D}^*(\Omega_2)$ is a continuous linear functional such that $\langle T^\dagger \rho, \phi \rangle = \langle \rho \circ T, \phi \rangle$ for all $\phi \in \mathscr{D}(\Omega_2)$ whenever $\rho \in \mathscr{D}(\Omega_1)$. Then, for any $\rho \in \mathscr{D}^*(\Omega_1)$, let $\{\rho_n\} \subset \mathscr{D}(\Omega_1)$ converge to ρ (in the $\mathscr{D}^*(\Omega_1)$ sense), so

$$\Big\langle (T^\sharp - T^\dagger)\rho, \phi \Big\rangle_{\Omega_2} = \lim \langle T^\sharp \rho_n, \phi \rangle_{\Omega_2} - \lim \langle T^\dagger \rho_n, \phi \rangle_{\Omega_2} = 0. \tag{2.12}$$

Hence $T^\sharp = T^\dagger$. We have established all of the properties in Theorem 2.0.2. □

Loosely speaking, the pullback by T represents a change of variables from (x, y) to (r, v) by extending expanding the domain of T to an invertible $T_{ij}(x, y)$ with the choice of T_{ij} somewhat arbitrary. The uniqueness allows us to freely choose any other change of variables such that $T \circ T_{ij}(r, v) = r$, and the analysis on T is still valid. I.e., another valid choice of T_{ij} is similar to polar-coordinates transformation $(T(x, y), \mathrm{Arg}(x, y))$. Our choice was such that the analysis

is straight-forward, although we will make use of the polar-coordinate variable transformation later to define the forward operator on linear representations.

The construction can be carried out much more generally for any smooth T and is outlined in Hörmander [1983]. 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 2.0.4. *For all $\rho \in \mathscr{D}^*(\Omega_1)$ and $\psi \in \mathscr{D}(\Omega_1)$, if we denote the $h^{-1'}$ as the derivative of the inverse of h in Theorem 2.0.2 then*

$$\langle T^\sharp \rho, \psi \circ T \rangle_{\Omega_2} = \pi \langle \rho, \psi \cdot h^{-1'} \rangle_{\Omega_1}. \tag{2.13}$$

Proof. First, note that both $\psi \circ T$ and $h^{-1'} \cdot \psi$ are elements of $\mathscr{D}(\Omega_1)$. From (2.6), observe

$$\begin{aligned} \int_0^{\sqrt{h^{-1}(r)}} |dT_{ij}(r,t)| dt &= \frac{h^{-1'}(r)}{2} \int_0^{\sqrt{h^{-1}(r)}} \left(h^{-1}(r) - t^2\right)^{-1/2} \\ &= \frac{\pi}{4} h^{-1'}(r). \end{aligned} \tag{2.14}$$

Invoking Theorem 2.0.1, let $\{\rho_n\}$ be a sequence in $\mathscr{D}(\Omega_1)$ converging to ρ in $\mathscr{D}^*(\Omega_1)$, then substituting ρ_n for ρ and $\psi \circ T$ for ϕ in (2.7), we have

$$\begin{aligned} \langle T^\sharp \rho_n, \psi \circ T \rangle_{\Omega_2} &= 4 \int_0^\infty \rho_n(r)\psi(r) \left(\int_0^{\sqrt{h^{-1}(r)}} |dT_{ij}| \, dt \right) dr \\ &= \pi \int_0^\infty \rho_n(r)\psi(r)\, h^{-1'}(r) dr \\ &= \pi \left\langle \rho_n, \psi \cdot h^{-1'} \right\rangle_{\Omega_1}. \end{aligned} \tag{2.15}$$

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

Corollary 2.0.5. *The map $T^\sharp : \mathscr{D}^*(\Omega_1) \to \mathscr{D}^*(\Omega_2)$ is injective.*

Proof. Suppose $\rho \in \mathscr{D}^*(\Omega_1)$ is such that $T^\sharp \rho = 0$. By the inverse function theorem, $h^{-1'}(r) =$

$\frac{1}{h'(r)} > 0$ since h is increasing. So, $h' \cdot h^{-1'}(r) = 1$. By Lemma 2.0.4, for an aribitrary $\psi \in \mathscr{D}(\Omega_1)$,

$$0 = \left\langle T^\sharp \rho, (h' \cdot \psi) \circ T) \right\rangle_{\Omega_2} = \left\langle \rho, \psi \right\rangle_{\Omega_1}, \tag{2.16}$$

thus, $T^\sharp$ has trivial kernel and as a linear map is injective.

□

2.0.3 The Space of Radially Symmetric Functions

What follows in this section are my thoughts as formally as I can state them for defining $\mathscr{K} \subset L^2(\Omega_2)$, the space of radially symmetric functions, and $\mathscr{P}$ the space of their radial representations. I have run into trouble showing the $\mathscr{K}$ is closed in $L^2(\Omega_2)$. I don't intend for this to be in the paper in this form, but I am trying to present it in the form that seems most reasonable for solving the problem. The gist is that I now believe that an extra condition is needed on h so that $\mathscr{K}$ is closed, but I am not sure what it should be. Alternatively, maybe I should change $\mathscr{K}$ or $\mathscr{P}$.

We will want to only consider radial representations, $\mathscr{P}$, that correspond to square integrable radially symmetric functions, $\mathscr{K}$. Therefore, a reasonable space for radial representations is $\mathscr{P} = T^{\sharp -1}(L^2(\Omega_2))$. If we restrict $T^\sharp$ to $\mathscr{P}$, then the image of this map is $T^\sharp(T^{\sharp -1}(L^2(\Omega_2)))$, and it is a straight-forward set theory argument to show that this is precisely $\mathscr{K} = T^\sharp(\mathscr{D}^*(\Omega_1) \bigcap L^2(\Omega_2)$. Here is the argument:

Proposition 2.0.6. *Let $f : X \to Y$ with $B \subseteq Y$, then $f(f^{-1}(B)) = f(X) \bigcap B$.*

Proof. Let $y \in f(f^{-1}(B))$, then there exists a $x \in f^{-1}(B)$ such that $f(x) = y$. Since $x \in f^{-1}(B)$, this implies that $f(x) = y \in B$. We have shown $y \in f(X)$ and $y \in B$, hence $f(f^{-1}(B)) \subseteq F(X) \bigcap B$.

On the other hand, suppose $y \in f(X)$ and $y \in B$. Then there exists $x \in X$ such that $y = f(x) \in B$. Hence $x \in f^{-1}(B)$. This implies $y = f(x) \in f(f^{-1}(B)$. □

Both $\mathscr{P}$ and $\mathscr{K}$ seem to be the natural choices for these spaces. Moreover, the map

$$T^\sharp|_{\mathscr{P}} : \mathscr{P} \to \mathscr{K} \tag{2.17}$$

is a surjection. Since $\mathscr{K} \subset L^2(\Omega)$, we can endow it with the subspace topology, and endow $\mathscr{P}$ with the topology whose open sets are of the form $T^{\sharp -1}(U)$. This makes $T^\sharp$ automatically continuous, and the correspondence of the open sets is given by Lemma 2.0.4

I have been struggling to prove that $\mathscr{K}$ is closed in $L^2(\Omega_2)$, and I believe that it might not be true in general for all increasing diffeomorphisms $h : \Omega_1 \to \Omega_1$.

One reason I believe h is not restrictive enough is the following sketch of a construction: Consider the sequence $\{\rho_n\} \subset \mathscr{D}(\Omega_1)$ given by the convolution

$$\rho_n(r) = r^{-1/3}\chi_{r\geq 1}(r) * \delta_n(r) \tag{2.18}$$

where $\delta_n \in \mathscr{D}(\Omega_1)$ is an approximate identity. This sequence converges to a distribution $\rho \in L^2(\Omega_1)$. Yet, if $h(t) = t^2$, then $h^{-1\prime}(r) = \frac{1}{2}r^{-1/2}$, then invoking Lemma 2.0.4,

$$\int_{\Omega_2} T^\sharp\rho_n \cdot \overline{T^\sharp\rho_n}\, dxdy = \Big\langle T^\sharp\rho_n, \overline{\rho_n} \circ T \Big\rangle_{\Omega_2} \tag{2.19}$$

$$= \pi\Big\langle \rho_n, h^{-1\prime} \cdot \overline{\rho_n} \Big\rangle_{\Omega_1} \tag{2.20}$$

$$= \pi \int_{\Omega_1} \rho_n(r)^2 h^{-1\prime}(r) dr \tag{2.21}$$

which defines a sequence diverging like $\int_1^\infty r^{-7/6}dr$. Hence $T^\sharp$ in this case is not continuous on $L^2(\Omega_1)$. Observe, however, that the choice of $h(t)$ has the wrong concavity to make points more dense near the origin. I think I can reconcile this with h of the form $h(t) = t^a$ where

$0 < a < 1/2$, or maybe even $0 < a < 1$, but is considering only such functions too restrictive? Although this doesn't say anything about whether or not $\mathscr{K}$ is closed in $L^2(\Omega_2)$, it precludes the following sketch of an argument that I initially wanted to use:

Ideally, I would like to have h so that I can prove that $\mathbb{R} \subseteq L^2(\Omega_1)$ and $T|_{L^2(\Omega_1)}$ continuously maps into $L^2(\Omega_2)$ (with respect to L^2 in both domain and range). Then, $\mathbb{R} = T^{\sharp -1}(L^2(\Omega_2))$ is closed in $L^2(\Omega_1)$. Moreover, by the open mapping theorem [Rudin pg. 48], $T^\sharp|_{\mathbb{R}}$ is an open map ($\mathbb{R}$ is a Hilbert space, thus an F-space). Since $T^\sharp|_{\mathbb{R}}$ is injective, it is also a closed map (Use $T^\sharp(U^c) = T^\sharp(U)^c$). Thus $\mathscr{K} = T^\sharp(\mathbb{R})$ is closed in $L^2(\Omega_2)$. It would also be nice to show that the topology induced by $T^\sharp|_{\mathbb{R}}$ is coarser than that of $L^2(\Omega_1)$ so that convergence in the first implies convergence in the latter.

I have other ideas about proving this directly by considering limit points of $\mathscr{K}$, but the above approach seems nice. As you suspected, this is much more complicated that I initially anticipated. Any comments, thoughts, suggestions as always are appreciated.

Chapter 3

Reconstruction on the Computer

In the last chapter, we established the theory for defining the inverse problem in an infinite dimensional Hilbert space. The development led to defining the functional $\Phi : \mathscr{H}_1 \times \mathscr{H}_2 \to \mathbb{R}$

$$\Phi(p; b) = \lambda \|\mathcal{G}p - b\|^2_{\mathscr{H}_2} + \delta \langle p, \mathcal{L}p \rangle_{\mathscr{H}_1}, \tag{3.1}$$

where $\mathcal{G}$ is the operator that takes a radial profile p to a line of blurred image of an edge b, and $\mathcal{L}$ is the induced radial Laplacian. We also showed that $\mathcal{L}^n$ was trace class for all n in $\mathscr{H}_1$. Hence, in the infinite dimensional Bayesian perspective, if $p \sim \mu_p$ where μ_p is a Gaussian measure with mean 0 and precision $\delta\mathcal{L}$, then in the presence of independent noise precision λI (with corresponding measure μ^b) the posterior is Gaussian with a Radon-Nykodym derivative with respect to μ_p

$$\frac{\mu^b}{\mu_p}(p) \propto \exp\left(-\Phi(p; b)\right), \tag{3.2}$$

where $\Phi(p; b)$ is unique up to scaling by a factor dependent on b. We called this the infinite dimensional posterior for p given b.

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 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 a Laplacian precision operator and a hierarchical Gamma prior for δ with a given estimate for λ. They analyzed the infinite dimensional Gibbs sampler, and showed that it had deficiencies in sampling δ that exacerbate as the discretization converges. They then introduced two algorithms that alleviate this issue. Although their analysis is not directly applicable to PSF reconstruction since our prior is the *radial* Laplacian, we take cues from their work to design the algorithm for exploring the *discrete* posterior density for PSF reconstruction. Also, by discretizing at this stage, we will be able to develop an algorithm that allows for the noise precision λ to be estimated.

The discrete representations will be point-based on equally spaced grids, and each operator defined in the Chapter 2 will be estimated using either numerical quadrature or finite differencing. Hence, to each operator we will define a matrix that defines their action on the point-wise estimates, and they will be derived in Section 3.1 We then develop the *discrete* probability spaces associated with the matrix-operators defined in Section 3.1, which will serve as our discrete approximation of the infinite dimensional space defined in Chapter 2. In this development, we will add "uninformative" 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. In Section 3.2, we will derive a standard Gibbs' sampling algorithm Geman and Geman [1984] and improve upon it using a technique called *partial collapse*, which can motivated by several recent theoretical and practial analyses [Van Dyk and Park, 2008, Agapiou et al., 2014, Fox and Norton, 2015]. We will also briefly review standard convergence diagnostics for comparing Markov Chain based sampling algorithms, which will show that our adapted algorithm is indeed an enhancement of standard Gibbs' sampling.

3.1 Discrete representation of the inference problem

Recall that the radial Laplacian $\mathcal{L} : \mathscr{H}_1^{\circ} \to \mathscr{H}_1^{\circ}$ acted by

$$[\mathcal{L}p](r) = r^{-1}\frac{d}{dr}\left(r\,\frac{d}{dr}p(r)\right). \tag{3.3}$$

Recall also that the inner product for the regularization is also affected by the change of variables, i.e.

$$\langle p_1, \mathcal{L}^n p_2\rangle_{\mathscr{H}_1^{\circ}} = 2\pi\int_0^{\infty} p_1(r)[\mathcal{L}^n p_2](r)\,r dr. \tag{3.4}$$

The differential operator R is discretized using centered differencing, i.e. for $h = 1/M$ denote $r_{j\pm1/2} = h\cdot j \pm h/2$, $x_j = x(r_j)$ and

$$[\boldsymbol{R}\boldsymbol{x}]_j \overset{\text{def}}{=} \frac{1}{h^2}\Big(r_{j+1/2}(x_{j+1}-x_j) - r_{j-1/2}(x_j - x_{j-1})\Big). \tag{3.5}$$

or as a matrix stencil

$$\frac{1}{h^2}\begin{bmatrix} -(r_{j-3/2}+r_{j-1/2}) & r_{j-1/2} & 0 \\ r_{j-1/2} & -(r_{j-1/2}+r_{j+1/2}) & r_{j+1/2} \\ 0 & r_{j+1/2} & -(r_{j+1/2}+r_{j+3/2}) \end{bmatrix}\begin{bmatrix} x_{j-1} \\ x_j \\ x_{j+1}\end{bmatrix}. \tag{3.6}$$

The discretization of $\boldsymbol{R}$ has a zero right boundary condition since $\lim_{r\to\infty} x(r) = 0$ (we assume the domain for the radial profile is sufficiently large so that r_M is less than machine epsilon) and a reflective left boundary condition because of the assumption of radial symmetry. We then take $\boldsymbol{L} = \boldsymbol{r}^{-1}\odot\boldsymbol{R}^2$, i.e., the coordinate wise multiplication of reciprocals of the grid points r_j composed with $\boldsymbol{R}^2$. Finally, the discretization of the transformed inverse problem is

$$\boldsymbol{b} = \mathbf{A}x + \boldsymbol{\varepsilon}, \tag{3.7}$$

with $\boldsymbol{\varepsilon}\sim\mathcal{N}(\mathbf{0},\lambda^{-1}\boldsymbol{I})$ and $\boldsymbol{x}\sim\mathcal{N}(\mathbf{0},\delta^{-1}(\boldsymbol{L}^T\boldsymbol{L})^{-1})$.

3.1.1 The discrete posterior distribution

3.2 Markov Chain Monte Carlo estimation

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 can be thought of as the analogous central limit theorem for Markov chains. These notions will be briefly overviewed in the next section, and complete treatments can be found in [Robert and Casella, 2013, Liu, 2008]. In the context of PSF reconstruction, samples will be taken from the posterior $p(\boldsymbol{x}, \delta, \lambda|\boldsymbol{b})$. 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 dependence between components of the joint density. Our use of partial collapse will be motivated by the marginal algorithm in [Agapiou et al., 2014], a similar infinite dimensional sampler.

In what follows, we will develop theory for a general p component u Note that the model PSF estimation has $p = 3$ components, and the general development is not necessary for this problem. We've developed this process in the general setting, with potential modifications to the hierarchical model in mind. In [Howard et al., 2015], they observed potential sensitivity to the uninformative hyper-prior parameters $\alpha_\delta, \beta_\delta, \alpha_\lambda$ and β_λ in a similar hierarchical Bayesian estimation problem. A possible extension would be to impose a conjugate prior distribution on these parameters, in which case $p = 7$.

3.2.1 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, $\mathcal{F}$ a sigma-algebra of events from Ω and $\mathbb{P}$ a measure on $\mathcal{F}$ into $[0, 1]$. We will be concerned with sampling a p component random variable $\boldsymbol{X} = (X_1 \dots X_p) : \Omega \to \mathbb{R}^N$ so that the measure (known as its *law*) induced by X on $\mathbb{R}^N$ by taking pre-images of Borel sets is absolutely continuous with respect to Lebesgue measure. Hence, each law corresponding to X_i has a density (its Radon-Nykodym derivative with respect to Lebesgue measure) $\pi_{\boldsymbol{X}}(\boldsymbol{x}) = \pi_{\boldsymbol{X}}(x_1 \dots x_p)$, where X_i take on values in $\mathbb{R}^{k_i}$ such that $\sum k_i = N$. 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 $\pi(\boldsymbol{x}) \stackrel{\text{def}}{=} \pi_{\boldsymbol{X}}(\boldsymbol{x})$. Denote the vector with the ith component removed $\boldsymbol{x}_{\widehat{i}} \stackrel{\text{def}}{=} (x_1 \dots x_{i-1}, x_{i+1} \dots x_p)$, then the *marginal distribution* is

$$\pi(\boldsymbol{x}_{\widehat{i}}) \stackrel{\text{def}}{=} \int_{x_i} \pi(x_1 \dots x_p) dx_i, \tag{3.8}$$

and *conditional distribution* is

$$\pi(\boldsymbol{x}_{\widehat{i}} | x_i) \stackrel{\text{def}}{=} \pi(\boldsymbol{x}) / \pi(\boldsymbol{x}_{\widehat{i}}). \tag{3.9}$$

A Markov chain is a stochastic process $\{\boldsymbol{X}^0, \boldsymbol{X}^1, \boldsymbol{X}^2, \dots\}$ with X_i defined on a common probability space such that

$$\mathbb{P}\left(\boldsymbol{X}^{k+1} \in A | \boldsymbol{X}^k = \boldsymbol{x}^k, \dots \boldsymbol{X}^0 = \boldsymbol{x}^0\right) = \mathbb{P}\left(\boldsymbol{X}^{k+1} \in A | \boldsymbol{X}^k = \boldsymbol{x}^k\right) \tag{3.10}$$

for all events A.

A family of probability densities $K(\boldsymbol{x}, \cdot)$ is a *transition kernel* for the Markov chain if

$$\int_A K(\boldsymbol{x}^k, \boldsymbol{x}')d\boldsymbol{x}' = \mathbb{P}\left(\boldsymbol{X}^{k+1} \in A | \boldsymbol{X}^k = \boldsymbol{x}^k\right), \tag{3.11}$$

and $K(\cdot, \boldsymbol{x}')$ is measurable (that is, an un-normalized probability density). For a transition kernel, the corresponding *transition operator* is $\mathcal{K} : L^1 \to L^1$ by

$$\mathcal{K}[\pi](\boldsymbol{x}') = \int K(\boldsymbol{x}, \boldsymbol{x}')\pi(\boldsymbol{x})d\boldsymbol{x}. \tag{3.12}$$

Now, consider the joint density $\pi(\boldsymbol{x}^N \dots \boldsymbol{x}^0)$ for the truncated chain $\{\boldsymbol{X}^1 \dots \boldsymbol{X}^n\}$ with $\pi_0(\boldsymbol{x})$

the density for $\boldsymbol{X}_0$ and observe

$$\begin{aligned}
\pi(\boldsymbol{x}^1) &= \int_{\boldsymbol{x}^0} \pi(\boldsymbol{x}^1, \boldsymbol{x}^0) d\boldsymbol{x}^0 \\
&= \int_{\boldsymbol{x}^0} \pi(\boldsymbol{x}^1 | \boldsymbol{x}^0) \pi(\boldsymbol{x}^0) d\boldsymbol{x}^0 \\
&= \int_{\boldsymbol{x}^0} K(\boldsymbol{x}^0, \boldsymbol{x}^1) \pi(\boldsymbol{x}^0) d\boldsymbol{x}^0 \\
&= \mathcal{K}[\pi_0](\boldsymbol{x}^1) \qquad (3.13) \\
\pi(\boldsymbol{x}^2) &= \int_{\boldsymbol{x}^1} \int_{\boldsymbol{x}^0} \pi(\boldsymbol{x}^2, \boldsymbol{x}^1, \boldsymbol{x}^0) d\boldsymbol{x}^0 d\boldsymbol{x}^1 \\
&= \int_{\boldsymbol{x}^1} \pi(\boldsymbol{x}^2 | \boldsymbol{x}^1) \int_{\boldsymbol{x}^0} \pi(\boldsymbol{x}^1, \boldsymbol{x}^0) d\boldsymbol{x}^0 d\boldsymbol{x}^1 \\
&= \int_{\boldsymbol{x}^1} K(\boldsymbol{x}^1, \boldsymbol{x}^2) \int_{\boldsymbol{x}^0} \pi(\boldsymbol{x}^1, \boldsymbol{x}^0) d\boldsymbol{x}^0 d\boldsymbol{x}^1 \\
&= \mathcal{K} \circ \mathcal{K}[\pi_0](\boldsymbol{x}^2) \qquad (3.14) \\
&\vdots \\
\pi(\boldsymbol{x}^N) &= \int_{\boldsymbol{x}^{N-1}} \dots \int_{\boldsymbol{x}^0} \pi(\boldsymbol{x}^N, \boldsymbol{x}^{N-1} \dots \boldsymbol{x}^0) d\boldsymbol{x}_0 \dots, d\boldsymbol{x}_{N-1} \\
&= \int_{\boldsymbol{x}^{N-1}} \dots \int_{\boldsymbol{x}^0} \pi(\boldsymbol{x}^N | \boldsymbol{x}^{N-1}) \pi(\boldsymbol{x}^{N-1} \dots, \boldsymbol{x}^0) d\boldsymbol{x}_0 \dots, d\boldsymbol{x}_{N-1} \\
&= \int_{\boldsymbol{x}^{N-1}} K(\boldsymbol{x}^{N-1}, \boldsymbol{x}^N) \dots \int_{\boldsymbol{x}^0} K(\boldsymbol{x}^0, \boldsymbol{x}^1) \pi(\boldsymbol{x}^0) d\boldsymbol{x}_0 \dots, d\boldsymbol{x}_{N-1} \\
&= \mathcal{K}^N[\pi_0](\boldsymbol{x}^N). \qquad (3.15)
\end{aligned}$$

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 $K(\boldsymbol{x}, \boldsymbol{x}')$. Moreover, 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 *irreducibly* 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 $\int_A \mathcal{K}^N(\boldsymbol{x}, \boldsymbol{x}')dx'$ [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.

For us, this condition is easily met because we will consider transition kernels that are given by probability density functions with full support over the range of the random variables of interest. Hence, the Markov chains relevant to our algorithms are irreducible with respect to Lebesgue measure, and thus irreducible with respect to all measures absolutely continuous with respect to Lebesgue measure, and in particular, those given by probability densities.

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

$$\mathcal{K}[\pi(\boldsymbol{x})](\boldsymbol{x}') = \pi(\boldsymbol{x}'). \tag{3.16}$$

Note that an invariant distribution π is a eigenvector for the transition operator $\mathcal{K}$ corresponding to the eigenvalue 1. Moreover, since transition operators consist of probability densities, $\|K\pi_0\|_{L^1} \leq \|\pi_0\|_{L^1}$, thus π is the eigenvector corresponding to the leading order eigenvalue. With this viewpoint, there is an interesting connection the power-iteration method for finding leading order eigenvalues and corresponding eigenvectors. Specifically, when the space of possible outcomes is finite, say $\{\omega_1 \dots \omega_n\}$, then all probability densities π_0 correspond to real $\{\alpha_1 \dots \alpha_n\}$ such that $\sum \alpha_i = 1$. Taking the states $\{\omega_1 \dots \omega_n\}$ as basis vectors, the probability densities form a finite dimensional vector space and transition operators correspond to multiplication by a *transition matrix*. The power-iteration method states that the sequence give by the recursive relation $\pi_k = K\pi_{k-1}/\|K\pi_{k-1}\|$ converges to the leading order eigenvector. Hence, the invariant density $\pi = \lim_{k\to\infty} K^n\pi$.

There are two last technical conditions, known as Harris recurrence and aperiodicity, that must be defined in order establish the hypotheses of the ergodic theorem for Markov chains.

When the Markov chain is discrete, the *period* of a state ω is the greatest common denominator of the set $\{m \geq 1; K^m(\omega, \omega > 0\}$; that is, if ω is d-periodic, then returns to state ω occur multiples of d. For example, the deterministic two state Markov chain associated with the transition matrix

$$K = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \tag{3.17}$$

has period 2. A chain is *aperiodic* if each state has period 1. Since the period is constant on all states that communicate with ω, an irreducible chain has a unique period for all states. Verifying rigorously the requirement of aperiodicity for continuous Markov chains is somewhat more technical, and we cite [Liu, 2008] who states that transition kernels associated with Gibbs sampling and Metropolis-Hastings are aperiodic, and the algorithms presented in this work are compositions of such transitions. See [Robert and Casella, 2013] for the technical definition and details.

Harris recurrence ensures that a Markov chain re-enters events often enough to "fill-out" π. Formally, for a Borel set A, the *average number of passages* of $(\boldsymbol{X}^n)$ in A is $\eta_A \overset{\text{def}}{=} \sum_n I_A(\boldsymbol{X}^n)$ 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 [Robert and Casella, 2013] who ensures that transitions associated with Gibbs sampling and Metropolis-Hastings 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 3.2.1. *[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 and for any initial density π_0 and all $\boldsymbol{x}$ but a subset whose measure under π is zero,*

(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

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

Equation (3.18) of the ergodic theorem 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 are to simulate a Markov chain *designed* so that it has a desired density π. In the context of our Bayesian hierarchical model, this will be the posterior density $\pi(\boldsymbol{x}, \lambda, \delta|\boldsymbol{b})$. A widely used method that can be used when sampling from full conditional distributions is available, known as *Gibbs sampling*, is presented in the following section.

3.2.2 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 modelling particles in a lattice system. They developed the following 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 for analyzing the posterior density $\pi(\boldsymbol{p}, \delta, \lambda|\boldsymbol{b})$.

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

Algorithm 1 Gibbs sampler

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

1. $X_1^k \sim \pi(x_1|x_2^{k-1}, x_3^{k-1} \dots x_p^{k-1})$
2. $X_2^k \sim \pi(x_2|x_1^k, x_3^{k-1} \dots x_p^{k-1})$
...
p. $X_p^k \sim \pi(x_p|x_1^k, x_2^k \dots x_{p-1}^k)$

Algorithm 1 simulates the outcomes from the transition kernel

$$K(\boldsymbol{x}, \boldsymbol{x}') = \pi(x_p'|\boldsymbol{x}'_{\widehat{p}}) \dots \pi(x_2'|x_1', \boldsymbol{x}_{\widehat{12}})\pi(x_1'|\boldsymbol{x}_{\widehat{1}}). \tag{3.20}$$

Note that we can view the action of the transition in steps since it is factors, i.e.

$$\begin{aligned}
\mathcal{K}[\pi_0](\boldsymbol{x}') &= \int K(\boldsymbol{x}, \boldsymbol{x}')\pi_0(\boldsymbol{x})d\boldsymbol{x} \\
&= \int_{x_p} \dots \int_{x_1} \pi(x_p'|\boldsymbol{x}'_{\widehat{p}}) \dots \pi(x_2'|x_1', \boldsymbol{x}_{\widehat{12}})\pi(x_1'|\boldsymbol{x}_{\widehat{1}})dx_1 \dots dx_p \\
&= \int_{x_p} \pi(x_p'|\boldsymbol{x}'_{\widehat{p}}) \int_{x_{p-1}} \pi(x_{p-1}'|\boldsymbol{x}'_{\widehat{p,p-1}}x_p) \dots \int_{x_1} \pi(x_1'|\boldsymbol{x}_{\widehat{1}})\pi_0(x_1 \dots x_p)dx_1 \dots dx_p.
\end{aligned} \tag{3.21}$$

Each integration in (3.21) can be thought of as a sub-transition on components of $\pi_0(x_1 \dots x_p)$; that is, let

$$\mathcal{K}_i[\pi_0(\boldsymbol{x})](\boldsymbol{x}') \stackrel{\text{def}}{=} \int_{x_i} \pi(x_i'|x_1', \dots x_{i-1}', x_{i+1} \dots x_p)\pi_0(\boldsymbol{x})dx_i, \tag{3.22}$$

then we can express $\mathcal{K} = \mathcal{K}_p \circ \mathcal{K}_{p-1} \circ \dots \circ \mathcal{K}_1$. We note that functionally each operator $\mathcal{K}_i$ depends on $(x_1 \dots x_{i-1}, x_{i+1} \dots x_p)$ being given, and that successively integrating removes dependence. For example $\mathcal{K}_1$ depends on $(x_2 \dots x_p)$, $\mathcal{K}_2 \circ K_1$ depends on $(x_3 \dots x_p)$, etc. until the full composition $\mathcal{K}$ 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 (analyzing the transition kernel as a composition of conditional operators) will be useful for designing and verifying the stationarity of the proceeding algorithms.

Proposition 3.2.2. *The transition kernel associated to Algorithm 1 produces a Markov chain that is invariant to the density π.*

Proof. Observe that

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

Moreover, given $\mathcal{K}_{i-1}\circ\cdots\circ\mathcal{K}_1 = \pi(x_1'\dots x_{i-1}',x_i\dots x_p)$, then

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

By induction, $\mathcal{K}[\pi(\boldsymbol{x})](\boldsymbol{x}') = \mathcal{K}_p\circ\cdots\circ\mathcal{K}_1[\pi(\boldsymbol{x})](\boldsymbol{x}') = \pi(x_1',x_2'\dots x_p')$. Hence π is invariant. □

Note that the partial composition $\mathcal{K}_i\circ\cdots\circ\mathcal{K}_1$ is invariant with respect to $\pi(x_1\dots x_i|x_{i+1},\dots,x_p)$ when $(x_{i+1}\dots x_p)$ are given, then $\pi(\boldsymbol{x})/\pi(x_{i+1}\dots x_p) = \pi(x_1\dots x_i|x_{i+1}\dots x_p)$ and since each integration does not depend on $(x_{i+1}\dots x_p)$ by (3.24)

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

Viewing Gibbs sampling as composed conditional sub-transitions allows for the flexibility to design and analyze algorithms that modify each sub-transition step. Proving that the modifications retain the invariant distribution of the chain can focus on the sub-transition operator associated with the modified step and fit generally into the simple argument presented above.

As will be seen in following sections, the Gibbs sampler will produce good Monte Carlo estimates for the discretized PSF $\boldsymbol{p}$ and the noise level given by λ. However, the δ component of the chain exhibits poor convergence, hence, the asymptotic application of the ergodic theorem for Markov chains for the joint variable $(\boldsymbol{p}, \lambda, \delta)$ is not available. In fact, [Agapiou et al., 2014] gave theory that showed that the infinite dimensional hierarchical Gibbs sampler for linear inverse problems with λ known and Laplacian regularization 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 "marginalization" process can be carried out in general and is known as partial collapse, and is presented in [Van Dyk and Park, 2008]. In their paper, they showed various examples of partially collapsing the Gibbs sampler and how it can lead to Markov chains that no longer have π as an invariant density. They also presented theory that this process in general improves chain 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.

3.2.3 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. However, in both [Van Dyk and Park, 2008, Van Dyk and Jiao, 2015], they highlight that application of partial collapse must be done with care, 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 of interest. They carefully outlined methods for ensuring that the pitfalls of improper sampling are avoided, although did not formally prove the invariance of the resulting Markov chains. In this section, we give rigorous arguments that show the Markov chains associated with proper partial collapse are indeed invariant.

Consider the following modification of step p-1. in Algorithm 1:

Algorithm 2 p-Conditioned Gibbs sampler

Given $\widetilde{\boldsymbol{x}}^{k-1} = (x_1^{k-1} \dots \widetilde{x}_p^{k-1}, x_p^{k-1})$, simulate

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

$\vdots$

p-1. $(X_{p-1}^k, \widetilde{X}_p^k) \sim \pi(x_{p-1}, x_p | x_1^k, x_2^k \dots x_{p-2}^k)$

p. $X_p^k \sim \pi(x_p | x_1^k, x_2^k \dots x_{p-1}^k)$

If $\widetilde{\boldsymbol{X}}^k = (X_1^k \dots X_{p-1}^k, \widetilde{X}_p^k, X_p^k)$, then the corresponding transition operator is

$$\mathcal{K}\pi_0 = \mathcal{K}_p \circ \widetilde{\mathcal{K}}_{p-1} \dots \mathcal{K}_2 \circ \mathcal{K}_1 \pi_0 \tag{3.26}$$

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

$$\widetilde{K}_{p-1}(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{x}}') \stackrel{\text{def}}{=} \widetilde{K}_{p-1}(\widetilde{\boldsymbol{x}}') = \pi(x'_{p-1}, \widetilde{x}'_p | x'_1, x'_2 \dots x'_{p-2}) \tag{3.27}$$

Algorithm 2 produces a Markov chain with $p+1$ components by drawing $(X_{p-1}^k, \widetilde{X}_p^k)$ jointly at step p-1, and note that the transition to the next state does not depend on previous values of

$\widetilde{X}_p$. 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 3.2.3. *The Markov chain associated with the transition kernel corresponding to Algorithm 2 is invariant with respect to* $\pi(\boldsymbol{x})\pi(\widetilde{x_p}|\boldsymbol{x})$.

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

$$\begin{aligned}
&\widetilde{\mathcal{K}}\Big[\pi(\boldsymbol{x})\pi(\widetilde{x}_p|\boldsymbol{x})\Big](\widetilde{\boldsymbol{x}}') \\
&\quad= \mathcal{K}_p\widetilde{\mathcal{K}}_{p-1}\mathcal{K}_{p-2}\circ\cdots\circ\mathcal{K}_1\big[\pi(\boldsymbol{x})\pi(\widetilde{x}_p|\boldsymbol{x})\big](\widetilde{\boldsymbol{x}}') \\
&\quad= \int_{x_p}\pi(x_p'|\boldsymbol{x}'_{\widehat{p}})\iint_{\widetilde{x}_p,x_{p-1}}\widetilde{K}_{p-1}(\widetilde{\boldsymbol{x}}')\int_{x_{p-2}\ldots x_1}\cdots\int\ldots\pi(\boldsymbol{x})\pi(\widetilde{x}_p|\boldsymbol{x})dx_1\ldots d\widetilde{x}_p dx_p \\
&\quad= \int_{x_p}\pi(x_p'|\boldsymbol{x}'_{\widehat{p}})\int_{x_{p-1}}\widetilde{K}_{p-1}(\widetilde{\boldsymbol{x}}')\int_{x_{p-2}}\ldots\int_{x_1}\pi(x_1'|\boldsymbol{x}_{\widehat{1}})\pi(\boldsymbol{x})\left(\int_{\widetilde{x}_p}\pi(\widetilde{x}_p|\boldsymbol{x})\widetilde{x}_p\right)d\boldsymbol{x} \qquad (3.28)
\end{aligned}$$

where we used Fubini's theorem to integrate first in $\widetilde{x}_p$ for which each kernel K_i does not depend, and $\int\pi(\widetilde{x}_p|\boldsymbol{x})d\widetilde{x}_p = 1$. Now, each of the inner $p-2$ integrations express the action of the first $p-2$ steps of the standard Gibbs sampler, so continuing from (3.28),

$$\begin{aligned}
\widetilde{\mathcal{K}}\Big[\pi(\boldsymbol{x})\pi(\widetilde{x}_p|\boldsymbol{x})\Big](\widetilde{\boldsymbol{x}}') &= \int_{x_p}\pi(x_p'|\boldsymbol{x}'_{\widehat{p}})\int_{x_{p-1}}\widetilde{K}_{p-1}(\widetilde{\boldsymbol{x}}')\cdot\mathcal{K}_{p-2}\circ\cdots\circ\mathcal{K}_1[\pi(\boldsymbol{x})](\boldsymbol{x}') \\
&= \int_{x_p}\pi(x_p'|\boldsymbol{x}'_{\widehat{p}})\int_{x_{p-1}}\pi(x'_{p-1},\widetilde{x}'_p|x'_1,x'_2\ldots x'_{p-2})\cdot\pi(x'_1\ldots x'_{p-2},x_{p-1},x_p) \\
&= \pi(x_p'|\boldsymbol{x}'_{\widehat{p}})\pi(x'_{p-1},\widetilde{x}'_p|x'_1,x'_2\ldots x'_{p-2})\cdot\pi(x'_1\ldots x'_{p-2}) \\
&= \frac{\pi(\boldsymbol{x}')\pi(x'_1,x'_2\ldots x'_{p-1},\widetilde{x}'_p)}{\pi(\boldsymbol{x}'_{\widehat{p}})} \\
&= \pi(\boldsymbol{x}')\pi(\widetilde{x}'_p|\boldsymbol{x}'_{\widehat{p}}). \qquad (3.29)
\end{aligned}$$

□

Note that it is essential that each sub-kernel K_i does not depend on $\widetilde{x}_p$, else the initial integration in $\widetilde{x}_p$ would involve products of kernels depending on $\widetilde{x}_p$ with $\pi(\widetilde{x}_p|\boldsymbol{x}_{\widehat{p}})$. Also, the placement

In some sense, this algorithm is artificial, as we do not need to sample the auxiliary variable $\widetilde{X}_p$. Moreover, if we integrate the invariance condition

$$\int_{\widetilde{x}'_p} \widetilde{\mathcal{K}}[\pi(\boldsymbol{x})\pi(\widetilde{x}_p|\boldsymbol{x}_{\widehat{p}})](\widetilde{\boldsymbol{x}})d\widetilde{x}'_p = \pi(\boldsymbol{x}') \int_{\widetilde{x}'_p} \pi(x'_p|\boldsymbol{x}'_{\widehat{p}})dx'_p = \pi(\boldsymbol{x}'), \tag{3.30}$$

This results in the same transition kernel as the p-Conditioned sampler except for at step $p-1$

$$\overline{K}_{p-1}(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{x}}') \stackrel{\text{def}}{=} \int_{x'_p} \pi(x'_{p-1}, \widetilde{x}'_p|x'_1, x'_2 \dots x'_{p-2}) = \pi(x'_{p-1}|x'_1, x'_2 \dots x'_{p-2}). \tag{3.31}$$

By (3.30), the corresponding Markov Chain is invariant with respect to π. The following algorithm simulates this chain:

Algorithm 3 p-Partially Collapsed Gibbs sampler

Given $\widetilde{\boldsymbol{x}}^{k-1} = (x_1^{k-1} \dots \widetilde{x}_p^{k-1}, x_p^{k-1})$, simulate

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

$\vdots$

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

p. $X_p^k \sim \pi(x_p|x_1^k, x_2^k \dots x_{p-1}^k)$

There is one last modification to the transition kernel that will be required. In many cases, as will be the case of PSF reconstruction, a simulation from $\pi(x_{p-1}|x_1 \dots x_{p-2})$ 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 review Metropolis-Hastings, and show that directly substituting a Metropolis-Hastings transition into the p-partially collapsed Gibbs sampler remains stationary.

3.2.4 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 random walk Metropolis-Hastings and how it can be incorporated into the partially collapsed Gibbs sampler. Again, see one of the books [Robert and Casella, 2013, Liu, 2008] and references there for a complete treatment.

Consider the following algorithm for simulating a transition

3.2.5 Partially collapsed Gibbs sampling for PSF estimation

The effect of this process is that we have removed conditioning of $X_p^{k-1} = x_p^{k-1}$ from the sample X_{p-1}^k. Note that steps of the algorithm can be cyclicly permuted with the appropriate re-labeling with respect to k without chaning the full transition kernel (each sub-transition *will* change). The merely changes the initial distribution of the chain.

We can generalize this process by removing the conditioning on either X_{p-1} or X_p on X_{p-2}. Without loss of generality, X_{p-2} can be choosen from $X_1 \dots X_{p-2}$ by relabeling. By noting that K_p and K_{p-1} do not depend on $\boldsymbol{x}$, a similar re-ordering of integration with the auxilary variable as in (3.28) will be valid. Hence, X_p can be partially collapsed out of any number of

proceeding variables.

In practice, one starts with the standard Gibbs sampler, and observes convergence of each component. If a component exhibits poor convergence (see Section 3.3), see if any conditioned variables can be partially collapsed. If it is possible to sample the density with one of the conditoined variable collapsed out, re-order the sampler so that the collapsed variable is last and each of the poorly converging variable directly proceeds it. The theory presented in [Van Dyk and Park, 2008] gaurantees that the convergence of $(\boldsymbol{X}^k)$ will be improved. If some components still exhibit poor convergence, continue be removing the conditioning of one of the previous $p-1$ variables. See [Van Dyk and Park, 2008] for examples and a further discussion of the general process of partially collapsing variables.

3.3 Evaluating Convergence

This section is to introduce all of the chain statistics we will use to see how our algorithm is better than Gibbs. We also compare substeps in the Metropolis step.

3.3.1 Estimating Burn-in

Re-read PC Gibbs on John's section and basically copy his ideas. Talk about how to interpret the ergodic theorem.

3.3.2 Integrated Autocorrelation

Talk about IAC and how one can then develop effective sample size. Pretty straight forward, again re-reading PC Gibbs.

3.4 Sampling the PSF posterior

3.4.1 Gibbs sampling the PSF posterior

3.4.2 Partially collapsed Gibbs sampling for PSF reconstruction

3.4.3 Blocking the sampler and a connection to Marginal then conditional sampling

Chapter 4

Results

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.

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.

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

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.

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

David H Griffel. *Applied functional analysis*. Courier Corporation, 2002.

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.

Lars Hörmander. *The analysis of linear partial differential operators I*. Springer-Verlag, 1983.

Marylesa Howard, Margaret Hock, and Aaron Luttman. Wishart sampling and sensitivity to hyper priors? *???*, 2015.

V. Hutson and J. S. Pym. *Applications of Functional Analysis and Operator Theory*, volume 146 of *Mathematics in Science and Engineering*. Academic Press, 1980.

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.

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.

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

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

Walter Rudin. *Functional analysis. International series in pure and applied mathematics*. McGraw-Hill, Inc., New York, 1991.

MR Showalter, MK Gordon, and D Olson. Vg1/vg2 saturn iss processed images v1.0. NASA Planetary Data System, 2006.

Robert S Strichartz. *The way of analysis*. Jones & Bartlett Learning, 2000.

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.

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.