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

Fix all the emphasis crap. Comma after first word stuff. Subject verb agreement.

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. 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 density.

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 is 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 charge coupled devices (CCD) calibrated to count photons at a specified spectral band.

establish a dichotomy instead of asking a question Having established the form of image data, i.e. *what* the data is, *how* does one analyze it? 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 large volume of data, their spatial nature make measurements at each pixel highly dependent on measurements 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 needs to change to emphasis GMRF. Sentences need to read as easy possible. 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. and blah blah 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 say what b is in this sentence

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

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 for purpose of modeling we assume each function is continuous

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

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

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

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. 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 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]. The additional assumption of *radial symmetry* on k allows 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. Be explicit about what we are solving for. Why has has it changed from f

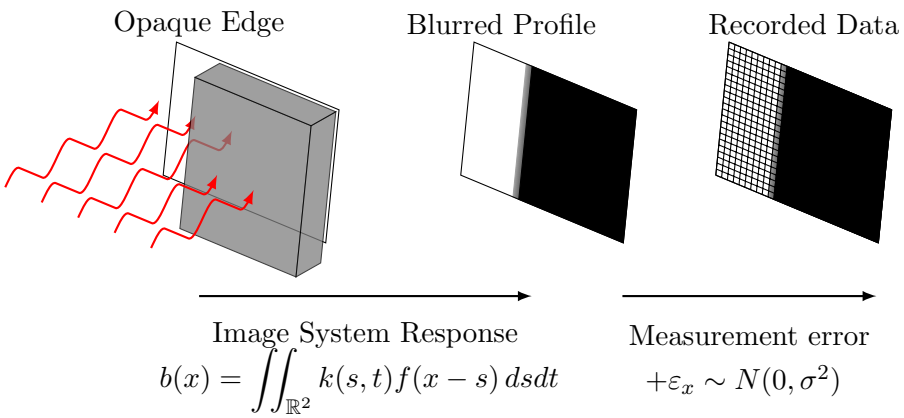

Figure 1.1: A schematic of the measurement model for an X-ray image of an edge. An opaque block aligned with the imaging plane blocks light on the half plane to produce a blurred edge.

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.

1.3 The Abel transform and a deterministic solution

Get rid of future tenses, like we will show to we show Analytic isn't the right word. 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,

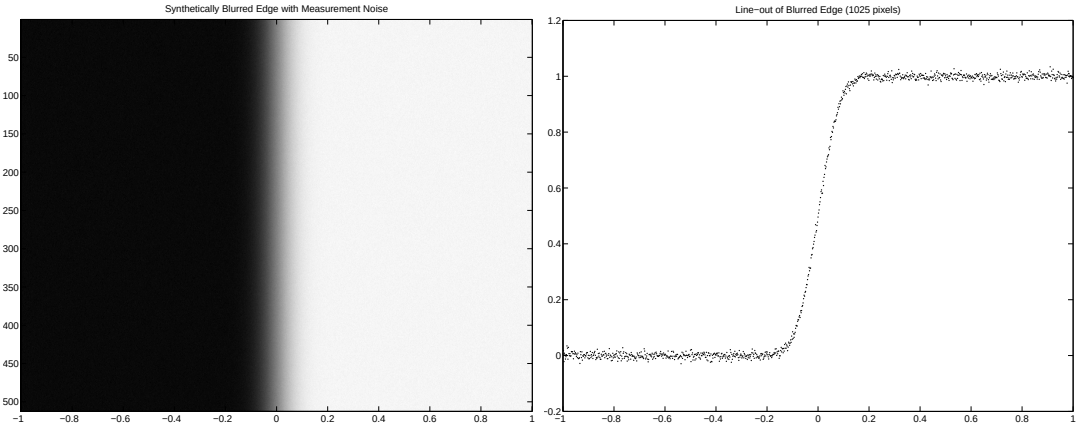

Figure 1.2: A synthetically blurred edge with simulated measurement error and a line-out (horizontal cross-section) from the data.

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 tomo-

graphic 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 $b(0)$. 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 $x > 0$

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

Since $\widetilde{b}(x)$ 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

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

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

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

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.

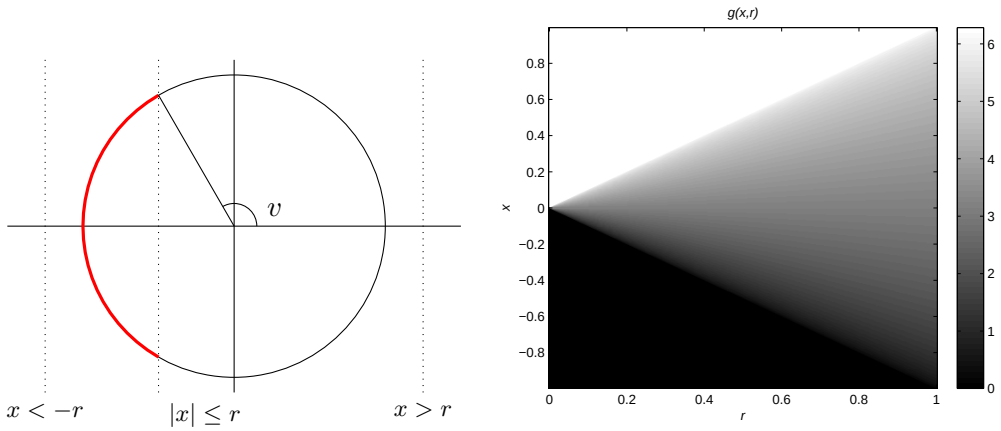

Figure 1.3: The PSF forward integral operator kernel $g(x,r)$ represented as the arc measure of v in $(-\pi,\pi)$ where $x \geq r\cos v$.

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

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

where

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

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.,

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

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 $\Phi : \mathscr{H}_1 \times \mathscr{H}_2 \to \mathbb{R}$

$$\Phi(p; b, \lambda, \delta) = \frac{1}{2}\left(\lambda \|\mathcal{G}p - b\|^2_{\mathscr{H}_2} + \delta \langle p, \mathcal{L}^n p \rangle_{\mathscr{H}_1}\right), \tag{3.1}$$

where $\mathcal{G}$ 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 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

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

where $\Phi(p; b, \lambda, \delta)$ 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 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 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 negative 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. 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 begin this chapter with a brief overview of the general theory for the Markov Chain Monte Carlo alorithms that will be used. In Section 3.1, we will derive the standard Gibbs' sampling algorithm first outlined by [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]. This section will give a full development from first principles, and prove the assertions of invariance stated in [Van Dyk and Park, 2008].

Section 3.2 returns to the PSF estimation problem, in which we discus how to transition from the continuum to numerical estimation on a computer. We then derive all of the necessary probability densities for carrying out the algorithms in Section 3.1. The discrete representations will be point-based on equally spaced grids, and each operator defined in 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 3.2. We then develop the *discrete* probability spaces associated with the matrix-operators defined in Section 3.2, 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. 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 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 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 $\pi(\boldsymbol{x}, \delta, \lambda|\boldsymbol{b})$. 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 density. Note that the

model for PSF estimation has $p = 3$ components, and the general development might seem unnecessary for this problem. We've developed this process in the general setting, with potential modifications to the hierarchical model in mind. In [**?**], 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 $p = 7$. Additionally, the partially collapsed Gibbs samplers presented in [Van Dyk and Jiao, 2015; Van Dyk and Park, 2008] do not argue that the resulting Markov chains remain invariant, and the following discussion fills that gap.

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, $\mathcal{F}$ 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

$$\pi(\boldsymbol{x}_{\widehat{i}}) \stackrel{\text{def}}{=} \int_{x_i} \pi(x_1 \dots x_p) dx_i, \tag{3.3}$$

and *conditional distribution* is

$$\pi(\boldsymbol{x}_{\widehat{i}}|x_i) \stackrel{\text{def}}{=} \pi(\boldsymbol{x})/\pi(\boldsymbol{x}_{\widehat{i}}). \tag{3.4}$$

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

$$\mathbb{P}\left(\boldsymbol{X}^{k+1} \in A | \boldsymbol{X}^k = \boldsymbol{x}^k, \dots \boldsymbol{X}^0 = \boldsymbol{x}^0\right) = \mathbb{P}\left(\boldsymbol{X}^{k+1} \in A | \boldsymbol{X}^k = \boldsymbol{x}^k\right) \tag{3.5}$$

for all events A.

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

$$\int_A K(\boldsymbol{x}^k, \boldsymbol{x}')d\boldsymbol{x}' \stackrel{\text{def}}{=} \mathbb{P}\left(\boldsymbol{X}^{k+1} \in A | \boldsymbol{X}^k = \boldsymbol{x}^k\right), \tag{3.6}$$

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

$$\mathcal{K}[\pi](\boldsymbol{x}') = \int K(\boldsymbol{x}, \boldsymbol{x}')\pi(\boldsymbol{x})d\boldsymbol{x}. \tag{3.7}$$

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

$$\begin{aligned}
\pi(\boldsymbol{x}^1) &= \int_{\boldsymbol{x}^0} \pi(\boldsymbol{x}^1, \boldsymbol{x}^0)d\boldsymbol{x}^0 \\
&= \int_{\boldsymbol{x}^0} \pi(\boldsymbol{x}^1|\boldsymbol{x}^0)\pi(\boldsymbol{x}^0)d\boldsymbol{x}^0 \\
&= \int_{\boldsymbol{x}^0} K(\boldsymbol{x}^0, \boldsymbol{x}^1)\pi(\boldsymbol{x}^0)d\boldsymbol{x}^0 \\
&= \mathcal{K}[\pi_0](\boldsymbol{x}^1)
\end{aligned} \tag{3.8}$$

$$
\begin{aligned}
\pi(\boldsymbol{x}^2) &= \int_{\boldsymbol{x}^1}\int_{\boldsymbol{x}^0} \pi(\boldsymbol{x}^2,\boldsymbol{x}^1,\boldsymbol{x}^0)d\boldsymbol{x}^0 d\boldsymbol{x}^1 \\
&= \int_{\boldsymbol{x}^1} \pi(\boldsymbol{x}^2|\boldsymbol{x}^1)\int_{\boldsymbol{x}^0} \pi(\boldsymbol{x}^1,\boldsymbol{x}^0)d\boldsymbol{x}^0 d\boldsymbol{x}^1 \\
&= \int_{\boldsymbol{x}^1} K(\boldsymbol{x}^1,\boldsymbol{x}^2)\int_{\boldsymbol{x}^0} \pi(\boldsymbol{x}^1,\boldsymbol{x}^0)d\boldsymbol{x}^0 d\boldsymbol{x}^1 \\
&= \mathcal{K}\mathcal{K}[\pi_0](\boldsymbol{x}^2) \qquad (3.9)\\
&\vdots \\
\pi(\boldsymbol{x}^N) &= \int_{\boldsymbol{x}^{N-1}}\dots\int_{\boldsymbol{x}^0} \pi(\boldsymbol{x}^N,\boldsymbol{x}^{N-1}\dots\boldsymbol{x}^0)d\boldsymbol{x}_0\dots,d\boldsymbol{x}_{N-1} \\
&= \int_{\boldsymbol{x}^{N-1}}\dots\int_{\boldsymbol{x}^0} \pi(\boldsymbol{x}^N|\boldsymbol{x}^{N-1})\pi(\boldsymbol{x}^{N-1}\dots,\boldsymbol{x}^0)d\boldsymbol{x}_0\dots,d\boldsymbol{x}_{N-1} \\
&= \int_{\boldsymbol{x}^{N-1}} K(\boldsymbol{x}^{N-1},\boldsymbol{x}^N)\dots\int_{\boldsymbol{x}^0} K(\boldsymbol{x}^0,\boldsymbol{x}^1)\pi(\boldsymbol{x}^0)d\boldsymbol{x}_0\dots,d\boldsymbol{x}_{N-1} \\
&= \mathcal{K}^N[\pi_0](\boldsymbol{x}^N). \qquad (3.10)
\end{aligned}
$$

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 $K(\boldsymbol{x},\boldsymbol{x}')$. 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 *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

$$\mathcal{K}[\pi(\boldsymbol{x})](\boldsymbol{x}') = \pi(\boldsymbol{x}'). \tag{3.11}$$

Note that an invariant distribution π is a eigenvector for the transition operator $\mathcal{K}$ corresponding to the eigenvalue 1. 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 given by the recursive relation $\pi_k = K\pi_{k-1}/\|K\pi_{k-1}\|$ converges to the leading order eigenvector. Hence, the invariant density $\pi = \lim_{k\to\infty} K^n\pi$. 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, 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 in multiples of d. For example, the deterministic two state Markov chain associated with the transition matrix

$$K = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \tag{3.12}$$

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 3.1.1. *[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 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

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} h(\boldsymbol{X}^n) = \int h(\boldsymbol{x})\pi(\boldsymbol{x})d\boldsymbol{x}. \tag{3.13}$$

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

$$\lim_{N\to\infty} \|\mathcal{K}^N \pi_0 - \pi\|_{TV} = 0. \tag{3.14}$$

Equation (3.13) of the ergodic theorem is the analogous notion of the Law of Large Numbers for independent samples and allows us to use chain averages to estimate statistics about π. Equation (3.14) 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.

3.1.2 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 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 simulates the outcomes from the transition kernel

$$K(\boldsymbol{x}, \boldsymbol{x}') = \pi(x'_p|\boldsymbol{x}'_{\widehat{p}}) \dots \pi(x'_2|x'_1, \boldsymbol{x}_{\widehat{12}})\pi(x'_1|\boldsymbol{x}_{\widehat{1}}). \tag{3.15}$$

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

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

$$\begin{aligned}
\mathcal{K}[\pi_0](\boldsymbol{x}') &= \int K(\boldsymbol{x}, \boldsymbol{x}') \pi_0(\boldsymbol{x}) d\boldsymbol{x} \\
&= \int_{x_p} \dots \int_{x_1} \pi(x'_p | \boldsymbol{x}'_{\widehat{p}}) \dots \pi(x'_2 | x'_1, \boldsymbol{x}_{\widehat{12}}) \pi(x'_1 | \boldsymbol{x}_{\widehat{1}}) dx_1 \dots dx_p \\
&= \int_{x_p} \pi(x'_p | \boldsymbol{x}'_{\widehat{p}}) \int_{x_{p-1}} \pi(x'_{p-1} | \boldsymbol{x}'_{\widehat{p,p-1}} x_p) \dots \int_{x_1} \pi(x'_1 | \boldsymbol{x}_{\widehat{1}}) \pi_0(x_1 \dots x_p) dx_1 \dots dx_p.
\end{aligned} \tag{3.16}$$

Each integration in (3.16) can be thought of as a sub-transition on components of $\pi_0(x_1 \dots x_p)$; that is, given $(x_1 \dots x_{i-1}, x_{i+1} \dots x_p)$, let

$$\mathcal{K}_i[\pi_0(\boldsymbol{x})](\boldsymbol{x}') \stackrel{\text{def}}{=} \int_{x_i} \pi(x'_i | x'_1, \dots x'_{i-1}, x_{i+1} \dots x_p) \pi_0(\boldsymbol{x}) dx_i, \tag{3.17}$$

then we can express $\mathcal{K} = \mathcal{K}_p \mathcal{K}_{p-1} \dots \mathcal{K}_1$. 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 only after successively integrating each sub-transition is the operator uniquely defined. For example $\mathcal{K}_1$ depends on $(x_2 \dots x_p)$, $\mathcal{K}_2 K_1$ depends on $(x_3 \dots x_p)$, 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 (analyzing the transition kernel as a composition of conditional operators) will be useful for designing and verifying the stationarity of the proceeding algorithms.

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

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

$$\begin{aligned}\mathcal{K}_1[\pi(\boldsymbol{x})](\boldsymbol{x}') &= \int_{x_1} \pi(x_1'|x_2, \ldots x_p)\pi(x_1 \ldots x_p)dx_1 \\ &= \int_{x_1} \frac{\pi(x_1', x_2, \ldots x_p)\pi(x_1 \ldots x_p)}{\pi(x_1', x_2 \ldots x_p)}dx_1 \\ &= \pi(x_i', x_2, \ldots x_p). \end{aligned} \tag{3.18}$$

Moreover, given for fixed $(x_{i+1}, \ldots x_p)$, and when $\mathcal{K}_{i-1} \ldots \mathcal{K}_1 = \pi(x_1' \ldots x_{i-1}', x_i \ldots x_p)$, then

$$\begin{aligned}\mathcal{K}_i \ldots \mathcal{K}_1[\pi(\boldsymbol{x})](\boldsymbol{x}') &= \int_{x_i} \pi(x_i'|x_1', \ldots x_{i-1}', x_{i+1} \ldots x_p)\pi(x_1' \ldots x_{i-1}', x_i \ldots x_p)dx_i \\ &= \int_{x_i} \frac{\pi(x_1', \ldots x_i' \ldots x_p)\pi(x_1' \ldots x_{i-1}', x_i \ldots x_p)}{\pi(x_1', \ldots x_{i-1}', x_{i+1} \ldots x_p)}dx_i \\ &= \pi(x_1', x_2' \ldots x_i' \ldots x_p). \end{aligned} \tag{3.19}$$

By induction, $\mathcal{K}[\pi(\boldsymbol{x})](\boldsymbol{x}') = \mathcal{K}_p \ldots \mathcal{K}_1[\pi(\boldsymbol{x})](\boldsymbol{x}') = \pi(x_1', x_2' \ldots x_p')$. 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 $\pi(x_1 \ldots x_i|x_{i+1}, \ldots, x_p)$. To see this, when $(x_{i+1} \ldots x_p)$ are given, then $\pi(\boldsymbol{x})/\pi(x_{i+1} \ldots x_p) = \pi(x_1 \ldots x_i|x_{i+1} \ldots x_p)$ and since each integration does not depend on $(x_{i+1} \ldots x_p)$, by (3.19)

$$\mathcal{K}_i \ldots \mathcal{K}_1[\pi(x_1 \ldots x_i|x_{i+1} \ldots x_p)](\boldsymbol{x}') = \frac{\pi(x_1' \ldots x_p')}{\pi(x_{i+1} \ldots x_p)} = \pi(x_1' \ldots x_i'|x_{i+1} \ldots x_p). \tag{3.20}$$

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 $\pi(x_1 \ldots x_i|x_{i+1} \ldots x_p)$. We state this result formally:

Corollary 3.1.3. *Suppose $\mathcal{K} = \mathcal{K}_p \ldots \mathcal{K}_1$ is the transition operator for Algorithm 1, and*

$\widetilde{\mathcal{K}}_i$ given $\boldsymbol{x}_{\widehat{i}}$ is an operator such that $\widetilde{\mathcal{K}}_i\left[\pi(x_i|\boldsymbol{x}_{\widehat{i}})\right] = \pi(x_1'|\boldsymbol{x}_{\widehat{i}})$, then $\mathcal{K}_p \dots \widetilde{\mathcal{K}}_i \mathcal{K}_{i-1} \dots \mathcal{K}$ is invariant with respect to π.

Proof. Using (3.19) twice, we have

$$\begin{aligned}
\left(\mathcal{K}_p \dots \widetilde{\mathcal{K}}_i\right) \mathcal{K}_{i-1} \dots \mathcal{K}_1 \pi &= \mathcal{K}_p \dots \widetilde{\mathcal{K}}_i \pi(x_1' \dots x_{i-1}', x_i \dots x_p) \\
&= \mathcal{K}_p \dots \widetilde{\mathcal{K}}_i \pi(x_i|x_1' \dots x_{i-1}', x_{i+1} \dots x_p) \pi(x_1' \dots x_{i-1}' x_{i+1} \dots x_p) \\
&= \mathcal{K}_p \dots \mathcal{K}_{i+1} \pi(x_i'|x_1' \dots x_{i-1}', x_{i+1} \dots x_p) \pi(x_1' \dots x_{i-1}' x_{i+1} \dots x_p) \\
&= \mathcal{K}_p \dots \mathcal{K}_{i+1} \pi(x_1' \dots, x_i', x_{i+1} \dots x_p) \\
&= \pi(x_1' \dots x_p'). \qquad (3.21)
\end{aligned}$$

□

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.

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.

3.1.3 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 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 outline 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:

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

$$\mathcal{K}\pi_0 = \mathcal{K}_p \widetilde{\mathcal{K}}_{p-1} \dots \mathcal{K}_2 \mathcal{K}_1 \pi_0 \tag{3.22}$$

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

$$\widetilde{K}_{p-1}(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{x}}') \stackrel{\text{def}}{=} \widetilde{K}_{p-1}(\widetilde{\boldsymbol{x}}') \stackrel{\text{def}}{=} \pi(x'_{p-1}, \widetilde{x}'_p | x'_1, x'_2 \dots x'_{p-2}). \tag{3.23}$$

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

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

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

$$\begin{aligned}
&\widetilde{\mathcal{K}}\Big[\pi(\boldsymbol{x})\pi(\widetilde{x}_p|\boldsymbol{x})\Big](\widetilde{\boldsymbol{x}}') \\
&\qquad = \mathcal{K}_p \widetilde{\mathcal{K}}_{p-1} \mathcal{K}_{p-2} \dots \mathcal{K}_1 \big[\pi(\boldsymbol{x})\pi(\widetilde{x}_p|\boldsymbol{x})\big](\widetilde{\boldsymbol{x}}') \\
&\qquad = \int_{x_p} \pi(x_p' | \boldsymbol{x}'_{\widehat{p}}) \iint_{\widetilde{x}_p, x_{p-1}} \widetilde{K}_{p-1}(\widetilde{\boldsymbol{x}}') \int_{x_{p-2} \dots x_1} \cdots \int \dots \pi(\boldsymbol{x})\pi(\widetilde{x}_p|\boldsymbol{x}) dx_1 \dots d\widetilde{x}_p dx_p
\end{aligned} \tag{3.24}$$

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

of the first $p-2$ steps of the standard Gibbs sampler, continuing from (**??**) results in

$$\begin{aligned}
\widetilde{\mathcal{K}}\Big[\pi(\boldsymbol{x})\pi(\widetilde{x}_p|\boldsymbol{x})\Big](\widetilde{\boldsymbol{x}}') &= \int_{x_p} \pi(x'_p|\boldsymbol{x}'_{\widehat{p}}) \int_{x_{p-1}} \widetilde{K}_{p-1}(\widetilde{\boldsymbol{x}}') \cdot \mathcal{K}_{p-2} \dots \mathcal{K}_1[\pi(\boldsymbol{x})](\boldsymbol{x}') \\
&= \int_{x_p} \pi(x'_p|\boldsymbol{x}'_{\widehat{p}}) \int_{x_{p-1}} \pi(x'_{p-1}, \widetilde{x}'_p|x'_1, x'_2 \dots x'_{p-2}) \cdot \pi(x'_1 \dots x'_{p-2}, x_{p-1}, x_p) \\
&= \pi(x'_p|\boldsymbol{x}'_{\widehat{p}})\pi(x'_{p-1}, \widetilde{x}'_p|x'_1, x'_2 \dots x'_{p-2}) \cdot \pi(x'_1 \dots x'_{p-2}) \\
&= \frac{\pi(\boldsymbol{x}')\pi(x'_1, x'_2 \dots x'_{p-1}, \widetilde{x}'_p)}{\pi(\boldsymbol{x}'_{\widehat{p}})} \\
&= \pi(\boldsymbol{x}')\pi(\widetilde{x}'_p|\boldsymbol{x}'_{\widehat{p}}). \qquad (3.25)
\end{aligned}$$

□

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 of the conditioned variable at step p 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 effect on implementations where the conditioned variable appears later, since the implementation can be viewed as a Markov chain with the same transition kernel, that differs only in the initial distribution and that at step N, the kernel has only partially completed.

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

$$\int_{\widetilde{x}'_p} \widetilde{\mathcal{K}}[\pi(\boldsymbol{x})\pi(\widetilde{x}_p|\boldsymbol{x}_{\widehat{p}})](\widetilde{\boldsymbol{x}})d\widetilde{x}'_p = \pi(\boldsymbol{x}') \int_{\widetilde{x}'_p} \pi(x'_p|\boldsymbol{x}'_{\widehat{p}})dx'_p = \pi(\boldsymbol{x}'), \qquad (3.26)$$

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

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

By (3.26), 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. $X_p^k \sim \pi(x_p | x_1^k, x_2^k \dots x_{p-1}^k)$

The effect of this process is that we have removed conditioning of $X_p^{k-1} = x_p^{k-1}$ from the simulation of X_{p-1}^k. Note that the first $p-2$ steps of the algorithm can be permuted with the appropriate re-labeling with respect to k without changing the transition kernel.

We can generalize 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 chosen from $X_1 \dots X_{p-2}$ by permuting and relabeling. Hence, X_p can be partially collapsed out of any number of proceeding variables, and subsequently, X_{p-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 3.4), see if any conditioned variables can be partially collapsed. This choice is likely not obvious, unless guided by the specific situation (as in the hierarchical Gibbs sampler for sampling δ). If it is possible to sample the density with one of the conditioned 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] guarantees 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.

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 give a brief overview of the random walk Metropolis-Hastings method, and show that directly substituting a Metropolis-Hastings transition into the p-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 reversible 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 treatment.

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 reversible 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 $Y^k \sim \rho(y|x^k)$ called the *proposal* transition. The idea of the Metropolis-Hastings is, first generate a 'proposal' from a given transition operator ($\pi(y|x)$) from your current state, then if your guess 'improves' how likely you your next step is to be from the desired distribution π, then move there, otherwise stay put. At first glance, this algorithm might not seem useful since it requires a computation of a ratio of π, which we initially assumed was difficult to simulate. However, it is precisely because it appears as a *ratio* that makes the method useful – we need only know π up to a constant of proportionality since they cancel 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 *if*

$$K(x, x')\pi(x) = K(x', x)\pi(x'). \tag{3.28}$$

If K satisfies the detailed balance condition, then the corresponding Markov chain is invariant with respect to π.

Proof. The corresponding transition operator has

$$\mathcal{K}[\pi](x') = \int K(x, x')\pi(x)dx = \int K(x', x)\pi(x')dx = \pi(x') \tag{3.29}$$

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

The Metropolis-Hastings kernel is, in some sense, designed to satisfy detailed balance and [Calvetti and Somersalo, 2007] presents the development of the Metropolis-Hastings algorithm with that perspective. We summarize that discussion, to develop to give an explicit description of the transition kernel corresponding to Algorithm 4.

Proposition 3.1.6. *The Markov chain generated by Algorithm 4 is invariant with respect to* π.

Proof. Let the Markov chain associated with Algorithm 4 have $X_k = x_k$ given and $U = 1$ if the proposal is accepted and $U = 0$ otherwise. Then, for any event A

$$\begin{aligned}
\mathbb{P}\left(X^{k+1} \in A | X^k = x^k\right) &= \mathbb{P}\left(X^{k+1} \in A \text{ and } U = 1 | X^k = x^k\right) \\
&\quad + \mathbb{P}\left(X^{k+1} \in A \text{ and } U = 0 | X^k = x^k\right) \\
&= \mathbb{P}\left(Y^k \in A \text{ and } U = 1 | X^k = x^k\right) \\
&\quad + \mathbb{P}\left(x^k \in A \text{ and } U = 0 | X^k = x^k\right). \qquad (3.30)
\end{aligned}$$

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

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

so continuing from (3.30),

$$= \int_A \alpha(x^k, y)\rho(y|x^k)I_A(x^k) \int (1 - \alpha(x^k, y'))\rho(y|x^k))dy' \tag{3.31}$$

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

$$K(x, y) = \alpha(x, y)\rho(y|x)\delta_x(y) \left(1 - \int \alpha(x, y')\rho(y'|x)dy'\right). \tag{3.32}$$

Now, in order to show that $K(x, y)$ satisfies the detailed balance equation, it suffices to show it for each term in (3.32). If $\pi(y) >= \pi(x)$ then $\alpha(x, y) = 1$ and $\alpha(y, x) = \pi(x)/\pi(y)$ so

$$\alpha(x, y)\rho(y|x)\pi(x) = \rho(x|y)\pi(x) = \frac{\pi(x)}{\pi(y)}\rho(x|y)\pi(y) = \alpha(y, x)\rho(x|y)\pi(y). \tag{3.33}$$

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

$$f(x) \int_A \delta_x(y)\pi(y)dy = f(x)I_A(x)\pi(x) = \pi(x) \int_A \delta_x(y)f(y)dy \tag{3.34}$$

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

Combining Proposition 3.1.6 and Corollary 3.1.3, proves the invariance with respect to π of Algorithm 5.

Algorithm 5 Metropolis Hastings within 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. Simulate X_{p-1}^k from Algorithm 4 for $\pi(x_{p-1} | x_1^k, x_2^k \dots x_{p-2}^k)$

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

One implementation question remains as to whether to iterate step p-1. to obtain 'better' simulations from $\pi(x_{p-1} | x_1^k, x_2^k \dots x_{p-2}^k)$. When implemented in standard Gibbs sampling, [Robert and Casella, 2013] recommend only one simulation. Although, in [Van Dyk and Jiao, 2015], they recommend that iterating the Metropolis step improves convergence rates. Can we view the iteration as a delayed rejection for a one step process? If so, then there is theory that says there is a convergence improvement (which is intuitively obvious, and in our case comes with an added computational cost). Moreover, if we add an adaptive step, is this related to the DRAM algorithm?

As we've shown, each scheme is invariant with respect to π, however, as we will see with PSF estimation, generating the proposal for Algorithm 4 may be computationally expensive, and the improvement in convergence may not 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 3.4, we will address these issues explicitly for PSF estimation.

We now return to the problem of PSF estimation, where we will explicitly implement and describe Gibbs sampling and Metropolis Hastings within partially collapsed Gibbs sampling for PSF reconstruction.

3.2 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 a 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.

3.2.1 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 $\mathscr{H}_1$ and $\mathscr{H}_2$. Both methods assume p and g are functions with known evaluations on a discrete grid. The error analysis associated with these methods are based on Taylor expansions, which assume 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 since p and b are not functions and Taylor's theorem is not available. In fact, any element with discrete support in $\mathscr{H}_1$ and $\mathscr{H}_2$ is equivalent to 0, since each is a subspace of an L^2 space. Yet, we justify our use of quadrature methods by recalling from Chapter 2 that smooth functions are dense in $\mathscr{H}_1$ and $\mathscr{H}_2$. In theory, one could use that result to construct a smooth approximation, then apply quadrature on the approximation in order to explicitly control the error of the approximation. 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 that is negligible.

We mention that there are other methods that are theoretically more appealing which use a truncation of orthonormal bases of $\mathscr{H}_1$ and $\mathscr{H}_2$, 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 eigenvector, hence elements of $\mathscr{H}_1$ 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 – both analytically difficult. I did this a while ago with MAP estimation for fixed lambda and delta, and the results were not good. I numerically estimated both roots (using a code I found on the internet) and numerically estimated the forward operator on the Bessel functions. It produces a crappy estimate (worse than what we have estimates for), but then again I didn't put a lot into estimating λ and δ. Should I mention this?

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 $h \stackrel{\text{def}}{=} \frac{1}{N}$. 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, the integral kernel for $\mathcal{G}$ was

$$g(x, r) = \begin{cases} 0 & x < -r \\ 2(\pi - \operatorname{acos}(x/r)) & |x| \leq r \\ 2\pi & x > r \end{cases} . \tag{3.35}$$

For a fixed $x_i < 0$, we have

$$[\mathcal{G}p](x_i) = \int_0^{-x_i} p(r) 2(\pi - \operatorname{acos}(x_i/r))\, dr. \tag{3.36}$$

As in [Bardsley, 2012], we discretize the integral using midpoint quadrature which guarantees a second-order integration method. Because the upper bound in (3.36) 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 $\operatorname{acos}(t) = 0$ for $t < -1$

$$\begin{aligned}[\mathcal{G}p](x_i) &\approx \sum_{j=1}^{-i} p(r_j) 2(\pi - \operatorname{acos}(x_i/r_j)) r_j h \\ &= \sum_{j=1}^{N} p(r_j) 2(\pi - \operatorname{acos}(x_i/r_j)) r_j h. \end{aligned} \tag{3.37}$$

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

$$\begin{aligned}[\mathcal{G}p](x_i) &= 2\pi \int_0^{x_i} p(r)\, r dr + \int_{x_i}^{\infty} p(r) 2(\pi - \operatorname{acos}(x_i/r))\, r dr \\ &= \int_0^{x_i} p(r) 2(\pi - \operatorname{acos}(x_i/r))\, r dr + \int_{x_i}^{\infty} p(r) 2(\pi - \operatorname{acos}(x_i/r))\, dr \\ &\approx \sum_{j=1}^{N} p(r_j) 2(\pi - \operatorname{acos}(x_i/r_j)) r_j h. \end{aligned} \tag{3.38}$$

Now, let $\boldsymbol{p}$ be the $N \times 1$ column vector with entries $\boldsymbol{p}_j = p(r_j)$, then using (3.37) and (3.38), if we define the matrix $\boldsymbol{G}$ with entries $\boldsymbol{G}_{ij} \stackrel{\text{def}}{=} 2(\pi - \operatorname{acos}(x_i/r_j)) r_j h$, the quadrature approximation can be expressed by the matrix-vector equation $\boldsymbol{b} \approx \boldsymbol{G}\boldsymbol{p}$. Finally, we use $\|f\|_{\mathscr{H}_2} \approx h\|\boldsymbol{f}\|_{\mathbb{R}^{2N+1}}$ to approximate the norm in $\mathscr{H}_2$, where elements of $\boldsymbol{f}$ are point-wise evaluations of f. Combining these approximations, we have for the first term in (3.1),

$$\lambda \|b - \mathcal{G}p\|_{\mathscr{H}_2} \approx \lambda h \|\boldsymbol{b} - \boldsymbol{G}\boldsymbol{p}\|_{\mathbb{R}^{2N+1}}. \tag{3.39}$$

The negative radial Laplacian $\mathcal{L} : \mathscr{H}_1 \to \mathscr{H}_1$ operates on continuous functions by

$$[\mathcal{L}p](r) = -r^{-1}\frac{d}{dr}\left(r\,\frac{d}{dr}p(r)\right). \tag{3.40}$$

For definiteness, we take $n = 2$ in the inner product in (3.1), so

$$\begin{aligned}\langle p, \mathcal{L}^2 p\rangle_{\mathscr{H}_1} &= 2\pi\int_0^\infty p(r)[\mathcal{L}^2 p](r)\, r dr \\ &= 2\pi\int_0^\infty p(r)\left[\frac{d}{dr}\left(r\,\frac{d}{dr}p(r)\right)\right]^2 r^{-1}dr.\end{aligned} \tag{3.41}$$

The discretization of (3.41) will occur in two steps. We will use quadrature to estimate the integral in (3.41), 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.

Following similar arguments in [Morton and Mayers, 2005] for approximating a general parabolic operator, let $r_{j\pm1/2} \stackrel{\text{def}}{=} r_j \pm \frac{h}{2}$, then a center differencing scheme leads to the approximation

$$\left[\frac{d}{dr}r\frac{d}{dr}p\right]_{r_j} \approx \frac{1}{h}\left(\left[r\frac{d}{dr}p\right]_{r_{j-1/2}} + \left[r\frac{d}{dr}p\right]_{r_{j+1/2}}\right). \tag{3.42}$$

The center difference approximation of the first term is

$$\left[r\frac{d}{dr}p\right]_{r_{j-1/2}} \approx r_{j-1/2}\frac{p_{j-1}-p_j}{h} \tag{3.43}$$

and of the second term is

$$\left[r\frac{d}{dr}p\right]_{r_{j+1/2}} \approx r_{j+1/2}\frac{p_j-p_{j+1}}{h}. \tag{3.44}$$

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

$$[\boldsymbol{Rp}]_j \stackrel{\text{def}}{=} \frac{1}{h^2}\Big(r_{j+1/2}(p_{j+1}-p_j) - r_{j-1/2}(p_j-p_{j-1})\Big). \tag{3.45}$$

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

$$\frac{1}{h^2}\begin{bmatrix} -(r_{j-3/2}+r_{j-1/2}) & r_{j-1/2} & 0 \\ r_{j-1/2} & -(r_{j-1/2}+r_{j+1/2}) & r_{j+1/2} \\ 0 & r_{j+1/2} & -(r_{j+1/2}+r_{j+3/2}) \end{bmatrix}\begin{bmatrix} p_{j-1} \\ p_j \\ p_{j+1} \end{bmatrix}. \tag{3.46}$$

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

$$[\boldsymbol{Rp}]_N = r_{N-1/2}p_{N-1} - (r_{N-1/2}+r_{N+1/2})p_N. \tag{3.47}$$

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

$$\begin{aligned}[\boldsymbol{Rp}]_1 &= r_{1/2}p_0 - (r_{1/2}+r_{3/2})p_1 + r_{3/2}p_2 \\ &= r_{3/2}p_1 + r_{3/2}p_2.\end{aligned} \tag{3.48}$$

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

We then take $\boldsymbol{L} \stackrel{\text{def}}{=} \left(\text{diag}(\boldsymbol{r}^{-1/2})\boldsymbol{R}\right)^2$ where $\text{diag}(\boldsymbol{r}^{-1/2})$ denotes the $N \times N$ diagonal matrix whose diagonal entires are $(r_j^{-1/2})$. 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 (3.41) with

$$\langle p, \mathcal{L}^2 p\rangle_{\mathscr{H}_1} \approx 2\pi h \langle \boldsymbol{p}, \boldsymbol{Lp}\rangle_{\mathbb{R}^N}. \tag{3.49}$$

So the complete approximation to (3.1) is

$$\Phi(p; b, \lambda, \delta) \approx \lambda h \|\boldsymbol{G}\boldsymbol{p} - \boldsymbol{b}\|^2_{\mathbb{R}^{2N+1}} + \delta 2\pi h \langle \boldsymbol{p}, \boldsymbol{L}\boldsymbol{p} \rangle_{\mathbb{R}^N}. \tag{3.50}$$

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

$$F(\boldsymbol{p}; \boldsymbol{b}, \lambda, \delta) \stackrel{\text{def}}{=} \frac{1}{2} \left(\lambda \|\boldsymbol{G}\boldsymbol{p} - \boldsymbol{b}\|^2_{\mathbb{R}^{2N+1}} + \delta \langle \boldsymbol{p}, \boldsymbol{L}\boldsymbol{p} \rangle_{\mathbb{R}^N} \right). \tag{3.51}$$

3.2.2 The discrete hierarchical posterior distribution

As in [Bardsley, 2012], we employ a hierarchical model for λ and δ that employ 'uninformative' independent prior distributions that form a natural conjugacy so that the resulting full conditional densities will be known 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, or integrate, 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. If clarification is needed, they will be appropriately subscripted. 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, should be clear in context.

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

$$\pi(\boldsymbol{p}|\delta) = (2\pi)^{-N/2} \, \delta^N \, |\det L|^{-1/2} \, |\exp\left(-\frac{\delta}{2} \langle \boldsymbol{p}, \boldsymbol{L}\boldsymbol{p} \rangle_{\mathbb{R}^N} \right), \tag{3.52}$$

$$\pi(\boldsymbol{b}|\boldsymbol{p},\lambda) = (2\pi)^{-(2N+1)/2}\,\lambda^N \exp\left(-\frac{\lambda}{2}\|\boldsymbol{G}\boldsymbol{p}-\boldsymbol{b}\|^2\right), \tag{3.53}$$

and

$$\pi(\boldsymbol{p}|\boldsymbol{b},\lambda,\delta) \propto \exp\Big(-F(\boldsymbol{p};\boldsymbol{b},\lambda,\delta)\Big). \tag{3.54}$$

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 $\pi(\boldsymbol{p}|\delta)$ is independent of the noise parameter given δ and that the likelihood $\pi(\boldsymbol{b}|\boldsymbol{b},\lambda)$ is independent of the prior parameter δ given $\boldsymbol{b}$ and λ, i.e.

$$\pi(\boldsymbol{p}|\lambda,\delta) = \pi(\boldsymbol{p}|\lambda,\delta) \tag{3.55}$$

$$\text{and } \pi(\boldsymbol{b}|\boldsymbol{p},\lambda,\delta) = \pi(\boldsymbol{b}|\boldsymbol{p},\lambda). \tag{3.56}$$

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 family that can be made 'uninformative' by appropriately chosen parameters. Assuming λ and δ are independent gamma-distributed random variables, they have probability density functions

$$\pi(\lambda) \propto \lambda^{\alpha-1}\,\exp(-\beta\lambda) \tag{3.57}$$

$$\text{and } \pi(\delta) \propto \delta^{\alpha-1}\,\exp(-\beta\delta), \tag{3.58}$$

where as recommended by [Gelman et al., 2014], we use parameter values $\alpha = 1$ and $\beta = 10^{-4}$. Now, applying Bayes' theorem and the definition of conditional probability, the *joint posterior*

density is

$$\begin{aligned}\pi(\boldsymbol{p},\lambda,\delta|\boldsymbol{b}) &= \frac{\pi(\boldsymbol{b}|\boldsymbol{p},\lambda,\delta)\pi(\boldsymbol{p},\lambda,\delta)}{\pi(\boldsymbol{b})} \\ &= \frac{\pi(\boldsymbol{b}|\boldsymbol{p},\lambda,\delta)\pi(\boldsymbol{p}|\delta,\lambda)\pi(\lambda,\delta)}{\pi(\boldsymbol{b})} \\ &= \frac{\pi(\boldsymbol{b}|\boldsymbol{p},\lambda)\pi(\boldsymbol{p}|\delta)\pi(\lambda)\pi(\delta)}{\pi(\boldsymbol{b})} \\ &\propto \lambda^{(2N+1)/2+\alpha-1}\delta^{N/2+\alpha-1}\exp\left(-\frac{\lambda}{2}\|\boldsymbol{G}\boldsymbol{p}-\boldsymbol{b}\|^2-\frac{\delta}{2}\langle\boldsymbol{p},\boldsymbol{L}\boldsymbol{p}\rangle-\beta\lambda-\beta\delta\right). \quad (3.59)\end{aligned}$$

Our primary goal for estimation and uncertainty quantification of $\boldsymbol{p}$ will be drawing inference from (3.59). 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

$$\begin{aligned}\pi(\lambda|\boldsymbol{b},\boldsymbol{p},\delta) = \frac{\pi(\boldsymbol{p},\lambda,\delta|\boldsymbol{b})}{\pi(\boldsymbol{p},\delta|\boldsymbol{b})} &\propto \lambda^{(2N+1)/2+\alpha-1}\exp\left(-\lambda\left(\frac{1}{2}\|\boldsymbol{G}\boldsymbol{x}-\boldsymbol{b}\|^2-\beta\right)\right), \\ \text{and } \pi(\delta|\boldsymbol{b},\boldsymbol{p},\lambda) = \frac{\pi(\boldsymbol{p},\lambda,\delta|\boldsymbol{b})}{\pi(\boldsymbol{p},\lambda|\boldsymbol{b})} &\propto \delta^{N/2+\alpha-1}\exp\left(-\delta\left(\frac{1}{2}\langle\boldsymbol{p},\boldsymbol{L}\boldsymbol{p}\rangle-\beta\right)\right), \quad (3.60)\end{aligned}$$

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

$$\boldsymbol{J}_{\lambda,\delta} \stackrel{\text{def}}{=} (\lambda\boldsymbol{G}^T\boldsymbol{G}+\delta\boldsymbol{L}) \quad (3.61)$$

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

$$\boldsymbol{m}_{\lambda,\delta} \stackrel{\text{def}}{=} \boldsymbol{J}_{\lambda,\delta}^{-1}\lambda\boldsymbol{G}^T\boldsymbol{b}, \tag{3.62}$$

then observe

$$\begin{aligned}
2F(\boldsymbol{p};\boldsymbol{b},\lambda,\delta) &= \lambda\|\boldsymbol{G}\boldsymbol{x}-\boldsymbol{b}\|^2 + \delta\,\langle \boldsymbol{p}, \boldsymbol{L}\boldsymbol{p}\rangle \\
&= \lambda\langle \boldsymbol{G}\boldsymbol{p}, \boldsymbol{G}\boldsymbol{p}\rangle - 2\lambda\langle \boldsymbol{G}\boldsymbol{p}, \boldsymbol{b}\rangle + \lambda\langle \boldsymbol{b},\boldsymbol{b}\rangle + \delta\langle \boldsymbol{p}, \boldsymbol{L}\boldsymbol{p}\rangle \\
&= \left\langle \boldsymbol{p}, (\lambda\boldsymbol{G}^T\boldsymbol{G} + \delta\boldsymbol{L})\boldsymbol{p}\right\rangle - 2\lambda\left\langle \boldsymbol{p}, \boldsymbol{G}^T\boldsymbol{b}\right\rangle + \lambda\|\boldsymbol{b}\|^2 \\
&= \langle \boldsymbol{p}, \boldsymbol{J}_{\lambda,\delta}\boldsymbol{p}\rangle - 2\left\langle \boldsymbol{p}, \boldsymbol{J}_{\lambda,\delta}\boldsymbol{m}_{\lambda,\delta}\right\rangle + \lambda\|\boldsymbol{b}\|^2 \\
&= \langle \boldsymbol{p}, \boldsymbol{J}_{\lambda,\delta}(\boldsymbol{p}-\boldsymbol{m}_{\lambda,\delta})\rangle - \langle \boldsymbol{p}, \boldsymbol{J}_{\lambda,\delta}\boldsymbol{m}_{\lambda,\delta}\rangle + \lambda\|\boldsymbol{b}\|^2 \\
&= \langle (\boldsymbol{p}-\boldsymbol{m}_{\lambda,\delta}), \boldsymbol{J}_{\lambda,\delta}(\boldsymbol{p}-\boldsymbol{m}_{\lambda,\delta})\rangle + \langle \boldsymbol{m}_{\lambda,\delta}, \boldsymbol{J}_{\lambda,\delta}\boldsymbol{p}\rangle \\
&\qquad - \langle \boldsymbol{m}_{\lambda,\delta}, \boldsymbol{J}_{\lambda,\delta}\boldsymbol{m}_{\lambda,\delta}\rangle - \langle \boldsymbol{p}, \boldsymbol{J}_{\lambda,\delta}\boldsymbol{m}_{\lambda,\delta}\rangle + \lambda\|\boldsymbol{b}\|^2 \\
&= \langle (\boldsymbol{p}-\boldsymbol{m}_{\lambda,\delta}), \boldsymbol{J}_{\lambda,\delta}(\boldsymbol{p}-\boldsymbol{m}_{\lambda,\delta})\rangle - \langle \boldsymbol{m}_{\lambda,\delta}, \boldsymbol{J}_{\lambda,\delta}\boldsymbol{m}_{\lambda,\delta}\rangle + \lambda\|\boldsymbol{b}\|^2,
\end{aligned} \tag{3.63}$$

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

$$\begin{aligned}
\pi(\boldsymbol{p}|\boldsymbol{b},\lambda,\delta) &= \frac{\pi(\boldsymbol{p},\lambda,\delta|\boldsymbol{b})}{\pi(\lambda,\delta|\boldsymbol{b})} \\
&\propto \exp\left(-\frac{1}{2}\left\langle (\boldsymbol{p}-\boldsymbol{m}_{\lambda,\delta}), \boldsymbol{J}_{\lambda,\delta}(\boldsymbol{p}-\boldsymbol{m}_{\lambda,\delta})\right\rangle\right)
\end{aligned} \tag{3.64}$$

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

3.3 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. 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.

The general idea for random variable generation involves transforming a uniform random variable from $[0, 1]$ into a random variable with the desired distribution. 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

$$\begin{aligned} \mathbb{P}(X \leq x) &= \mathbb{P}(F^{-1}(U) \leq x) \\ &= \mathbb{P}(U \leq F(x)) \\ &= \int_0^{F(x)} ds \\ &= F(x). \end{aligned} \tag{3.65}$$

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 simulations 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.

3.3.1 Gibbs sampling the PSF posterior

Here, we describe how to explicitly obtain simulations from the full conditional densities $\pi(\lambda|\boldsymbol{b},\boldsymbol{p},\delta)$, $\pi(\lambda|\boldsymbol{b},\boldsymbol{p},\delta)$, and $\pi(\boldsymbol{p}|\boldsymbol{b},\lambda,\delta)$. The equations in (3.60) imply that both $\pi(\lambda|\boldsymbol{b},\boldsymbol{p},\delta)$, $\pi(\delta|\boldsymbol{b},\boldsymbol{p},\lambda)$ are gamma-distributed as

$$\begin{aligned}&\Gamma\left((2N+1)/2+\alpha,\frac{1}{2}\|\boldsymbol{G}\boldsymbol{x}+\boldsymbol{b}\|^2+\beta\right)\\ \text{and }&\Gamma\left(N/2+\alpha,\frac{1}{2}\langle\boldsymbol{p},\boldsymbol{L}\boldsymbol{p}\rangle+\beta\right)\end{aligned}\tag{3.66}$$

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 simulations from $\mathcal{N}(0,1)$. Hence, $\boldsymbol{z}$ is a simulation of a multivariate Gaussian $\mathcal{N}(\boldsymbol{0},\boldsymbol{I}_{N\times N})$. An important feature of positive definite matrices, $\boldsymbol{A}$, is that they have an eigenvalue decomposition of the form $\boldsymbol{U}\Lambda\boldsymbol{U}^*$ (here * denotes the conjugate transpose since columns of $\boldsymbol{U}$ may be complex valued), where the columns $\boldsymbol{U}^*$ are mutually orthonormal with $\boldsymbol{\Lambda}$ a diagonal matrix of positive eigenvalues. Therefore, there exists a matrix $\boldsymbol{M}=\boldsymbol{U}\boldsymbol{\Lambda}^{-1/2}\boldsymbol{U}^*$, 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 transform $\boldsymbol{M}\boldsymbol{z}\sim\mathcal{N}(0,\boldsymbol{A}^{-1})$. 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 $\boldsymbol{A}=\boldsymbol{M}\boldsymbol{M}^T$ 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

$$\boldsymbol{R}_{\lambda,\delta}\boldsymbol{R}_{\lambda,\delta}^T\overset{\text{def}}{=}\boldsymbol{J}_{\lambda,\delta}.\tag{3.67}$$

With $\boldsymbol{R}_{\lambda,\delta}$ in hand, it serves two purposes – first we can solve $\boldsymbol{J}_{\lambda,\delta}\boldsymbol{m}_{\lambda,\delta}=\boldsymbol{R}\boldsymbol{R}^T\boldsymbol{m}_{\lambda,\delta}=$

$\boldsymbol{G}^T\boldsymbol{b}$ 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 simulation from $\pi(\boldsymbol{p}|\boldsymbol{b},\lambda,\delta)$.

Note that each time a simulation 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 my 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 $\boldsymbol{R}_{\lambda^{k+1},\delta^{k+1}}(3.67), \boldsymbol{m}_{\lambda^{k+1},\delta^{k+1}}(3.62)$, and set $\boldsymbol{p}^{k+1} = \boldsymbol{R}_{\lambda^{k+1},\delta^{k+1}}^{-1}\boldsymbol{z} + \boldsymbol{m}_{\lambda^{k+1},\delta^{k+1}}$ where $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{N\times N})$.

3.3.2 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.

This will require a simulation from the density $\pi(\delta|\boldsymbol{b},\lambda)$. To express the kernel of this

density, note (3.64) is the kernel of a Gaussian with mean $\boldsymbol{m}_{\lambda,\delta}$ and variance $\boldsymbol{J}_{\lambda,\delta}^{-1}$, thus the normalized density is

$$\begin{aligned}\frac{\pi(\boldsymbol{p},\lambda,\delta|\boldsymbol{b})}{\pi(\lambda,\delta|\boldsymbol{b})} &= \pi(\boldsymbol{p}|\boldsymbol{b},\lambda,\delta)\\ &= (2\pi)^{-N/2}|\det \boldsymbol{J}_{\lambda,\delta}|^{1/2}\exp\left(-\frac{1}{2}\langle \boldsymbol{p}-\boldsymbol{m}_{\lambda,\delta}, \boldsymbol{J}_{\lambda,\delta}(\boldsymbol{p}-\boldsymbol{m}_{\lambda,\delta})\rangle\right). \end{aligned} \tag{3.68}$$

Dividing (3.59) by (3.68), one obtains

$$\begin{aligned}\pi(\lambda,\delta|\boldsymbol{b}) &= \frac{\pi(\boldsymbol{p},\lambda,\delta|\boldsymbol{b})}{\pi(\boldsymbol{p}|\boldsymbol{b},\lambda,\delta)}\\ &\propto \lambda^{\frac{2N+1}{2}+\alpha-1}\delta^{\frac{N}{2}+\alpha-1}|\det \boldsymbol{J}_{\lambda,\delta}|^{-1/2}\\ &\quad\times\exp\left(\frac{1}{2}\langle \boldsymbol{p}-\boldsymbol{m}_{\lambda,\delta}, \boldsymbol{J}(\boldsymbol{p}-\boldsymbol{m}_{\lambda,\delta})\rangle - F(\boldsymbol{p};\boldsymbol{b},\lambda,\delta)-\beta\lambda-\beta\delta\right)\\ &=\propto \lambda^{\frac{2N+1}{2}+\alpha-1}\delta^{\frac{N}{2}+\alpha-1}|\det \boldsymbol{J}_{\lambda,\delta}|^{-1/2}\\ &\quad\times\exp\left(-\frac{1}{2}\left(\lambda\|\boldsymbol{b}\|^2-\langle \boldsymbol{m}_{\lambda,\delta}, \boldsymbol{J}_{\lambda,\delta}\boldsymbol{m}_{\lambda,\delta}\rangle\right)-\beta\lambda-\beta\delta\right). \end{aligned} \tag{3.69}$$

Finally,

$$\begin{aligned}\pi(\delta|\boldsymbol{b},\lambda) &= \frac{\pi(\lambda,\delta|\boldsymbol{b})}{\pi(\lambda|\boldsymbol{b})}\\ &\propto \delta^{\frac{N}{2}+\alpha}|\det \boldsymbol{J}_{\lambda,\delta}|^{-1/2}\exp\left(-\frac{1}{2}\left(\lambda\|\boldsymbol{b}\|^2-\langle \boldsymbol{m}_{\lambda,\delta}, \boldsymbol{J}_{\lambda,\delta}\boldsymbol{m}_{\lambda,\delta}\rangle\right)-\beta\delta\right). \end{aligned} \tag{3.70}$$

The two terms $|\det \boldsymbol{J}_{\lambda,\delta}|$ and $\langle \boldsymbol{m}_{\lambda,\delta}, \boldsymbol{J}_{\lambda,\delta}\boldsymbol{m}_{\lambda,\delta}\rangle$ in (3.70) make the density depend in a complicated way in δ, so a direct simulation is not available. Additionally, they are potentially 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 computed with (3.70). 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 (3.70) into terms that depend only on expensive quantities $\boldsymbol{R}_{\lambda,\delta}$ and $\boldsymbol{m}_{\lambda,\delta}$. That is,

$$\begin{aligned} -\frac{1}{2}\left(\lambda\|\boldsymbol{b}\|^2 - \langle \boldsymbol{m}_{\lambda,\delta}, \boldsymbol{J}_{\lambda,\delta}\boldsymbol{m}_{\lambda,\delta}\rangle\right) &= -\frac{1}{2}\left(\lambda\langle \boldsymbol{b},\boldsymbol{b}\rangle - \langle \boldsymbol{m}_{\lambda,\delta}, \lambda\boldsymbol{G}^T\boldsymbol{b}\rangle\right) \\ &= -\frac{\lambda}{2}\langle \boldsymbol{b} - \boldsymbol{G}\boldsymbol{m}_{\lambda,\delta}, \boldsymbol{b}\rangle \\ &\stackrel{\text{def}}{=} -\frac{\lambda}{2}a(\boldsymbol{m}_{\lambda,\delta}), \end{aligned} \tag{3.71}$$

and

$$\begin{aligned} \ln(|\det \boldsymbol{J}_{\lambda,\delta}|^{-1/2}) &= -\frac{1}{2}\ln(|\det \boldsymbol{R}_{\lambda,\delta}\boldsymbol{R}_{\lambda,\delta}^T|) \\ &= -\frac{1}{2}\ln(|\det \boldsymbol{R}_{\lambda,\delta}|^2) \\ &= -\frac{1}{2}\ln\left(\prod_{i=1}^{N}|\boldsymbol{R}_{\lambda,\delta_{ii}}|^2\right) \\ &= -\sum_{i=1}^{N}\ln|\boldsymbol{R}_{\lambda,\delta_{ii}}| \\ &\stackrel{\text{def}}{=} -b(\boldsymbol{R}_{\lambda,\delta}), \end{aligned} \tag{3.72}$$

where we used the fact that $\boldsymbol{R}_{\lambda,\delta}$ is lower triangular to compute the determinant. Substituting these expressions into (3.70)

$$\begin{aligned} \pi(\delta|\boldsymbol{b},\lambda) &\propto \delta^{\frac{N}{2}+\alpha}|\det \boldsymbol{J}_{\lambda,\delta}|^{-1/2}\exp\left(-\frac{\lambda}{2}\langle \boldsymbol{b}-\boldsymbol{G}\boldsymbol{m}_{\lambda,\delta},\boldsymbol{b}\rangle - \beta\delta.\right) \\ &= \exp\left(\left(\frac{N}{2}+\alpha-1\right)\ln\delta - b(\boldsymbol{R}_{\lambda,\delta}) - \frac{\lambda}{2}a(\boldsymbol{m}_{\lambda,\delta}) - \beta\delta\right) \\ &\stackrel{\text{def}}{=} \exp\Big(c(\boldsymbol{R}_{\lambda,\delta},\boldsymbol{m}_{\lambda,\delta},\delta)\Big). \end{aligned} \tag{3.73}$$

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

$$\alpha(\delta, \delta') = \min\left\{1, \frac{\pi(\delta'|\boldsymbol{b}, \lambda)}{\pi(\delta|\boldsymbol{b}, \lambda)}\right\} \tag{3.74}$$

is equivalent to accepting with log uniform probability

$$\ln \alpha(\delta, \delta') = \min\left\{0, c(\boldsymbol{R}_{\lambda,\delta'}, \boldsymbol{m}_{\lambda,\delta'}, \delta') - c(\boldsymbol{R}_{\lambda,\delta}, \boldsymbol{m}_{\lambda,\delta}, \delta)\right\} \tag{3.75}$$

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 the Metropolis-Hastings within PCG in Algorithm 5. The full implementation is described in Algorithm 7. Note that we were able to re-use the factorization $\boldsymbol{R}_{\lambda,\delta}$ and $\boldsymbol{m}_{\lambda,\delta}$ to sample $\boldsymbol{p}^{k+1}$, so there are $M+1$ 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 $\boldsymbol{R}_{\lambda,\delta}(3.67), \boldsymbol{m}_{\lambda,\delta}(3.62)$, then $c(\boldsymbol{R}_{\lambda,\delta}, \boldsymbol{m}_{\lambda,\delta}, \delta)(3.73)$.
   For $j = 1 \ldots M$
   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,

$$4.\ \text{Set } \gamma^2 = \frac{1}{k+2}\sum_{i=1}^{k+1}(\delta^i - \overline{\boldsymbol{\delta}_k})^2 \tag{3.76}$$

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 Chapter 4 with the algorithm with a tuned γ^2, with much less 'tunning-effort.' 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 estimates.

3.3.3 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 (3.59) factors in λ and δ, and since $\pi(\lambda, \delta|\boldsymbol{bp}) \propto \pi(\lambda, \delta, \boldsymbol{p}|\boldsymbol{b})$, 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 (3.73), by using (3.58)

in

$$\begin{aligned}\pi(\lambda,\delta|\boldsymbol{b}) &= \pi(\delta|\boldsymbol{b},\lambda)\pi(\lambda|\boldsymbol{b}) \\ &\propto \pi(\delta|\boldsymbol{b},\lambda)\pi(\lambda) \\ &= \exp\left(\left(\frac{2N+1}{2}+\alpha-1\right)\ln\lambda + \left(\frac{N}{2}+\alpha-1\right)\ln\delta - b(\boldsymbol{R}_{\lambda,\delta}) - \frac{\lambda}{2}a(\boldsymbol{m}_{\lambda,\delta}) - \beta\lambda - \beta\delta\right) \\ &\stackrel{\text{def}}{=} \exp\Big(c(\boldsymbol{R}_{\lambda,\delta},\boldsymbol{m}_{\lambda,\delta},\lambda,\delta)\Big). \end{aligned} \tag{3.77}$$

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 $\boldsymbol{R}_{\lambda,\delta}(3.67), \boldsymbol{m}_{\lambda,\delta}(3.62)$, then $c(\boldsymbol{R}_{\lambda,\delta},\boldsymbol{m}_{\lambda,\delta},\lambda,\delta)(3.77)$.
   For $j = 1 \ldots M$
   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 are addressed in Section 3.4 on evaluating chain convergence. 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 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 (3.71), using (3.62) then (3.61)

$$\begin{aligned} a(\boldsymbol{m}_{\lambda,\delta}) &= \langle \boldsymbol{b} - \boldsymbol{G}\boldsymbol{m}_{\lambda,\delta}, \boldsymbol{b} \rangle \\ &= \left\langle \boldsymbol{b}, \left(I - \boldsymbol{G}\boldsymbol{J}_{\lambda,\delta}^{-1} \lambda \boldsymbol{G}^T \right) \boldsymbol{b} \right\rangle \\ &= \left\langle \boldsymbol{b}, \left(I - \boldsymbol{G} \left(\boldsymbol{G}^T\boldsymbol{G} + \frac{\delta}{\lambda}\boldsymbol{L} \right)^{-1} \boldsymbol{G}^T \right) \boldsymbol{b} \right\rangle. \end{aligned} \tag{3.78}$$

The Woodbury matrix identity [Woodbury, 1950] states

$$(\boldsymbol{A} + \boldsymbol{U}\boldsymbol{C}\boldsymbol{V})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{U} \left(\boldsymbol{V}\boldsymbol{A}^{-1}\boldsymbol{U} + \boldsymbol{C}^{-1} \right)^{-1} \boldsymbol{V}\boldsymbol{A}^{-1}, \tag{3.79}$$

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

$$\left(I - \boldsymbol{G} \left(\boldsymbol{G}^T\boldsymbol{G} - \frac{\delta}{\lambda}\boldsymbol{L} \right)^{-1} \boldsymbol{G}^T \right) = \left(I + \frac{\lambda}{\delta}\boldsymbol{G}\boldsymbol{L}^{-1}\boldsymbol{G}^T \right)^{-1}. \tag{3.80}$$

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,

$$\delta^{-1}b(\lambda,\delta) = \ln\left(\left|\det\left(\frac{\lambda}{\delta}\boldsymbol{G}^T\boldsymbol{G}+\boldsymbol{L}\right)\right|\right) \tag{3.81}$$

can also be computed offline.

3.4 Evaluating Convergence

There are two final issues to address before presenting the numerical results of the algorithms on synthetic and real data. They are related to how close is the convergence to the invariant density and how closely do realizations of the chain represent uncorrelated samples. The steps in the chain where π_0 is far from converging to is referred to as burn-in and the tool for accounting for correlation is the integrated auto correlation time (IACT). Both issues inform how long to run the MCMC algorithm in order to effectively analyze chain as a robust sample for the PSF posterior. The first being that we must detect when 'burn-in' has completed, and the second is for how long must the MCMC algorithm run in order to

3.4.1 Estimating Burn-in

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

3.4.2 Integrated Autocorrelation

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

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.

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.

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.

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.

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.

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 solution of partial differential equations: an introduction*. Cambridge university press, 2005.

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.

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