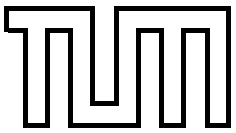
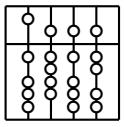


Technische Universität München
Fakultät für Informatik



Diplomarbeit in Informatik

# Image Deconvolution for Microscopy

Thomas Kasper

| | |
|---|---|
| Aufgabensteller: | Prof. Dr. Nassir Navab |
| Betreuer (TUM): | Dipl.-Tech. Math. Andreas Keil, Dipl.-Inf. Andreas Hofhauser |
| Betreuer (FIIS): | Dipl.-Inf. Thorsten Zerfaß, Dipl.-Ing. Stephan Rupp |
| Abgabedatum: | 12. Oktober 2006 |

Ich versichere, daß ich diese Diplomarbeit selbständig verfaßt und nur die angegebenen Quellen und Hilfsmittel verwendet habe.

München, den 12. Oktober Autor

**Acknowledgement**

I thank the Fraunhofer Institute (IIS) in Erlangen, Germany, and in particular the department Image Processing and Medical Engineering (BMT), for their friendly support during the six months I worked on this thesis.

## Abstract

Image restoration deals with recovering the original scene from the raw data recorded by a flawed optical device subject to systematic and random degradations. Modelled as a (convolution-) integral equation, it belongs to the class of notoriously ill-posed inverse problems characterized by their pathologic sensitivity to perturbations in the data. Casting the problem in an abstract framework of (robust) parameter estimation, we elaborate the theoretical background and discuss different regularization techniques. A selection of algorithms is reviewed and adapted for microscopy. Finally, their performance is evaluated on sets of both synthetically generated and real-world data.



## Zusammenfassung

Ziel der im Englischen als 'image restoration' bezeichneten Disziplin ist die möglichst wirlichkeitsgetreue Rekonstruktion von Bildern aus dem Datensatz eines optischen Geräts, das verschiedensten Beeinträchtigungen systematischer und zufälliger Natur unterworfen ist. Modelliert als (Faltungs-) Integralgleichung gehört sie in mathematischer Hinsicht zur Klasse der schlecht-gestellten inversen Probleme, die sich durch eine Überempfindlichkeit gegenüber Störungen im Datensatz auszeichnen. Die Arbeit beleuchtet den theoretischen Hintergrund aus der Perspektive sogenannter 'robuster' Parameter-Schätzung und behandelt verschiedene Regularisierungstechniken. Eine Auswahl von Algorithmen wird ausführlich besprochen, um sie dann speziell für den Anwendungsbereich der Mikroskopie zu adaptieren. Ihre Leistungsfähigeit schließlich wird anhand von künstlich generierten wie echten Datensätzen evaluiert.

# Contents

# 1 Introduction

Empirical data are rarely exact. This is true, in particular, for imagery. Optical devices are not perfect; the conditions under which they are put to use may be even less so. This observation, rather trivial in itself, has motivated a field of research that undertakes to mend the various effects of degradation — both systematic and random —, and to infer an estimate as faithful as possible of the original scene.

Mathematically, image deconvolution is a prime example of linear inverse problems. While the meaning of linearity is beyond dispute and requires no special elucidation, the very opposite is true for the second adjective. What determines the orientation of a problem and occasionally earns it the label 'inverse' is not clear in the first place. With notions that pertain to the realm of physical phenomena, we are told that inverse problems try to infer the unknown 'cause' giving rise to the observed 'effect'. While it is not difficult to match this description with the particular case of image deconvolution, it falls short of providing a rigorous definition. Clearly 'cause' and 'effect' are sound concepts within a framework of physical phenomena, but tend to lose their meaning when transferred into abstract mathematical language. In spite of its intuitivity, then, the classification may not be argued from a strictly formalistic point of view.

Whence their status as a separate class of their own? — one might wonder. A superficial assessment, indeed, will spot little difference to ordinary problems of parameter estimation. And yet, recurring to otherwise well-tried methods like Maximum-Likelihood estimation is ill-advised and bound to fail. Without anticipating an in-depth analysis to be conducted in due course, it seems as if inverse problems were characterized by their pathologic sensitivity to perturbations in the data — a property for which deterministic math has coined the notion of ill-posedness with its generalization as condition number. Owing to the genuinely probabilistic nature of the problem, however, we shall prefer a description in terms more specifically pertinent to the field of estimation theory and use statistical concepts like variance, bias, and mean-square-error performance instead. Although parallel explanations, very often, coexist in either domain, this framework strikes us as both more powerful and elegant.

The thesis is organized as follows. The first part of the second chapter is dedicated to the forward problem. The challenge, here, is to develop an adequate model of the image formation process, sophisticated enough to capture all relevant aspects without forfeiting practicability by an overly complex design. Simplifying assumptions — the better ones holding up to scrutiny, others more precarious and with less backing from empirical observation — are necessary to keep the problem commensurable in terms of computational cost. We review the standard models prevailing in the literature and discuss alternatives where we see fit.

The second part of the chapter takes a closer look at the inverse problem and the difficulties that inevitably arise from an ingenuous or naive approach. In order to motivate the subsequent inquiry into possible alternatives, it will be demonstrated both analytically and experimentally

why direct Maximum-Likelihood estimates, unless regularized by additional constraints, fail to produce meaningful results.

Chapter three launches into a general discussion of regularization techniques — typically one layer of abstraction underneath the concrete algorithm — to be used in what is commonly referred to as 'robust' parameter estimation. Wherever possible without forcing the matter, we have opted for a probabilistic perspective. Following a distinction that is rather philosophical than technical we have divided the chapter into sections on traditional and more specifically Bayesian approaches.

A selection of derived algorithms is presented and discussed in chapter four. The attempted survey is a by no means complete, yet deliberately eclectic in the variety of approaches considered. The algorithms, all of certain renown in their category, have been chosen so as to give a reasonably broad overview, covering both direct and iterative approaches. The comparably recent branch known as 'blind' deconvolution is represented by an EM-based method.

Chapter five is concerned with adapting and customizing the acquired instrumentarium for a given set of optical hardware at the Fraunhofer IIS Research Center, comprised of a light-transmitting confocal microscope with attached CCD camera for digital recording. As most of the best-performing algorithms to date do not fall into the blind category, their effective use entails the supply of reliable a-priori information about the optical system. With the relevant hardware at our disposition — an untypical situation, perhaps, and a privilege to take advantage of — both impulse response and read-out noise will be subject to a careful examination. In this context we derive a novel covariance estimator to be used for wide-sense-stationary random processes and experimentally verify its predicted accuracy.

An evaluation of the results finally can be found in chapter six. The implemented algorithms have been tested both on synthetic data — allowing a precise quantification of their performance in terms of square error — and 'authentic' images with various degrees of out-of-focus blur. In the former category, different levels of noise have been simulated, ranging from ideal to worst-case conditions.

# 2 Problem Statement

In this chapter we derive a mathematical model of the image formation process and lay the groundwork for the subsequent formulation of the inverse problem. It is clear that restoration can only succeed to the point where the model is a truthful representation of the underlying physical process. On the other extreme, an over-sophisticated model is pointless if it cannot be computed with reasonable effort. We will endeavor to hold a judicious balance here, striving for as faithful a model as possible, while being considerate of the available resources and their limits.

## 2.1 Forward Problem

As a first step toward modelling the degradation it makes sense to distinguish between deterministic and random components. Henceforth we will refer to the former as 'blur' and to the latter as 'noise'. Identifying the pristine image and its altered, degraded version with the variables $f$ and $g$ respectively, we have

$$g = \underbrace{A(f)}_{\text{blur}} + \underbrace{\nu}_{\text{noise}} \tag{2.1}$$

Although in practice no such representation will ever be available, the original scene may be assumed to be a real-valued function $f : \mathbb{R}^2 \longrightarrow [0,1]$ with compact support $\Omega_0 \subset \mathbb{R}^2$. Without loss of generality we will restrict ourselves to one-channel or greyscale images with intensities normalized to the unit interval, where it is implicitly understood that color images can be dealt with by processing each channel separately.

### 2.1.1 Blur-Model

In this section we review the standard model prevailing both in practical applications and pertinent literature. Occasional deviations shall be discussed where we see fit.

**Assumption 1: Linearity** The vast majority falls into this category. Except for very special applications, the few known exceptions have reported little, if any, improvement over linear models. Without making a substantial difference, however, non-linear models can hardly compete with their much less costlier counterparts of the linear category.

Let $\delta_{x,y}(\xi,\eta) := \delta(\xi - x, \eta - y)$ for $(x,y) \in \mathbb{R}^2$ denote shifted instances of the two-dimensional Dirac Impulse. Exploiting linearity, the blurred image $\tilde{f} := A(f)$ can be expressed as a superposition integral of the form

$$\tilde{f}(x,y) = \iint h(x,y,\xi,\eta) \cdot f(\xi,\eta)\, d\xi\, d\eta \tag{2.2}$$

where $h(x, y, \xi, \eta) = \langle A(\delta_{\xi,\eta}), \delta_{x,y} \rangle$ and $\langle \cdot, \cdot \rangle$ the canonical inner product in $\mathcal{L}^2$. In practice, all we have is an $m$ by $n$ grid of equally spaced samples, where $m$ and $n$ denote the number of rows and columns respectively. Replacing the integrals in (2.2) by sums we obtain the discrete approximation

$$\tilde{f}(x, y) = \sum_{\xi=1}^{m} \sum_{\eta=1}^{n} h(x, y, \xi, \eta) \cdot f(\xi, \eta) \tag{2.3}$$

for $1 \leq x \leq m$, $1 \leq y \leq n$. The cumbersome book-keeping of multiple indices can be eluded by using compact matrix-vector notation which is less prone to confusion. To this effect let $f_k = \big(f(1, k), ..., f(m, k)\big)^T$ denote the vector containing the $k$'th column of $f$. By replacing the outer sum in (2.3) we get

$$\tilde{f}_y = \sum_{\eta=1}^{n} A_{y,\eta} \cdot f_\eta \tag{2.4}$$

where

$$A_{y,\eta} = \{h(x, y, \xi, \eta)\}_{x,\xi=1...m} \in \mathbb{R}^{m \times m} \tag{2.5}$$

is the $m$ by $m$ square matrix formed by the kernel coefficients with fixed horizontal indices. Likewise, by vertically stacking the $n$ columns of the image, $f$ can be economically represented by one single vector $f = \big(f_1, ..., f_n\big)^T \in \mathbb{R}^{mn}$. Proceeding in the same manner with the remaining sum we find that

$$\tilde{f} = Af \tag{2.6}$$

with

$$A = \{A_{y,\eta}\}_{y,\eta=1...n} \in \mathbb{R}^{mn \times mn} \tag{2.7}$$

the block matrix composed of the $n \times n$ chunks previously defined in (2.5) with $y, \eta$ running independently from $1...n$.

It will be noted that without further simplification the dimension of this matrix would be a serious issue even for moderately sized images. With just under 69 billion components taking up 256 GB of physical memory for a square image of 512 pixels length — provided we settle for single precision! — it is not likely to fit into a customary desktop PC.

**Assumption 2: Shift-Invariance** A decisive simplification can be achieved by assuming that the blur is spatially invariant, meaning that the projection of an impulse at location $(\xi, \eta)$ onto $\delta_{x,y}$ is a function of (2-dimensional) distance

$$h(x, y, \xi, \eta) = h(x - \xi, y - \eta) \tag{2.8}$$

which makes (2.2) a convolution integral and $A$ a block Toeplitz matrix with Toeplitz blocks (BTTB). Indeed, by looking at (2.5) and (2.7) we find that both at the level of macro- and microstructure elements are constant along the diagonals

$$A_{i,j} = \{h(x - \xi, i - j)\}_{x,\xi=1...m} = A_{i-j} = A_k \tag{2.9}$$

with $k := i - j$ and likewise

$$A_l(i,j) = h(i-j,l) =: a_l^{(k)} \tag{2.10}$$

Due to its particular structure, $A$ is completely determined by $(2n-1)\cdot(2m-1) = 4mn - 2(n+m)+1$ parameters, which make for a storage cost that is linear in the number of pixels, as opposed to the quadratic behaviour of (2.7).

**Assumption 3: Periodic Boundaries** The assumption of shift-invariance falls short of providing an answer as to how boundaries should be dealt with. This question arises from the following observation. The aforementioned free parameters can be thought of as weights in a $2m-1$ by $2n-1$ filter mask, just big enough to have it cover the whole image for any possible alignment. It follows that roughly a quarter of the filter coefficients contribute for any given pixel while the remaining ones map to intensities outside the area covered by the recording system

$$\begin{bmatrix} h_{-(m-1),-(n-1)} & h_{-(m-1),-(n-1)} & \cdots & h_{-(m-1),n-2} & h_{-(m-1),n-1} \\ h_{-(m-2),-(n-1)} & \ddots & & & h_{-(m-1),n-1} \\ \vdots & & & & \vdots \\ h_{m-2,-(n-1)} & & & \ddots & h_{m-2,n-1} \\ h_{m-1,-(n-1)} & h_{m-1,-(n-2)} & \cdots & h_{m-1,n-2} & h_{m-1,n-1} \end{bmatrix}$$

These masks, usually normalized with weights summing up to one, tend to have much smaller support than the actual image. Typically filter coefficients have a peak at the anchor and decay more or less rapidly toward the perimeter. Though the output is always a linear combination of $mn$ values, the number of 'active' non-zero weights may decrease significantly as we approach the borders of the image, causing it to leak some of its energy there.

To mend this effect, the unavailable information usually is predicted according to one of the following patterns

- periodic extension/wrap-around

$$f(-x,y) \approx f(m-x,y) \qquad h_{-i,j} = h_{m-i,j} \tag{2.11a}$$
$$f(x,-y) \approx f(x,n-y) \qquad h_{i,-j} = h_{i,n-j} \tag{2.11b}$$

- reflexive extension/axial symmetry

$$f(-x,y) \approx f(x,y) \qquad h_{-i,j} = h_{i,j} \tag{2.12a}$$
$$f(x,-y) \approx f(x,y) \qquad h_{i,-j} = h_{i,j} \tag{2.12b}$$

Obviously, each of the above heuristics — stated in this generality — is bound to find itself at fault with reality, for most images of interest are neither periodic nor symmetric. Usually they are just good enough to prevent the image from leaking too much of its energy at the borders without introducing strong artefacts.

Opinions diverge as to how important boundary conditions are for the restoration. While [26] maintains that the reflexive variant, on average, yields better results, superficial tests

have not been able to confirm this and rather suggest that, especially for images of big or moderate size, boundary condition have not a great impact on the result.

In the following we shall opt for periodic variant which has the advantage of being easy to compute. Provided the filter mask does not exceed the size of the image, a linear convolution may still be simulated by padding the image with an appropriately sized strip of zeros in both dimensions. With the inclusion of boundary condition (2.11a) we may specialize the BTTB-structure of $A$ even further, by noting that the matrix is now block circulant

$$A = \begin{pmatrix} A_0 & A_{n-1} & & \dots & A_1 \\ A_1 & A_0 & A_{n-1} & & \vdots \\ \vdots & A_1 & A_0 & \ddots & \vdots \\ \vdots & & \ddots & \ddots & A_{n-1} \\ A_{n-1} & & & A_1 & A_0 \end{pmatrix} \in \mathbb{R}^{mn \times mn} \tag{2.13}$$

with circulant blocks (BCCB)

$$A_k = \begin{pmatrix} a_k^{(0)} & a_k^{(m-1)} & & \dots & a_k^{(1)} \\ a_k^{(1)} & a_k^{(0)} & a_k^{(m-1)} & & \vdots \\ \vdots & a_k^{(1)} & a_k^{(0)} & \ddots & \vdots \\ \vdots & & \ddots & \ddots & a_k^{(m-1)} \\ a_k^{(m-1)} & & & a_k^{(1)} & a_k^{(0)} \end{pmatrix} \in \mathbb{R}^{m \times m} \tag{2.14}$$

This matrix, or any of its columns by which it is completely defined, sometimes is referred to as point-spread-function (PSF). The actual shape of the PSF is determined by the mechanical properties of the optical device or induced by other physical phenomena, such as

- Diffraction
- Atmospheric turbulence
- Out-of-focus blur
- Camera motion
- ...

As a detailed discussion of these phenomena is beyond the scope of this thesis, we refer the interested reader to any of the numerous textbooks on optical physics.

### 2.1.2 Noise-Model

Unlike the pristine image, read-out noise, from the very outset, is a discrete phenomenon. It can be modelled as a random process or — to emphasize that it unfolds in space rather than time — as a random field.

**Assumption 1: Gaussianity** The statistical properties of the noise are device dependent. For CCD cameras, which have been used for testing and evaluation, light intensities are known to have a Poissonian distribution

$$g_i = (Af)_i + \nu_i \sim P_o((Af)_i) \tag{2.15}$$

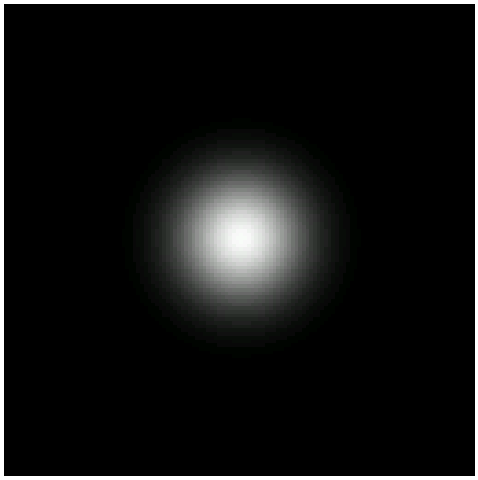


Figure 2.1: Example of a Gaussian PSF

A discrete distribution, however, proves unwieldy for further analysis. Since $\mathrm{E}\{X\} = \mathrm{Var}[X]$ for a Poissonian random variable $X$, the model (2.15) can be approximated up to the second (central) moment by letting

$$\nu_i = \sqrt{(Af)_i} \cdot X \qquad\qquad X \sim \mathcal{N}(0,1) \tag{2.16}$$

where the square-root is well-defined due to the non-negativity of light-intensities. The effect of measurement errors can thus be modelled as additive Gaussian noise with (multivariate) distribution

$$\nu \sim \mathcal{N}(0, \Phi_\nu) \tag{2.17}$$

where the variance $\Phi_\nu$ is specified only up to the entries on the main diagonal by (2.16)

$$\Phi_{ij} = \begin{cases} (Af)_i & i = j \\ \mathrm{Cov}(\nu_i, \nu_j) & i \neq j \end{cases} \tag{2.18}$$

Note that $\nu$ is not stationary, which would require that $\mathrm{Var}[\nu_1] = \mathrm{Var}[\nu_2] = ... = \mathrm{Var}[\nu_{mn}]$. For any non-trivial combination of $(A, f)$, however, this is unlikely to hold.

**Assumption 2: Wide-Sense Stationarity (WSS)** Although Poissonian models like (2.15) and their approximation through non-stationary Gaussian processes (2.16) are current in astronomical imaging — Richardson-Lucy being the most noteworthy example of this class — they have not been able to prevail on a larger scale. The limited success of these models is due, in part, to their comparably high complexity. Expensive processing — of minor relevance in astronomical imagery, where it still relates favourably to the cost of collecting the data — is often unproportionate or simply not affordable.

For the sake of simplicity, albeit against better knowledge, it has become *de-facto* standard to model the noise as a wide-sense stationary (WSS) zero-mean Gaussian process, meaning that

$$\forall x, y : \mathrm{E}\{\nu(x,y)\} = 0 \qquad\qquad \mathrm{E}\{\nu(x,y) \cdot \nu(\hat{x},\hat{y})\} = R_{\nu\nu}(x - \hat{x}, y - \hat{y}) \tag{2.19}$$

This property induces a variance matrix that, again, is block Toeplitz

$$\Phi_\nu = \begin{pmatrix} C_1 & C_2 & & \dots & C_n \\ C_2^T & C_1 & C_2 & & \vdots \\ \vdots & C_2^T & C_1 & \ddots & \vdots \\ \vdots & & \ddots & \ddots & C_2 \\ C_n^T & & & C_2^T & C_1 \end{pmatrix} \in \mathbb{R}^{mn \times mn} \tag{2.20}$$

with Toeplitz blocks

$$C_i = \begin{pmatrix} c_{i,m} & c_{i,m+1} & & \dots & c_{i,2m-1} \\ c_{i,m-1} & c_{i,m} & c_{i,m+1} & & \vdots \\ \vdots & c_{i,m-1} & c_{i,m} & \ddots & \vdots \\ \vdots & & \ddots & \ddots & c_{i,m+1} \\ c_{i,1} & & & c_{i,m-1} & c_{i,m} \end{pmatrix} \in \mathbb{R}^{m \times m} \tag{2.21}$$

Note that due to the symmetry of the variance matrix, $C_1 = C_1^T$ and hence $c_{1,2m-i} = c_{1,i}$ for $1 \leq i \leq m$.

**Assumption 3: Uncorrelatedness** Even more restrictively, noise is sometimes assumed to be uncorrelated

$$R_{\nu\nu}(x, y, \hat{x}, \hat{y}) = \sigma_\nu^2 \, \delta_{x\hat{x}} \, \delta_{y\hat{y}} \tag{2.22}$$

with $\sigma_\nu$ the standard deviation of any one component which makes $\Phi_\nu = \sigma_\nu^2 \cdot I_n$ a scalar multiple of the identity matrix. This case is often referred to as additive 'white' Gaussian noise (AWGN), in analogy to the flat power spectrum of daylight, where all frequencies are typically represented to an equal extent.

According to [26] algorithms based on Possionian noise distribution do not perform significantly better in a large number of scenarios that typically arise in practice. Therefore, unless specific knowledge suggests beforehand that these models will provide superior results, we shall opt, by default, for the Gaussian variant $\nu \sim \mathcal{N}(0, \Phi_\nu)$ with $\Phi_\nu$ BTTB and specialize in the sense of (2.22) as we see fit.

## 2.2 Convolution Theorem

Due to its huge size, the matrix representation of $A$ does not lend itself to actual computations and is hardly ever constructed in practice. Fortunately, we find that linear systems involving BCCB matrices are substantially less complex than matrix-vector algebra in general and can be efficiently solved using discrete Fourier Transform. Such is, in quintessence, the finding usually referred to as convolution theorem. Loosely formulated it states the equivalence of the following operations

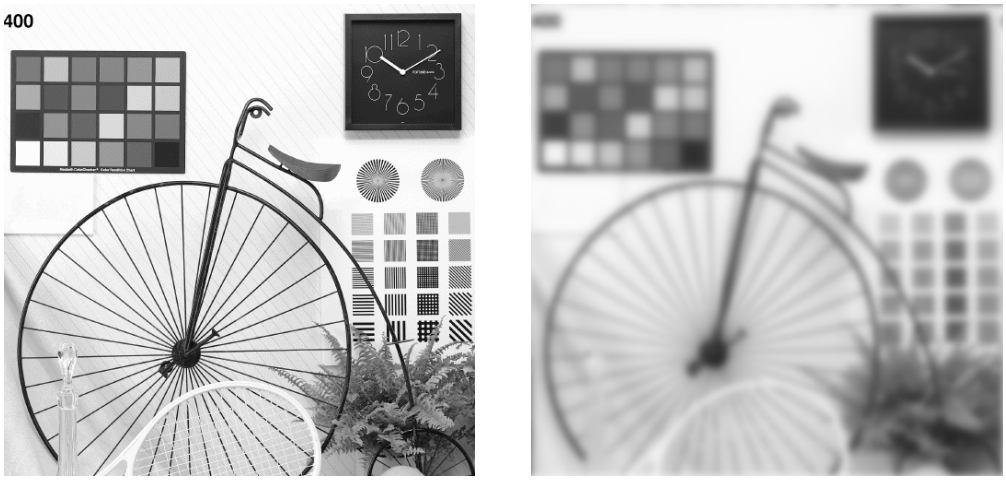



Figure 2.2: Convolution with a Gaussian blur kernel

Discrete 1D [2D]-Convolution
⇕
Multiplication with [block-]circulant matrix [with circulant blocks]
⇕
Componentwise multiplication in the 1D [2D] Fourier-domain

where the parts in square brackets refer to the 2-dimensional case. We review the theorem here in some detail both for its general importance in signal processing and the extensive use we are going to make of it throughout the following sections.

**Convolution Theorem.** *Let* $C = \sum_{k=1}^{m} c_i \cdot R^{k-1} \in \mathbb{R}^{m\times m}$ *be a circulant matrix, with* $c = (c_1, ..., c_m)^T = Ce_1$ *its first column and* $R_m = (e_2, ..., e_m, e_1) \in \mathbb{R}^{m\times m}$ *the downshift operator. Let further* $F_m = \{\frac{1}{\sqrt{m}} \cdot \omega_m^{(i-1)(j-1)}\}_{i,j=1,..m}$ *the one-dimensional DFT matrix, where* $\omega_m = \exp(-2\pi\imath/m)$ *and* $\imath = \sqrt{-1}$ *the imaginary unit. Then*

$$F_m C F_m^H = \operatorname{diag}(F_m C e_1) \tag{2.23}$$

*Now let* $F_{mn} = F_n \otimes F_m$ *be the two-dimensional DFT matrix,* $\otimes$ *denoting Kronecker-product. Then for a block-circulant matrix* $A = \sum_{k=1}^{n} R^{k-1} \otimes C_k \in \mathbb{R}^{mn\times mn}$ *with circulant blocks* $C_k \in \mathbb{R}^{m\times m}$ *it holds that*

$$F_{mn} A F_{mn}^H = \operatorname{diag}(F_{mn} A e_1) \tag{2.24}$$

**Proof:** As a first step in proving the theorem we show that the $k$'th column of $F_m^H$ is an

eigenvector of $R$ with corresponding eigenvalue $\omega_m^{k-1}$.

$$R_m(F_m^H e_k) = R_m \cdot \frac{1}{\sqrt{m}} \begin{pmatrix} 1 \\ \omega_m^{*(k-1)} \\ \vdots \\ \omega_m^{*(k-1)(m-1)} \end{pmatrix} = \frac{1}{\sqrt{m}} \begin{pmatrix} \omega_m^{*(k-1)(m-1)} \\ 1 \\ \vdots \\ \omega_m^{*(k-1)(m-2)} \end{pmatrix} = \omega_m^{(k-1)} \cdot F_m^H e_k \tag{2.25}$$

where the asterisk $(*)$ denotes complex conjugation. It follows immediately from (2.25) that

$$F_m R_m F_m^H = \operatorname{diag}\left(1, \omega_m, ..., \omega_m^{m-1}\right) \tag{2.26}$$

Exploiting the representation of $C$ as a polynomial in the downshift operator $R_m$, we find that

$$\begin{aligned} F_m C F_m^H &= F_m \left( \sum_{k=1}^{m} c_k \cdot R_m^{k-1} \right) F_m^H \\ &= \sum_{k=1}^{m} c_k \cdot F_m R_m^{k-1} F_m^H \\ &= \sum_{k=1}^{m} c_k \left( F_m R_m F_m^H \right)^{k-1} \\ &= \sum_{k=1}^{m} c_k \cdot \operatorname{diag}\left(1, \omega_m, ..., \omega_m^{m-1}\right)^{k-1} \\ &= \operatorname{diag}\left( \sum_k c_k, \sum_k c_k \omega_m^{k-1}, ..., \sum_k c_k \omega_m^{(m-1)(k-1)} \right) \\ &= \operatorname{diag}\left(F_m C e_1\right) \end{aligned} \tag{2.27}$$

It remains to verify the theorem for the 2-dimensional case. To resolve ambiguities, superscripts in parenthesis indicate the dimension of the canonical base vectors. Exploiting the identities $(A \otimes B)^H = A^H \otimes B^H$ and $(A \otimes B) \cdot (C \otimes D) = (AC \otimes BD)$ we get

$$\begin{aligned} F_{mn} A F_{mn}^H &= \sum_{k=1}^{n} F \left( R_n^{k-1} \otimes C_k \right) F^H \\ &= \sum_{k=1}^{n} \left(F_n \otimes F_m\right) \left( R_n^{k-1} \otimes C_k \right) \left( F_n^H \otimes F_m^H \right) \\ &= \sum_{k=1}^{n} \left( F_n R_n^{k-1} \otimes F_m C_k \right) \left( F_n^H \otimes F_m^H \right) \\ &= \sum_{k=1}^{n} \left( F_n R_n^{k-1} F_n^H \right) \otimes \left( F_m C_k F_m^H \right) \\ &= \sum_{k=1}^{n} \operatorname{diag}\left( F_n R_n^{k-1} e_1^{(n)} \right) \otimes \operatorname{diag}\left( F_m C_k e_1^{(m)} \right) \end{aligned} \tag{2.28}$$

which, by using $\mathrm{diag}\,(A) \otimes \mathrm{diag}\,(B) = \mathrm{diag}\,(A \otimes B)$, may be reduced to the identity

$$\begin{aligned} F_{mn} A F_{mn}^H &= \sum_{k=1}^{n} \mathrm{diag}\left(F_n R_n^{k-1} e_1^{(n)} \otimes F_m C_k e_1^{(m)}\right) \\ &= \sum_{k=1}^{n} \mathrm{diag}\left((F_n \otimes F_m) \cdot (R_n^{k-1} e_1^{(n)} \otimes C_k e_1^{(m)})\right) \\ &= \mathrm{diag}\left((F_n \otimes F_m) \cdot (\sum_{k=1}^{n} R_n^{k-1} C_k) \cdot (e_1^{(m)} \otimes e_1^{(n)})\right) \\ &= \mathrm{diag}\left(F_{mn} A e_1^{(mn)}\right) \end{aligned} \tag{2.29}$$

as claimed. ■

The importance of the above theorem is not duly appreciated unless we consider implementational issues. From (2.24) it seems as if the $O(n^3), n = dim(C)$ operations required by a convolution in the spatial domain were merely replaced by the DFT, without effectively reducing the overall-complexity. Here the matrix-vector notation used for convenience is obfuscating in the sense that it makes (2.24) look more expensive than it actually is. In practice, we use recursive Fast Fourier Transform (FFT) which is $O(n \log n)$ and readily available in an excellent and well-documented open source implementation as FFTW3. For details on this issue, including a survey of different algorithmic approaches, see [20].

## 2.3 Inverse Problem

We consider the inverse problem within a probabilistic framework of parameter estimation. Given the impulse response or blurring kernel and the observation $g$ the goal is to find as faithful an approximation of the pristine image $f$ as possible.

In order to motivate the subsequent inquiry into so-called robust parameter estimation we start by illustrating the problem arising from an ingenious, or at any rate, less sophisticated approach. In particular it will be shown why otherwise well-tried methods, so long as they do not take into account the particular nature of the problem, are bound to fail.

**Maximum Likelihood Estimator and Pseudo Inverse** Perhaps the most intuitive way to go about constructing an estimator is the Maximum-Likelihood principle. Suppose the noise is Gaussian with known distribution

$$\nu \sim \mathcal{N}(\mu_\nu, \Phi) \tag{2.30}$$

All probabilities involved being strictly positive, the conditional log-likelihood is well-defined and given by

$$\log p(g|f) = c - \frac{(g - Af - \mu_\nu)^T \Phi^{-1} (g - Af - \mu_\nu)}{2} \tag{2.31}$$

where $c = -\log|2\pi\,\Phi|/2$ is a constant independent of $f$. Consider first the special case of white zero-mean noise with standard deviation $\sigma_\nu$ for any one component

$$\mu_\nu = 0 \qquad\qquad \Phi = \sigma_\nu^2 I \tag{2.32}$$

Then the second term on the right-hand-side of (2.31) represents simply the 2-norm of the residual, scaled by a strictly positive factor $\sigma^{-2} > 0$. Finding the maximizer of (2.31) with respect to $f$ thus is equivalent with the least squares problem (LS)

$$\arg\max \log(g|f) = \arg\min \|g - Af\|_2 \tag{2.33}$$

whose well-known solution

$$\hat{f}_{ML} = A^{\dagger} g \tag{2.34}$$

defines the unbiased ML-estimator for $f$. Depending on whether $A \in \mathbb{R}^{m \times n}$ is over- or underdetermined, an explicit representation of the pseudo-inverse is given by

$$A^{\dagger} = \begin{cases} (A^T A)^{-1} A^T & m \geq n \\ A^T (A A^T)^{-1} & m \leq n \end{cases} \tag{2.35}$$

Using Choleski-Factorization for positive definite matrices, the general case may be reduced to (2.32) by the following change of variables

$$\hat{g} = \Phi^{-1/2}(g - \mu_\nu) \qquad\qquad \hat{A} = \Phi^{-1/2} A \tag{2.36}$$

Premultiplying the variables with $\Phi^{-1/2}$ acts as a prewhitening filter and effectively decorrelates the noise. Transforming (2.30) in this way eventually yields the Generalized Least Squares (GLS) problem

$$\hat{f}_{ML} = \hat{A}^{\dagger} \hat{g} = (A^T \Phi^{-1} A)^{-1} A^T \Phi^{-1} g \tag{2.37}$$

for $m \geq n$.

Without loss of generality we shall henceforth consider only the case of uncorrelated zero-mean Gaussian noise

$$\nu \sim \mathcal{N}(0, \sigma_\nu^2 I) \tag{2.38}$$

where it is implicitly understood that the general case (2.30) can be reduced to this form by changing the variables as in (2.36). For the time being, we conclude that for normally distributed noise Maximum Likelihood (ML) estimation is conceptually equivalent to (Generalized) Least Squares.

Maximum-Likelihood (ML) estimation
⇕
Generalized Least Squares (GLS)

**Singular Value Decomposition** Sometimes it is convenient to express the Pseudo-Inverse in terms of Singular Value Decomposition (SVD), without having to discriminate between over- and underdetermined case. Let

$$A = USV^H \tag{2.39}$$

with $U \in \mathbb{C}^{m\times m}$ and $V \in \mathbb{C}^{n\times n}$ unitary matrices, $S = \operatorname{diag}(\sigma_1, ..., \sigma_r, 0, ...) \in \mathbb{R}^{m\times n}$ the diagonal matrix of singular values, conventionally arranged in decreasing order such that

$$\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_r > 0 \tag{2.40}$$

and $r = \operatorname{rank}(A) \leq \min(m, n)$. Define $S^{-1} := \operatorname{diag}\left(\sigma_1^{-1}, ..., \sigma_r^{-1}, 0, ...\right) \in \mathbb{R}^{n\times m}$ the matrix obtained by transposal of $S$ and inversion of non-zero entries on the main diagonal. Then

$$A^\dagger g = VS^{-1}U^H g \tag{2.41}$$

is sometimes referred to as *principal solution*. Writing out (2.41) as a sum we find that, equivalently

$$A^\dagger g = \sum_{i=1}^{r} v_i \frac{\langle u_i, g\rangle}{\sigma_i} \tag{2.42}$$

where the subscripted lowercase letters $u_i$ and $v_i$ refer to the left and right singular vectors which comprise the columns of $U$ and $V$ respectively.

**Error Analysis** It is common to assess the goodness of an estimator $\hat{f}$ in terms of its *Mean (Integrated) Square Error* (MISE), defined as

$$\operatorname{MISE}\left[\hat{f}\right] = \mathrm{E}\left\{ \left\| f - \hat{f} \right\|_2^2 \;\middle|\; f \right\} \tag{2.43}$$

Of all unbiased estimators, the Maximum Likelihood-estimate (2.34) and its generalization for arbitrary variance matrices (2.37) may be shown to be optimal in the sense of (2.43)

$$\hat{f}_{ML} = \underset{\{\hat{f}|\mathrm{E}\{\hat{f}\}=f\}}{\arg\min} \operatorname{MISE}\left[\hat{f}\right] \tag{2.44}$$

which makes it also the Best Linear Unbiased Estimator (BLUE).

Although unbiased, the Maximum Likelihood estimator turns out to be a poor choice for its variance and mean square error performance. What the label 'Best Linear Unbiased' is actually worth can be seen in figure 2.3. It represents the estimate obtained by straight inversion of the same square PSF-matrix previously used to blur the image without adding a whatsoever small amount of noise. Due to its huge size, the system of linear equations has been solved using Fast-Fourier-Transform (FFT) by application of the convolution theorem (2.24). Those who suspect this particular method to be instable and would like to hold it responsible for the disastrous result 2.3, will recall that the DFT is numerically as well-behaved as can be (with condition number $\kappa = 1$ for orthogonal transforms). The only noise then, if such we want to call it, is introduced by inevitable rounding errors in the order of machine precision $\epsilon \approx 10^{-16}$ for 64-bit IEEE floating point numbers. The real issue, let us be clear about it, is the system itself, rather than a particular way of solving it. Looking at (2.42) should give a good idea of what can go wrong. For although $\sigma_i \neq 0$, its absolute value can be arbitrarily small. So long as the data are not corrupted by noise, dividing by this quantity is innocuous, for any contribution of this subspace in the original image has been obliterated by the forward mapping, so that division and previous scaling cancel

$$\frac{\langle v_i, g\rangle}{\sigma_i} = \frac{\langle v_i, Af\rangle + \langle v_i, \nu\rangle}{\sigma_i} = \langle v_i, f\rangle + \frac{\langle v_i, \nu\rangle}{\sigma_i} \tag{2.45}$$

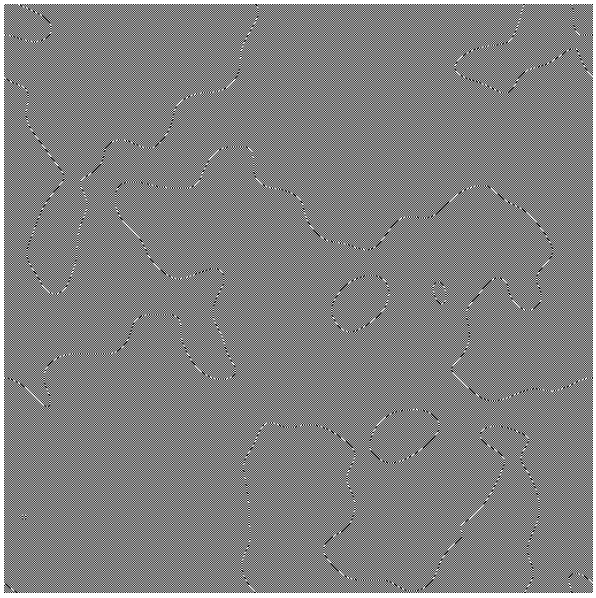


Figure 2.3: Maximum Likelihood Estimate and BLUE

In the presence of noise, however, we see that (2.45) is largely dominated by the second term for singular values with small magnitude $\sigma_i \approx 0$. However moderate the level of noise, it will end up hugely amplified in the corresponding subspaces of the restored inverse, introducing artefacts that usually distort it beyond recognition. The estimate, then, will not be anywhere close to the pristine image and void of any real meaning.

This pathological sensitivity to small perturbations in the data can be formalized by analyzing the error performance of the estimator according to (2.43). Using that for a zero-mean random variable $X \in \mathbb{R}^n$

$$\mathrm{E}\left\{\|X\|_2^2\right\} = \mathrm{E}\left\{\sum_{i=1}^{n} X_i^2\right\} = \sum_{i=1}^{n} \mathrm{Var}\left[X_i\right] = \mathrm{trace}\left(\mathrm{Var}\left[X\right]\right) \tag{2.46}$$

holds and letting $A = USV^H$ as in (2.39) we find that the mean (integrated) square error of the Maximum Likelihood estimator is given by

$$\begin{aligned} \mathrm{E}\left\{\left\|f - A^{\dagger}g\right\|_2^2 \mid f\right\} &= \mathrm{E}\left\{\left\|VS^{-1}U^H\nu\right\|_2^2\right\} \\ &= \sigma_\nu^2 \,\mathrm{trace}\left(S^{-2}\right) \end{aligned} \tag{2.47}$$

and hence

$$\mathrm{MISE}\left[\hat{f}_{ML}\right] = \sigma_\nu^2 \sum_{i=1}^{r} \frac{1}{\sigma_i^2} \approx \infty \tag{2.48}$$

for small singular values $\sigma_i \approx 0$.

Unfortunately, we have to reckon with singular values close to zero, not accidentally but as a necessary consequence of $A$ being a bounded linear operator. The spectrum $(\sigma_1, ..., \sigma_r)$ of $A$

which coincides with basically the Fourier Transform of the PSF, is also refered to as Optical Transfer Function (OTF) and characterizes its response to different frequencies. Though OTFs are device specific and vary accordingly, they typically exhibit a more or less rapid decay of magnitude toward high frequency components which is indicative of a poor or, at any rate, limited resolution capacity of the device.

Examples of this phenomenon are manifold. Consider, for instance, that the Fourier-Transform of a Gaussian

$$f(x) = e^{-x^2/(2\sigma^2)} \quad \overset{FT}{\longleftrightarrow} \quad F(\omega) = \sqrt{2\pi}\sigma e^{\omega^2\sigma^2/2} \tag{2.49}$$

is another Gaussian with variance antiproportional to that of the former. In other words, the larger the spread in the PSF, the faster will be the decay in the magnitudes of the spectrum toward high frequencies, leading to the critical situation described above.

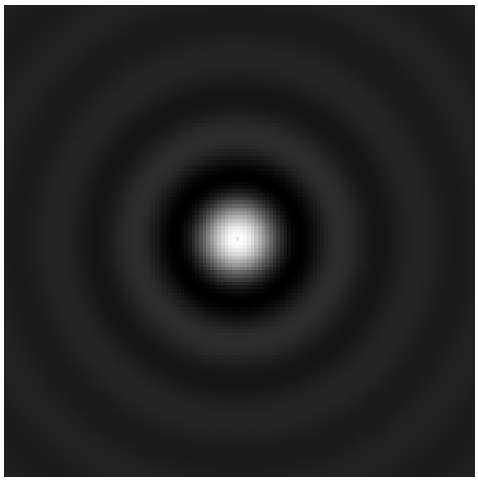


Figure 2.4: OTF of a diffraction limited microscope (Airy disk)

Having identified the extreme volatility of the unbiased estimator as its critical flaw, we shall now explore ways to make it more 'robust'.

# 3 Robust Parameter Estimation

Estimates obtained by straight inversion of the blur-operator have been shown to be useless because of their enormous sensitivity to noise. They will exhibit wild oscillations — both in the sense that high (spatial) frequency components dominate in the reconstructed image as well as in the sense of arbitrarily disparate solutions, should the experiment be repeated several times. As a straight inversion will almost certainly fail to yield a meaningful solution to the problem, other approaches have to be considered.

In order to obtain a more regular estimate, conjectures have to make up for what is missing in the data. In other words, we exploit what we know — or believe to know — about the pristine image independent of the data and even before any measurements are made. This is not as forlorn as it may seem at first sight. The image — by all reason — will exhibit features of some spatial extension; regions of interest will typically stretch out over many pixels. The contrary is very unlikely at best and may be dismissed on the grounds that otherwise the resolution would be inappropriate to capture relevant information anyway. Notwithstanding singularities (edges and regions of sharp contrast in the image), it seems reasonable to expect a minimum of correlation among the pixels, typically more pronounced for immediate neighbors and roughly decaying as a function of their distance. In other words, smoothness and stability are distinguishing properties of a plausible solution. 'Most' elements — whatever that means for an incountably infinite set — contained in $\{f : \mathbb{R}^2 \longrightarrow [0, 1]\}$ can be ruled out as violating those constraints.

In this chapter we consider different ways to make the estimator prefer solutions which are deemed more plausible than others. The structure is bipartite, following a distinction that is much rather philosophical than practical. As for concrete recipes, there may be little difference, indeed; being derived from a distinct theoretical framework in either case, however, we shall treat Bayesian and Non-Bayesian techniques separately. While the latter ones (spectral filters, for their majority) remain consistent with the setting of conventional parameter estimation — all they do, in quintessence, is trade variance for bias — Bayesian techniques require a recast of the whole problem. Their remarkable flexibility, extending beyond the subset of linear problems, stems from the fact that the parameter to be estimated is itself changed into a random variable obeying a distribution known as *prior probability*. Both approaches have a long history of successful application to inverse problems.

Literature and online material pertinent to this subject is prolific; most recent and comprehensive publications include [25] and [17].

## 3.1 Non-Bayesian Regularization

**Idea** Clearly, an ideal estimator would combine the best of all possible worlds and join a small variance to a zero bias. In reality, however, even the best unbiased estimator has been

shown to possess a variance exceeding all reasonable bounds. To comprehend how either of the those quantities contributes to the mean (integrated) square error, consider that

$$\begin{aligned} \mathrm{MISE}\left[\hat{f}\right] &= \mathrm{E}\left\{\left\|f - \mathrm{E}\left\{\hat{f}\right\} + \mathrm{E}\left\{\hat{f}\right\} - \hat{f}\right\|_2^2 \;\middle|\; f\right\} \\ &= \left\|\mathrm{Bias}\left[\hat{f}\right]\right\|_2^2 + \mathrm{trace}\left(\mathrm{Var}\left[\hat{f}\right]\right) \end{aligned} \tag{3.1}$$

According to (3.1) constructing a well-performing estimator is equivalent to minimizing the sum of the above terms. One may go about this task using the following heuristic. Starting with the Maximum Likelihood estimator for which

$$\left\|\mathrm{Bias}\left[\hat{f}_{ML}\right]\right\|_2^2 = 0 \qquad\qquad \mathrm{trace}\left(\mathrm{Var}\left[\hat{f}_{ML}\right]\right) \approx \infty \tag{3.2}$$

the idea is to move along an appropriately parameterized trade-off curve away from zero-bias toward smaller variance. In other words, we accept a systematic error to achieve greater stability in turn. Such is, summarized in one sentence, the deal underlying non-Bayesian regularization techniques.

### 3.1.1 Spectral Filters

One way to parameterize the trade-off curve is to apply a filter that operates on the spectrum of $A$. Let $A = USV^H$ the SVD as in (2.39). Letting $\alpha$ denote the amount of regularization we wish to inflict upon the solution, the resulting estimate is given by

$$\hat{f}_\alpha = VR_\alpha S^{-1}U^H g \tag{3.3}$$

with $R_\alpha$ the diagonal matrix of filter coefficients

$$R_\alpha = \begin{pmatrix} r_\alpha(\sigma_1) & & 0 \\ & \ddots & \\ 0 & & r_\alpha(\sigma_n) \end{pmatrix} \tag{3.4}$$

and $r_\alpha : [0, \infty] \longrightarrow [0, 1]$ appropriately. Basically, this may be any function satisfying

$$r_0(\sigma) = 1 \qquad\qquad \lim_{\alpha\to\infty} \frac{r_\alpha(\sigma)}{\sigma} = 0 \tag{3.5}$$

Otherwise its design is somewhat arbitrary. Two of the most commonly used filter functions are

$$r_\alpha(\sigma) = \begin{cases} \mathbf{1}_{(\sigma^2>\alpha)} & \text{Truncated SVD ('Spectral-Cut-Off')} \\ \dfrac{\sigma^2}{\sigma^2+\alpha} & \text{Ridge Regression (Tikhonov-Regularization)} \end{cases} \tag{3.6}$$

To our knowledge there is no real difference between Ridge Regression and Tikhonov Regularization other than, perhaps, their pertinence to different theoretical frameworks — probabilistic the former, deterministic the latter. As far as we can see it's merely two names for the same thing.

The effect of spectral filters is to dampen or completely obliterate contributions of subspaces associated with small singular values, while those characterized by a sound signal-to-noise-ratio (SNR) and well-determined by the data pass virtually unaltered. Unlike the plain thresholding of spectral cut-off — also known as Truncated Singular Value Decomposition (TSVD) — the Tikhonov filter function smoothly interpolates between the boundary constraints 0 and 1 (see the graph in figure 3.1).

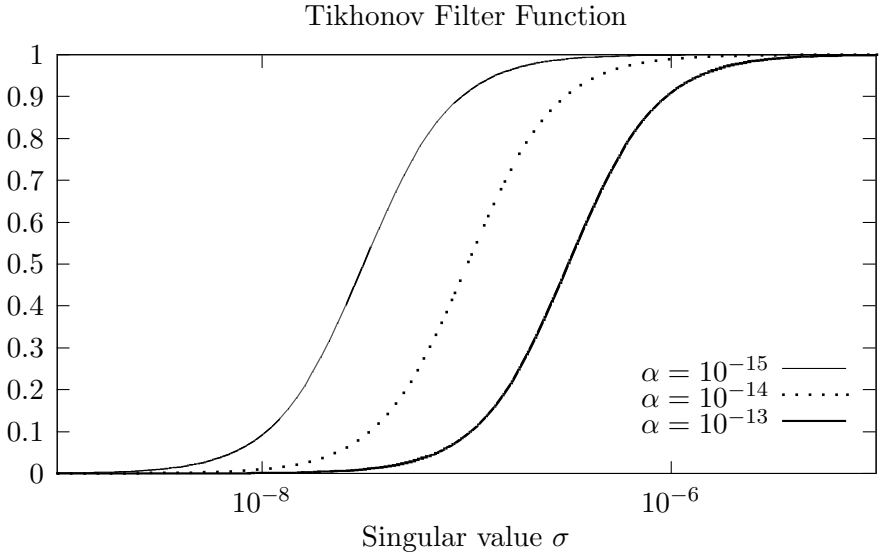



Figure 3.1: Tikhonov Filter Function for different values of $\alpha$

**Error Analysis** It will be noted that the resulting estimator $\hat{f}_\alpha$ has bias

$$\begin{aligned} \left\| \text{Bias}\left[\hat{f}_\alpha\right] \right\|_2^2 &= \left\| V(I - R_\alpha)V^H f \right\|_2^2 \\ &\leq \|f\|_2^2 \max_{1\leq i\leq r} (1 - r_\alpha(\sigma_i))^2 \end{aligned} \tag{3.7}$$

As for the other quantity on the right-hand-side of (3.1) it holds that

$$\begin{aligned} \text{trace}\left(\text{Var}\left[\hat{f}_\alpha\right]\right) &= \sigma_\nu^2 \,\text{trace}\left((RS^{-1})^2\right) \\ &\leq \sigma_\nu^2 \, n \max_{1\leq i\leq r} \left(\frac{r_\alpha(\sigma_i)}{\sigma_i}\right)^2 \end{aligned} \tag{3.8}$$

Using properties (3.5), we see that the regularized solution interpolates between the Maximum-Likelihood estimate (obtained as a special case of (3.3) by letting $\alpha = 0$) and the uniformly 'black' image of zero-intensities

$$\lim_{\alpha\to\infty} \left\| \text{Bias}\left[\hat{f}_\alpha\right] \right\|_2^2 = \|f\|_2^2 \qquad\qquad \lim_{\alpha\to\infty} \text{trace}\left(\text{Var}\left[\hat{f}_\alpha\right]\right) = 0 \tag{3.9}$$

To see that the regularized solution effectively makes for a better estimate, consider the example of Ridge Regression. By substituting into (2.43) and using that the 2-norm is invariant

under orthogonal transform, we find that

$$\begin{aligned} \mathrm{MISE}\left[\hat{f}_\alpha\right] &= \sum_{i=1}^{r}\left(\frac{\sigma_i^2}{\sigma_i^2+\alpha}-1\right)^2 + \sigma_\nu^2\sum_{i=1}^{r}\left(\frac{\sigma_i^2}{\sigma_i^2+\alpha}\frac{1}{\sigma_i}\right)^2 \\ &= \sum_{i=1}^{r}\frac{\sigma_\nu^2\sigma_i^2-\alpha^2\langle v_i, f\rangle}{(\sigma_i^2+\alpha)^2} \end{aligned} \tag{3.10}$$

Differentiating with respect to $\alpha$ yields

$$\frac{\partial}{\partial\alpha}\,\mathrm{MISE}\left[\hat{f}_\alpha\right] = -2\sum_{i=1}^{r}\frac{\alpha\sigma_i^2\langle v_i, f\rangle + \sigma_\nu^2\sigma_i^2}{(\sigma_i^2+\alpha)^3} < 0 \tag{3.11}$$

for $\alpha$ nearby zero. Hence moving along the trade-off curve away from the extremely volatile estimate at $\alpha = 0$ effectively reduces the mean square error.

**Consistency and Convergence Rate** Either of the above filter functions, by the way, can be shown to satisfy the inequality $r_\alpha(\sigma_i)/\sigma_i \le \alpha^{-1/2}$. Now let $\alpha = \sigma_\nu^p$ for some $p \in (0,2)$ and suppose the noise level can be made arbitrarily small. Considering the asymptotic behaviour of the resulting estimator as $\sigma_\nu \to 0$, we find that

$$\begin{aligned} 0 \le \mathrm{MISE}\left[\hat{f}\right] &\le \max_{1\le i\le n}\left(1-r_\alpha(\sigma_i)\right)^2\|f\|_2^2 + n\,\sigma_\nu^2\,\alpha^{-1} \\ &= O(\sigma_\nu^p) + O(\sigma_\nu^{(2-p)}) \end{aligned} \tag{3.12}$$

where the Landau-symbol is used in its conventional acceptation as $g = O(h) \Leftrightarrow \limsup_{h\to 0}|g/h| \le c \in \mathbb{R}$ ('Big Oh notation'). Then for $0 < p < 2$

$$\lim_{\sigma_\nu\to 0}\mathrm{MISE}\left[\hat{f}\right] = 0 \tag{3.13}$$

Evidently this requires that the variance of the noise be known beforehand, which is rarely the case. A filter function together with a parameter-rule that ensures asymptotic behaviour as in (3.13) is called consistent. A lot of research has been dedicated to quantify the rate of convergence achieved under certain conditions, see e.g. [27], [5], [11] and [3].

### 3.1.2 Parameter Rules

Equation (3.13) is unrealistic, not only in the sense that the noise level cannot be made arbitrarily small; in reality we are not even likely to know its variance; for hardly any real-world problem ever comes with such detailed information. To get better idea where on the trade-off curve to look for the optimal estimate, parameter rules are a valuable guideline. They can be grouped into the following two categories

- a-priori rules, assuming precise knowledge of the signal-to-noise ratio
- a-posteriori rules, based exclusively on the available data

One of the more successful techniques that falls into the latter class is called (General) Cross-Validation, where the adjective refers to its isotropic variant.

#### 3.1.2.1 Generalized Cross Validation (GCV)

The idea of cross-validation originally was developed by Grace Whaba [28] for underdetermined linear systems as arising in spline-smoothing. While it is clear that constructing a curve, and hence a continuous object, from a necessarily finite set of samples $(x_1, y_1), ..., (x_m, y_m)$ is ill-posed and requires a reasonable amount of regularization, the optimal choice of $\alpha$ is less so and remains to be determined. Given this particular background, we note that the paper deals with problems of the form

$$y = Ax + \nu \tag{3.14}$$

where $A$ is a discrete×continous matrix. We follow the paper in that we stick with the underdetermined case, at least initially, while translating it to finite dimensional spaces for simplicity.

In the course of derivation we shall need the following lemma, which we review here for completeness.

**Inverse of a block-partitioned Matrix.** *Let $P \in \mathbb{R}^{(m+n)\times(m+n)}$ invertible, partitioned in blocks $A \in \mathbb{R}^{m\times m}, B \in \mathbb{R}^{m\times n}, C \in \mathbb{R}^{n\times m}$ and $D \in \mathbb{R}^{n\times n}$ with $|D| \neq 0$. Let further*

$$E_1 = \begin{pmatrix} I & 0 \\ -D^{-1}B & D^{-1} \end{pmatrix} \quad E_2 = \begin{pmatrix} (A - BD^{-1}C)^{-1} & -(A - BD^{-1}C)^{-1}BD^{-1} \\ 0 & I \end{pmatrix} \tag{3.15}$$

*be elementary block-matrices. Then the inverse is given by (column-wise GE)*

$$\underbrace{\begin{pmatrix} A & B \\ C & D \end{pmatrix}}_{P} \cdot \underbrace{\begin{pmatrix} (A - BD^{-1}C)^{-1} & -(A - BD^{-1}C)^{-1}BD^{-1} \\ -D^{-1}C(A - BD^{-1}C)^{-1} & D^{-1} + D^{-1}C(A - BD^{-1}C)^{-1}BD^{-1} \end{pmatrix}}_{E_1E_2 = P^{-1}} = I$$

**Derivation** Let $A \in \mathbb{R}^{m\times n}$ with $m < n$ a matrix of maximal rank and $y = (y_1, ..., y_m)^T$ the data vector. Then the regularized pseudo inverse is given by

$$x_\alpha^\dagger = \arg\min \|Ax - y\|^2 + \alpha \|x\|^2 \tag{3.16}$$

Now suppose you eliminate one single datum $(x_k, y_k)$ from the set of measurements. For this purpose let $A_{\not k}$ denote the matrix obtained by deleting the $k$'th row in $A$ and likewise for vectors. The regularized estimate obtained by the reduced data

$$x_{\not k,\alpha}^\dagger := \arg\min \|A_{\not k}x - y_{\not k}\|^2 + \alpha \|x\|^2 \tag{3.17}$$

may then be assessed by comparing the predicted value at $x_k$ with the actual measurement $y_k$. If we let $a_k^T = e_k^T A$ denote the $k$'th row of $A$, we can define

$$r_{k,\alpha} := a_k^T x_{\not k,\alpha}^\dagger - y_k \tag{3.18}$$

to be the residual at sample $k$ for parameter $\alpha$. This proceeding is sometimes referred to as leave-out-one prediction. Effectively it amounts to splitting the data in two parts, the first of which is used to actually construct the solution that is being validated, subsequently,

against the latter one. The roles, of course, are completely arbitrary and therefore reversible. From this symmetry one quickly arrives at the idea of Cross-Validation, where each datum is singled out in turn to be validated against the remainder in the aforementioned way. We define the optimal parameter $\hat{\alpha} \in (0, \infty)$ to be the minimizer of the CV-function, defined as the square sum of the residuals

$$CV(\,\alpha\,) = \frac{1}{m} \sum_{k=1}^{m} r_{k,\alpha}^2 \tag{3.19}$$

A good solution — obtained as a function of the regularization parameter — should be able to predict new measurement points fairly well. In other words, adding or omitting a single datum should not result in a dramatic change of the trajectory, nor should the desired stability go to the detriment of an accurate approximation by making the trajectory lie far off the sample points. In either case we would pay a high penalty in terms of a large Cross-Validation residual. Minimizing the CV-function thus effectively results in a solution 'just smooth enough' by adjudicating an optimal compromise between stability and accuracy.

While equations (3.18) and (3.19) best render the idea of Cross-Validation on the conceptual level, they are impractical for computation. In fact, evaluating the CV function as above would entail the solution of $m$ linear systems, each of nearly the same size as the actual problem (3.14) — a cost clearly unproportional for the mere identification of a secondary parameter, important as it may be. However, it turns out that the computation can be greatly simplified by some basic observations.

First of all, note that in the underdetermined case presently discussed

$$x_{\not k,\alpha}^{\dagger} = A_{\not k}^T (A_{\not k} A_{\not k}^T + \alpha I_{m-1})^{-1} y_{\not k} \tag{3.20}$$

Now let $Q = AA^T$ and $M = Q + \alpha I_m$ for notational convenience and consider the case $k = 1$ By partitioning the matrices as follows

$$\underbrace{\left(\begin{array}{c} a_1^T \\ \hline A_{\not 1} \end{array}\right)}_{A} \cdot \underbrace{\left(\begin{array}{c|c} a_1 & A_{\not 1}^T \end{array}\right)}_{A^T} = \underbrace{\left(\begin{array}{c|c} q_{11} & q_{21}^T \\ \hline q_{21} & Q_{22} \end{array}\right)}_{Q} \tag{3.21}$$

we see that

$$\begin{aligned} A_{\not 1} A_{\not 1}^T &= Q_{22} \\ A_{\not 1} A_{\not 1}^T + \alpha I_{m-1} &= M_{22} \\ a_1^T A_{\not 1}^T &= q_{21}^T = m_{21}^T \end{aligned} \tag{3.22}$$

Using the identities (3.22) we find that the residual at the first sample point for a given parameter $\alpha$ is

$$\begin{aligned} r_{1,\alpha} &= a_1^T A_{\not 1} (A_{\not 1} A_{\not 1}^T + \alpha I_{m-1})^{-1} y_{\not 1} - y_1 \\ &= m_{21}^T M_{22}^{-1} y_{\not 1} - y_1 \end{aligned} \tag{3.23}$$

Now set $T := M^{-1}$ partitioned as above. Then according to the lemma it holds that $-t_{11}^{-1}t_{21}^T = m_{21}^T M_{22}^{-1}$, for $t_{11} \neq 0$. Substituting into (3.23) yields

$$\begin{aligned} r_{1,\alpha} &= -t_{11}^{-1} t_{12}^T y_{\not 1} - y_1 \\ &= -t_{11}^{-1} \left( e_1^T T y - t_{11} y_1 \right) - y_1 \\ &= -t_{11}^{-1} e_1^T T y \end{aligned} \tag{3.24}$$

We shall now consider the case where $k \neq 1$. To this effect let

$$P := (e_k^T, ..., e_m^T, e_1^T, ..., e_{k-1}^T) \in R^{m \times n} \tag{3.25}$$

be the (circular) shift operator, composed by row-wise stacking of the canonical base vectors in $\mathbb{R}^n$. By orthonormal transformation of the variables

$$\begin{aligned} \tilde{A} &:= PA & \tilde{Q} &:= \tilde{A}\tilde{A}^T = PQP^T \\ \tilde{y} &:= Py & \tilde{T} &:= (\tilde{Q} + \alpha I_m)^{-1} = PTP^T \end{aligned}$$

and exploiting the identities $A_{\not k} = \tilde{A}_{\not 1}$, $y_{\not k} = \tilde{y}_{\not 1}$ we find that

$$\begin{aligned} r_{k,\alpha} &= \tilde{a}_1^T \tilde{A}_{\not 1} (\tilde{A}_{\not 1} \tilde{A}_{\not 1}^T + \alpha I_{m-1})^{-1} \tilde{y}_{\not 1} - \tilde{y}_1 \\ &= -\tilde{t}_{11}^{-1} \cdot e_1^T \tilde{T} \tilde{y} \\ &= -(e_1^T P T P^T e_1)^{-1} \cdot e_1^T P T P^T \tilde{y} \\ &= -t_{kk}^{-1} \cdot e_k^T T y \end{aligned} \tag{3.26}$$

Putting together (3.24) and (3.26) we thus have

$$r_{k,\alpha} = -\left[ (Q + \alpha I_m)^{-1} y \right]_k / \left[ (Q + \alpha I_m)^{-1} \right]_{kk} \tag{3.27}$$

for $1 \leq k \leq m$. Now let $C := QT$, then

$$\begin{aligned} I - C &= I - Q(Q + \alpha I_m)^{-1} \\ &= \big( (Q + \alpha I_m) - Q \big) \big( Q + \alpha I_m \big)^{-1} \\ &= \alpha \cdot (Q + \alpha I_m)^{-1} \end{aligned} \tag{3.28}$$

and therefore $r_{k,\alpha} = [(I - C)y]_k \,/\, [I - C]_{kk}$. Note further that $Cy = Ax_\alpha^\dagger$ and hence

$$\mathrm{CV}(\,\alpha\,) = \frac{1}{m} \sum_{k=1}^{m} \frac{(y - Ax_\alpha^\dagger)_k^2}{(1 - c_{kk})^2} \tag{3.29}$$

with $C = \{c_{ij}\}_{ij}$ depending on $\alpha$ which is not emphasized in (3.29).

Certainly the above shortcut is more suitable for practical implementation; yet, without restrictive assumptions like shift-invariance, computing trace($C$) may still be difficult for problems of very large scale. (See [8] for a possible remedies)

Equation (3.29) identifies the CV function as a weighted Euclidean norm. Note, however, that while summing over the terms in either the numerator or denominator separately is invariant under orthogonal transform the sum of their componentwise ratio is not. A simple

rotation of the problem thus most likely results in a different minimizer, which is an unpleasant property, after all.

This observation has motivated a variant of CV that keeps its minimum under orthogonal transform, called Generalized Cross Validation (GCV). To this effect, the diagonal elements of $I - C$ in the denominator of (3.29) are simply replaced by their average $\text{trace}(I - C)/m$ which effectively corresponds to a weighted sum of the residuals

$$\text{GCV}(\,\alpha\,) = \frac{1}{m}\sum_{k=1}^{m} r_{k,\alpha}^2\, w_k \tag{3.30}$$

with weights

$$w_k = \frac{\left[(AA^T + \alpha I_m)^{-1}\right]_{kk}^2}{\left[\frac{1}{m}\,\text{trace}\left((AA^T + \alpha I_m)^{-1}\right)\right]^2} \tag{3.31}$$

By modifying (3.29) in this sense we finally obtain a compact and computationally efficent representation of the GCV function as follows

$$\text{GCV}(\,\alpha\,) = m \cdot \frac{\left\| y - Ax_\alpha^\dagger \right\|_2^2}{\text{trace}(I - C_\alpha)^2} \tag{3.32}$$

Sometimes the dimensionality factor $m$ is dropped in (3.32), being irrelevant for the minimization problem. We recall that $C_\alpha = AA^T(AA^T + \alpha I_m)^{-1}$ for the underdetermined case presently discussed, but it is not difficult to show that the above equations hold true as well for the overdetermined case, if we let $C = A^T(A^TA + \alpha I_m)^{-1}A^T$.

A particularly useful representation of the GCV function can be obtained by spectral factorization of the matrix $A$. Thus let $A = USV^H$ the SVD, with singular values $\sigma_1, ..., \sigma_m$ and left singular vectors $u_1, ..., u_m$ the columns of $U$. Then

$$\text{GCV}(\,\alpha\,) = m \cdot \sum_{k=1}^{m} \left(\frac{|\langle u_i, y\rangle|}{\sigma_i^2 + \alpha}\right)^2 \Big/ \sum_{k=1}^{m} \left(\frac{1}{\sigma_i^2 + \alpha}\right)^2 \tag{3.33}$$

can be minimized using standard optimization techniques. Figure 3.2 shows the generalized cross-validation function for the testimage 2.2 blurred with a Gaussian kernel in the presence of additive white Gaussian noise (AWGN). (The eigenvalues defining the Optical Transfer Function were $\sigma_{x,y} = \exp(-\theta r^{5/3})$ with $r = \sqrt{x^2 + y^2}$ the radius of frequency and $\theta = 0.005$). The GCV-function assumes its minimum at $\alpha_0 = 3.237 \times 10^{-7}$. Figure 3.5 on page 36 represents a sequence of restored images for varying $\alpha$ obtained by stepping through an interval around the numerically determined minimizer of (3.32). In agreement with the GCV-criterion, solutions for $\alpha < \alpha_0$ strike the beholder as too rough, while choosing $\alpha > \alpha_0$ results in an oversmoothed estimate and significant loss of detail. If one were to choose 'by eye', based exclusively on the visual impression, most would probably pick the image on the right-hand-side in the middle row which is fact the solution for $\alpha = \alpha_0$. So far we have argued on a purely heuristic level without a strict definition of optimality — and all the less a formal proof. As an in-depth discussion is beyond the scope of this paper, we leave it at the statement that under certain, not too restrictive, conditions

$$\lim_{m\to\infty} \text{E}\left\{\text{GCV}(\alpha)\right\} = \text{E}\left\{\frac{1}{m}\left\| Ax - Ax_\alpha^\dagger \right\|_2^2\right\} \tag{3.34}$$

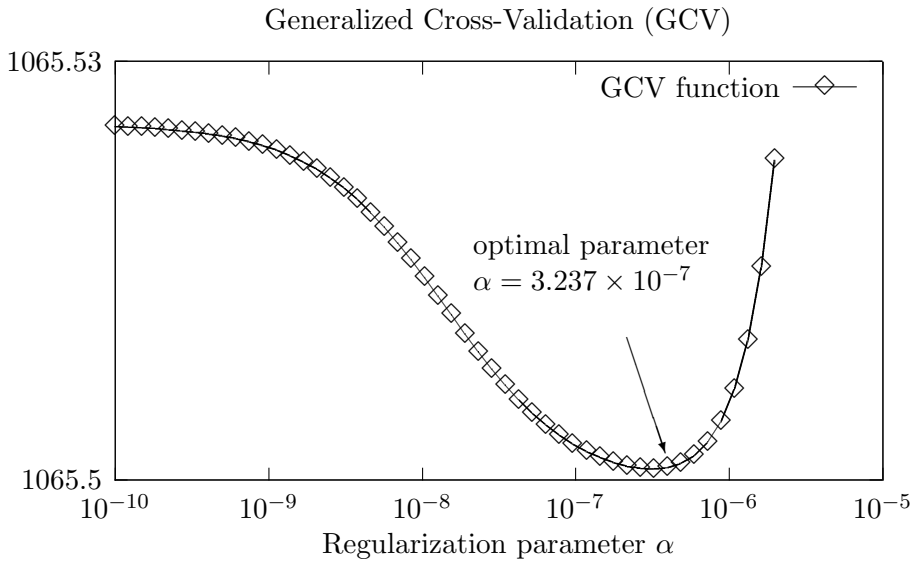



Figure 3.2: Generalized Cross Validation (GCV) function

The claim for goodness, then, is based on the asymptotic convergence to the mean square true prediction error given by the right-hand-side of (3.34). For details and proof the interested reader is referred to the aforementioned paper by Whaba.

#### 3.1.2.2 L-Curve

Another technique to predict the optimal regularization parameter is named after the typical shape of the trade-off curve between bias and variance — or between residual and solution norm, for that matter. Figure 3.3 shows a logscale plot of $\|g - Af\|_2$ versus $\|f\|_2$ evoquative of the capital letter 'L', each sample point representing a solution for a different value of $\alpha$. Again the experiment has been conducted with the same test-image 2.2 as in the part on GCV (Gaussian OTF, AWGN with $\sigma = 3$). According to this heuristic the optimal choice for the parameter $\alpha$ would be at the corner or 'knee' of the 'L'. Translating this informal description into mathematical language it has been suggested by [10] to look for a maximum in the second derivative of the parameterized graph, which characterizes the point of greatest curvature. This particular method has not been extensively tested but it seems as if there was an inherent tendency toward overly smooth estimates by predicting a greater value for $\alpha$ than necessary. For an in-depth discussion and possible alternatives see, again, [10] and [9]. Figure 3.4 shows, once more but this time in parallel, residual and solution norm as a function of the regularization parameter.

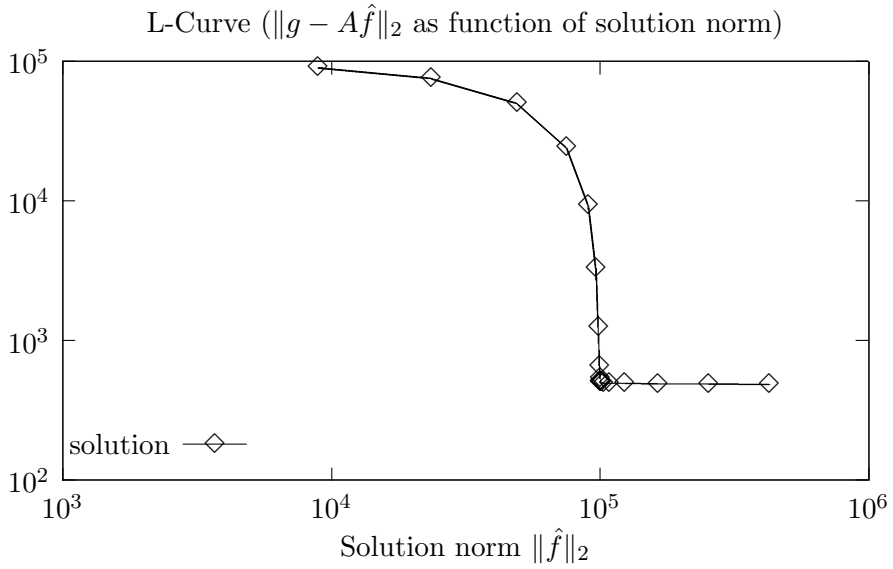



Figure 3.3: L-Curve (trade-off between data-fit and solution norm)

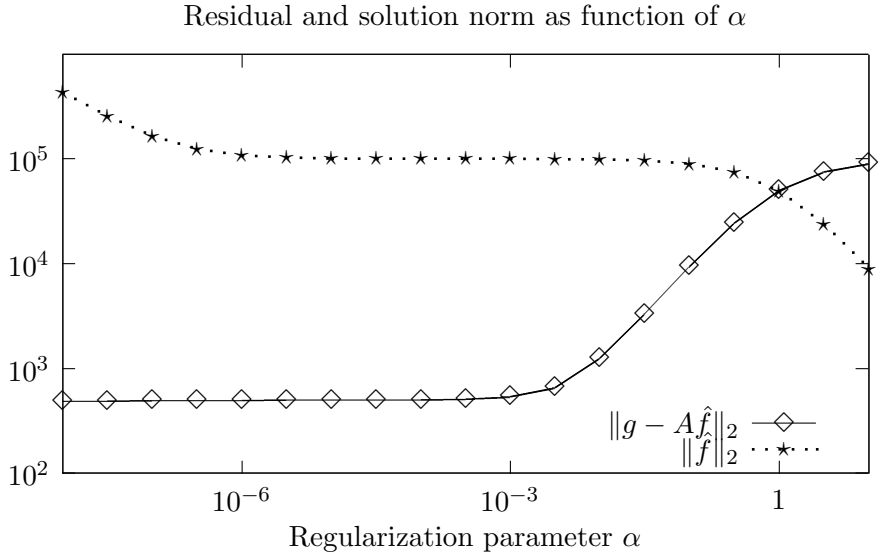



Figure 3.4: Residual and solution norm, varying with the regularization parameter

### 3.1.3 Generalized Tikhonov Regularization

Classical Tikhonov Regularization, as presented in (3.6), is susceptible of an important generalization that readily extends beyond spectral filters. To this effect it is useful to recast it first as an optimization problem. In fact, the Tikhonov estimate

$$f_\alpha^{(Tv)} = V R_\alpha S^{-1} U^H g \qquad\qquad R_\alpha = (S^2 + \alpha I)^{-1} S^2 \tag{3.35}$$

can be shown to be the minimizer of

$$\phi(f) = \varphi_d(f) + \alpha\,\varphi_m(f) \tag{3.36}$$

where $\varphi_d$ and $\varphi_m$ are cost-functionals representing data misfit and model norm respectively

$$\varphi_d(f) = \|g - Af\|_2 \qquad\qquad \varphi_m(f) = \|f\|_2 \tag{3.37}$$

To see that $f_\alpha^{(Tv)} = \arg\min_f \phi(f)$ we differentiate (3.36) with respect to $f$ and set

$$0 = \nabla_f \phi(\hat{f}) = 2\left((A^H A + \alpha I)\hat{f} - A^H g\right) \tag{3.38}$$

which leads to the modified system of normal equations $(A^H A + \alpha I)\hat{f} = A^H g$. Using the SVD representation we can solve for the minimizer $\hat{f}$ and get

$$\begin{aligned} \left(VS^2V^H + \alpha I\right)\hat{f} &= VS^HU^Hg \\ V\left(S^2 + \alpha I\right)V^H\hat{f} &= VS^2S^{-1}U^Hg \\ \hat{f} &= V\left(S^2 + \alpha I\right)^{-1}S^2S^{-1}U^Hg \end{aligned} \tag{3.39}$$

as claimed. Thus Tikhonov Regularization has been shown to be equivalent with minimizing a weighted linear combination of two cost- or penalty-functionals. While the misfit $\varphi_d$ is an objective criterion ensuring a minimum of consistency with the observed data, $\varphi_m$ is less so. Reflecting our conjectures about a plausible solution, it is supposed to penalize 'rough' estimates with high energy. An obvious way to generalize this pattern is by letting

$$\varphi_m(f) = \|f - f_0\|_{\mathbb{L}} \tag{3.40}$$

with $\|\cdot\|_{\mathbb{L}} = \|L(\cdot)\|_2$ for a linear map $L$ defining a (semi-)norm on the domain space and $f_0 \in \mathbb{R}^n$ a default solution. Note that the classical variant (3.35) is obtained as a special case by letting $(L, f_0) = (I, 0)$. So long as $L$ is square and invertible, the general case (3.40) may reduced to the standard form (3.35) by the following change of variables

$$\begin{aligned} A' &:= AL^{-1} \\ g' &:= g + Af_0 \\ f' &:= L(f - f_0) \end{aligned} \tag{3.41}$$

After solving for the minimizer of the transformed problem, say, $\hat{f}'$, the solution to the original problem is readily obtained by computing $\hat{f} = L^{-1}(\hat{f}' + f_0)$.

Some difficulties arise, however, when $L$ is underdetermined. A prominent example of this case is so-called k'th order regularization, where $L$ is a discrete approximation of the derivative operator. For one-dimensional data this will most likely be a finite difference matrix of the corresponding order, like

$$D_n^{(1)} = \begin{pmatrix} 1 & -1 & & \\ & 1 & -1 & \\ & & \ddots & \ddots \\ & & & 1 & -1 \end{pmatrix}, \qquad D_n^{(2)} = \begin{pmatrix} 1 & -2 & 1 & & \\ & 1 & -2 & 1 & \\ & & \ddots & \ddots & \ddots \\ & & & 1 & -2 & 1 \end{pmatrix}$$

etc, or any linear combination we see fit. The pattern naturally extends to higher orders and is rendered by the following line of MATLAB pseudo-code

```
%% construct finite difference matrix of order k
%% and dimension (n-k) x n
D(n,k) = eye(n-k, k) * (eye(n) - circshift(eye(n), [0, 1]))^k
```

Effectively, k'th order regularization gives some control as to the rate of change in the resulting estimate (constant, linear, quadratic, etc). Note that this framework contains the classical variant (3.35) with $L = I = D_n^{(0)}$ as a special case. For $k > 0$, however, we see that $D_n^{(k)} \in \mathbb{R}^{(n-k)\times n}$ is a rectangular matrix and, by design, has a non-trivial null-space spanned by the first $k$ moments of the sequence $0, ..., (n-1)$

$$\begin{aligned} \ker(D_n^{(k)}) &= \operatorname{span}\langle \{v_n^{(0)}, v_1^{(1)}, ..., v_n^{(k-1)}\} \rangle \\ v_n^{(i)} &= (0^i, 1^i, ..., (n-1)^i)^H \end{aligned} \tag{3.42}$$

Hence $\|\cdot\|_{\mathbb{L}}$ constitutes merely a semi-norm for $L = D_n^{(k)}$ and $k > 0$. Inclusion of boundary conditions would lend the matrix square shape without however correcting rank-deficiency. Since the transformation (3.41) will not do under these conditions, we deal with a particular case in some detail. For a more general discussion of possible strategies see [9] and the references therein.

It will have been noted that finite difference matrices, essentially, are circulant matrices representing convolution with a high-pass filter kernel. This principle expands very naturally to the case of interest, if we replace the stencils $[\ 1 \quad -1\ ]$ and $[\ -1 \quad 2 \quad -1\ ]$ by their two-dimensional counterparts

$$\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \qquad\qquad \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$$

approximating first order directional derivatives and

$$\begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

for the second order derivative. Now suppose that both $A$ and $L$ represent periodic 2D-convolution with a PSF and one of the above kernels respectively. Then $U = V = F^H$ the inverse Fourier Transform-Matrix and

$$A = FSF^H \qquad S = \operatorname{diag}(\sigma_1, ..., \sigma_n) \tag{3.43a}$$

$$L = F\Lambda F^H \qquad \Lambda = \operatorname{diag}(\lambda_1, ..., \lambda_n) \tag{3.43b}$$

Setting $\nabla\phi(\hat{f};\ L, f_0) = 0$ yields

$$(A^H A + \alpha L^H L)\hat{f} = A^H g + L^H L f_0 \tag{3.44}$$

. By use of (3.43) we find that

$$\begin{aligned}\left(FS^2F^H + \alpha F\Lambda^2F^H\right)\hat{f} &= FS^HF^Hg + \alpha F\Lambda^2F^Hf_0 \\ F\left(S^2 + \alpha\Lambda^2\right)F^H\hat{f} &= F\left(S^HF^H\,g + \alpha\Lambda^2F^H\,f_0\right) \\ F^H\hat{f} &= \left(S^2 + \alpha\Lambda^2\right)^{-1}\left(S^2S^{-1}F^Hg + \alpha\Lambda^2F^Hf_0\right)\end{aligned} \tag{3.45}$$

By writing out (3.45), we get

$$\langle v_i, \hat{f}\rangle = r_\alpha(\sigma_i, \lambda_i)\,\frac{1}{\sigma_i}\langle v_i, g\rangle + (1 - r_\alpha(\sigma_i, \lambda_i))\,\langle v_i, f_0\rangle \tag{3.46}$$

for $1 \leq i \leq r$, where

$$r_\alpha(\sigma_i, \lambda_i) = \frac{\sigma_i^2}{\sigma_i^2 + \alpha\lambda_i^2} \tag{3.47}$$

generalizes the classical Tikhonov filter function. Equation (3.46) characterizes the resulting estimate as linearly interpolating between the principal and the default solution. Note that due to the degeneracy of $L$ some of the eigenvalues $\lambda_i$ will be zero and consequently no regularization is performed for the corresponding subspace — which is innocuous so long as only frequencies of typically high signal-to-noise ratio are concerned.

**Maximum Entropy (ME)** While sticking to the principle of composite cost-functionals, further generalization may be achieved by lifting the linearity constraint on $L$ and take $\varphi_m$ to be an arbitrary functional. One such method with a non-linear penalty-function is Maximum Entropy, where

$$\varphi_m^{ME}(f) = \sum_{i=1}^{n} f_i \log f_i \tag{3.48}$$

Without trying to elucidate the philosophical background or discussing information theory, we simply motivate (3.48) by the following observation.

Suppose the probability for any one intensity unit $X$ to end up at location $i$ in the image plane is $P(X = i) = p_i$ with $\sum_{i=1}^n p_i = 1$. Then the image $f$, given by the total of $\sum_{i=1}^n f_i = c$ independent trials with $\dim(f) = n$ possible outcomes each, obeys a multinomial distribution with pmf

$$\rho(f) = \frac{c!}{\prod_{i=1}^n f_i!}\prod_{i=1}^{n} p_i^{f_i} \tag{3.49}$$

Now the implicit assumption leading to the Maximum Entropy functional (3.48) lets

$$p_1 = p_2 = ... = p_n = \frac{1}{n} \tag{3.50}$$

Then $\prod_{i=1}^n p_i^{f_i} = n^{-c}$ and hence

$$\log\rho(f) = \log c! - \sum_{i=1}^{n}\log f_i! - c\log n \tag{3.51}$$

Using Stirling's approximation of the factorial $n! \approx e^{n \log n + O(n)}$ and neglecting inconsequential constants, we see that maximizing the log-likelihood (3.51) is roughly equivalent to minimizing the Maximum Entropy functional (3.48)

$$\arg\max_f \rho(f) = \arg\min_f \varphi_m^{ME}(f) \tag{3.52}$$

For a more in-depth analysis of the ME principle and contextual background the reader is referred to [12], [24] or any of the standard works on information theory.

## 3.2 Bayesian MAP Estimation

So far, we have considered ways to estimate a *deterministic* parameter of an otherwise known distribution. When it comes to inverse — and thus notoriously ill-conditioned — problems, regularization has been shown to be crucial for success. This principle of gearing the solution toward more plausible candidates while discarding others which disagree with our presumed knowledge about the 'true' image is rendered within a Bayesian framework by the concept of *prior probability*. What makes this model slightly different from those considered so far, then, is the status of the parameter as a random variable of its own. The idea that some solutions by itself are more likely than the remaining ones translates as a non-flat prior pdf $p(f)$.

In Baysian parameter estimation, the 'solution' to the problem is actually given in terms of the conditional density or *posterior pdf*

$$p(f|g) = \frac{p(g|f) \cdot p(f)}{p(g)} \tag{3.53}$$

where 'prior' and 'posterior' refer to the state of information before and after the observation has occurred. For most real-world applications, however, we are asked to give a single estimate rather than a whole distribution. Basically, there are two strategies which do not necessarily agree

- Mode of the posterior pdf
  $\rightsquigarrow$ Maximum A-Posteriori estimate $\hat{f}_{\text{map}} = \arg\max_f p(f|g)$
- Mean of the posterior pdf
  $\rightsquigarrow$ Minimum Mean Square Error estimate $\hat{f}_{\text{mmse}} = \mathrm{E}\{f|g\}$

It seems to be common policy to choose $f_{\text{mmse}}$ so long as the pdf is unimodal and $f_{\text{map}}$ otherwise.

**MMSE Estimate** Note that by modelling the 'true' image as a random variable, we have

$$\begin{aligned} \mathrm{MISE}\left[\hat{f}\right] &= \mathrm{E}_f\left\{\mathrm{E}_g\left\{\left\|f - \hat{f}(g) \,\middle|\, f\right\|_2^2\right\}\right\} \\ &= \int_F \int_G \left\|f - \hat{f}(g)\right\|_2^2 \rho(f,g)\, dg\, df \end{aligned} \tag{3.54}$$

with $F$, $G$ the vector spaces associated with the respective lowercase variables and $\rho(f,g)$ the joint probability density function. It remains to prove that $\hat{f}_{\text{mmse}}(g) = \mathrm{E}\{f|g\}$ has minimal

mean square error. To this effect let $\hat{f}(g)$ be an arbitrary estimate. Then

$$\begin{aligned}\left\|f-\hat{f}(g)\right\|_2^2 &= \left\|f-\hat{f}_{\text{mmse}}(g)+\hat{f}_{\text{mmse}}(g)-\hat{f}(g)\right\|_2^2 \\ &= \left\|f-\hat{f}_{\text{mmse}}(g)\right\|_2^2+\left\|\hat{f}_{\text{mmse}}(g)-\hat{f}(g)\right\|_2^2-2\left(f-\hat{f}_{\text{mmse}}(g)\right)^H\left(\hat{f}_{\text{mmse}}(g)-\hat{f}(g)\right)\end{aligned}$$

Taking expectation for fixed $g$ on both sides and using $\mathrm{E}_f\{(f-\hat{f}_{\text{mmse}}(g))\}=0$ yields

$$\begin{aligned}\mathrm{E}\left\{\left\|f-\hat{f}(g)\right\|_2^2 \;\middle|\; g\right\} &= \mathrm{E}\left\{\left\|f-\hat{f}_{\text{mmse}}(g)\right\|_2^2 \;\middle|\; g\right\}+\mathrm{E}\left\{\left\|\hat{f}_{\text{mmse}}-\hat{f}(g)\right\|_2^2 \;\middle|\; g\right\} \\ &\geq \mathrm{E}\left\{\left\|f-\hat{f}_{\text{mmse}}(g)\right\|_2^2 \;\middle|\; g\right\}\end{aligned} \tag{3.55}$$

for the second term on the right-hand-side is non-negative. Since $g$ was chosen arbitrary, we obtain, by taking expectation with respect to $g$ and exploiting monotony of the integral

$$\mathrm{MISE}\left[\hat{f}\right] \geq \mathrm{MISE}\left[\hat{f}_{\text{mmse}}\right] \tag{3.56}$$

which proves our claim that the mean of the posterior pdf is optimal in the sense of (3.54).

**Gaussian Case** Due to its flexibility, Bayesian MAP estimation is extremely powerful and versatile. Owing to the practical interest of this thesis and its limited scope, however, we shall only consider the standard case where both the prior $p(f)$ and — consistent with what has been done so far — the data for a given parameter $p(g|f)$ are Gaussian. We start by stating the following theorem:

**Conditional pdf of random variables with jointly Gaussian distribution.** *Let $X$ and $Y$ be two random variables with mean $\mu_X$ and $\mu_Y$ respectively and jointly Gaussian distribution. Then for a given observation $Y=Y_0$ the conditional mean of $X$ is given by*

$$\mathrm{E}\left\{X \mid Y=Y_0\right\} = \mu_X + \mathrm{Cov}\left(X,Y\right)\mathrm{Var}\left[Y\right]^{-1}\left(Y_0-\mu_Y\right) \tag{3.57}$$

*with conditional variance equal to*

$$\mathrm{Var}\left[X \mid Y=Y_0\right] = \mathrm{Var}\left[X\right] - \mathrm{Cov}\left(X,Y\right)\mathrm{Var}\left[Y\right]^{-1}\mathrm{Cov}\left(Y,X\right) \tag{3.58}$$

**Proof:** Let $Z=(X^T\,Y^T)^T$ be the compound of the two random variables. By assumption $Z\sim\mathcal{N}(\mu_Z,\Lambda_Z)$ is normally distributed with

$$\mu_Z = \begin{pmatrix}\mu_X\\ \mu_Y\end{pmatrix} \qquad\qquad \Lambda_Z = \begin{pmatrix}\mathrm{Var}\left[X\right] & \mathrm{Cov}\left(X,Y\right)\\ \mathrm{Cov}\left(Y,X\right) & \mathrm{Var}\left[Y\right]\end{pmatrix} \tag{3.59}$$

and conditional density $f_{X|Y}(X|Y=Y_0)=f_Z(Z)/f_Y(Y_0)$

$$f_{X|Y}(X|Y=Y_0) = \sqrt{\frac{|\mathrm{Var}\left[Y\right]|}{(2\pi)^{\dim(Y)}|\Lambda_Z|}}\exp\left(-\frac{1}{2}\left(\bar{Z}^H\Lambda_Z^{-1}\bar{Z}-\bar{Y}_0^H\,\mathrm{Var}\left[Y\right]^{-1}\bar{Y}_0\right)\right) \tag{3.60}$$

where $\bar{Y}_0=Y_0-\mu_Y$ and $\bar{Z}=Z-\mu_Z$ denote the corresponding zero-mean variables. Now let $P=\Lambda_Z$ the block-partitioned matrix in lemma 3.15 on page 23, so that $\Lambda^{-1}=E_1E_2=:L$.

Due to the symmetry of the variance matrix $A = A^H, B = C^H$ and $D = D^H$ and therefore $L_{12} = L_{21}^H$. Using this equality, we want to write the first term of the exponential in (3.60)

$$\begin{aligned} \bar{Z}\Lambda_Z^{-1}\bar{Z} &= \bar{X}^H L_{11}\bar{X} + 2\bar{X}^H L_{12}\bar{Y}_0 + \bar{Y}_0^H L_{22}\bar{Y}_0 \\ &= (\bar{X} - M)^H L_{11}(\bar{X} - M) + R \end{aligned} \tag{3.61}$$

as a quadratic form in a shifted variable $(\bar{X} - M)$ plus a remainder $R$ independent of $X$. By comparing the coefficient of the linear term in $X$ we find that

$$\begin{aligned} M &= -(L_{11})^{-1}L_{12}\,\bar{Y}_0 & R &= \bar{Y}_0^H\,L_{22}\,\bar{Y}_0 - M^T\,L_{11}\,M \\ &= BD\,\bar{Y}_0 & &= \bar{Y}_0^H(L_{22} - L_{21}L_{11}^{-1}L_{12})\bar{Y}_0 \\ &= \mathrm{Cov}\,(X,Y)\,\mathrm{Var}\,[Y]^{-1}\;\bar{Y}_0 & &= \bar{Y}_0^H D^{-1}\bar{Y}_0 \end{aligned}$$

Thus the exponential in (3.60) simplifies to

$$\begin{aligned} \bar{Z}^H\Lambda_Z^{-1}\bar{Z} - \bar{Y}_0^H\,\mathrm{Var}\,[Y]^{-1}\,\bar{Y}_0 &= (\bar{X} - M)^H L_{11}(\bar{X} - M) + R - \bar{Y}_0\,\mathrm{Var}\,[Y]^{-1}\,\bar{Y}_0 \\ &= (X - \mu_X - M)^H L_{11}(X - \mu_X - M) \\ &= (X - \mu_{X|Y})^H\Lambda_{X|Y}^{-1}(X - \mu_{X|Y}) \end{aligned} \tag{3.62}$$

with

$$\begin{aligned} \mu_{X|Y} &= \mu_X + M = \mu_X + \mathrm{Cov}\,(X,Y)\,\mathrm{Var}\,[Y]^{-1}\,(Y_0 - \mu_Y) \\ \Lambda_{X|Y} &= (L_{11})^{-1} = A - BD^{-1}C = \mathrm{Var}\,[X] - \mathrm{Cov}\,(X,Y)\,\mathrm{Var}\,[Y]^{-1}\,\mathrm{Cov}\,(Y,X) \end{aligned} \tag{3.63}$$

Finally consider that

$$\begin{aligned} |P| &= |D| \cdot |PE_1| \\ &= |D| \cdot |A - BD^{-1}C| \end{aligned} \tag{3.64}$$

and therefore $|\Lambda_Z| = |\mathrm{Var}\,[Y]| \cdot |\Lambda_{X|Y}|$. By substituting into (3.60) we obtain for the conditional density

$$f_{X|Y}(X|Y = Y_0) = \frac{1}{\sqrt{(2\pi)^{\dim(Y)}|\Lambda_{X|Y}|}}\exp\left(-\frac{1}{2}(X - \mu_{X|Y})^H\Lambda_{X|Y}^{-1}(X - \mu_{X|Y})\right) \tag{3.65}$$

which is Gaussian with parameters (3.63) as claimed. ■

Now consider the case of a linear mapping $g = Af$ with Gaussian noise and prior

$$g|f \sim \mathcal{N}(Af, \Phi_\nu) \qquad\qquad f \sim \mathcal{N}(\mu_f, \Phi_f) \tag{3.66}$$

By theorem (3.57) we know that the posterior pdf is also Gaussian with

$$\hat{f}_{\mathrm{map}} = \hat{f}_{\mathrm{mmse}} = \mu_f + \Phi_f A^H\big(A\Phi_f^{-1}A^H + \Phi_\nu^{-1}\big)\big(g - A\mu_f\big) \tag{3.67}$$

where the identities $\mathrm{Cov}\,(f,g) = \Phi_f A^H$ and $\mathrm{Var}\,[g] = A\Phi_f A^H + \Phi_\nu$ have been exploited. However, by substituting into (3.53) and directly maximizing the log-likelihood in the usual manner

$$\hat{f}_{\max} = \arg\max_f \log p(f|g) \tag{3.68}$$

we obtain

$$\hat{f}_{\max} = \left(A^H \Phi_\nu^{-1} A + \Phi_f\right)^{-1} \left(A^H \Phi_\nu^{-1} g + \Phi_f^{-1} \mu_f\right) \tag{3.69}$$

which looks different at first sight. To see that (3.67) and (3.69) are actually identical, consider that

$$\begin{aligned} \left(A^H \Phi_\nu^{-1} A + \Phi_f^{-1}\right) \operatorname{Cov}(f, g) &= A^H \Phi_\nu^{-1} A \Phi_f A^H + A^H \\ &= A^H \Phi_\nu^{-1} \left(A \Phi_f A^H + \Phi_\nu\right) \\ &= A^H \Phi_\nu^{-1} \operatorname{Var}[g] \end{aligned} \tag{3.70}$$

or, equivalently, $\operatorname{Cov}(f, g) \operatorname{Var}[g]^{-1} = Q\, A^H \Phi_\nu^{-1}$ with $Q := (A^H \Phi_\nu^{-1} A + \Phi_f^{-1})^{-1}$. Using this representation, we can finally establish the identity

$$\begin{aligned} \hat{f}_{\text{map}} &= \operatorname{Cov}(f, g) \operatorname{Var}[g]^{-1} (g - A\mu_f) + \mu_f \\ &= Q\left(A^H \Phi_\nu^{-1} (g - A\mu_f) + Q^{-1} \mu_f\right) \\ &= Q\left(A^H \Phi_\nu^{-1} (g - A\mu_f) + (A^H \Phi_\nu^{-1} A + \Phi_f^{-1}) \mu_f\right) \\ &= Q\left(A^H \Phi_\nu^{-1} g + \Phi_f^{-1} \mu_f\right) \\ &= \hat{f}_{\max} \end{aligned} \tag{3.71}$$

**Connection to Tikhonov Regularization** Bayesian MAP estimation provides a flexible and at the same time elegant framework for the incorporation of regularization constraints. For our simple case where the model is linear and the densities involved all Gaussian, however, the Bayesian approach does not result in a substantial advance compared with previously discussed methods. We conclude this chapter by noting that for

$$\begin{aligned} \Phi_\nu &= \sigma_\nu^2 I \\ (\Phi_f, \mu_f) &= (\sigma_f^2 I, 0) \end{aligned} \tag{3.72}$$

the Bayesian MAP estimate

$$\begin{aligned} \hat{f}_{\text{map}} &= A^H \left(A A^H + \frac{\sigma_\nu^2}{\sigma_f^2} I\right)^{-1} g \\ &= V S^2 (S^2 + \alpha)^{-1} S^{-1} U^H g \end{aligned} \tag{3.73}$$

is indeed equivalent to the classical Tikhonov estimate with $\alpha = \sigma_\nu^2 / \sigma_f^2$ the reciproque of the signal-to-noise ratio.

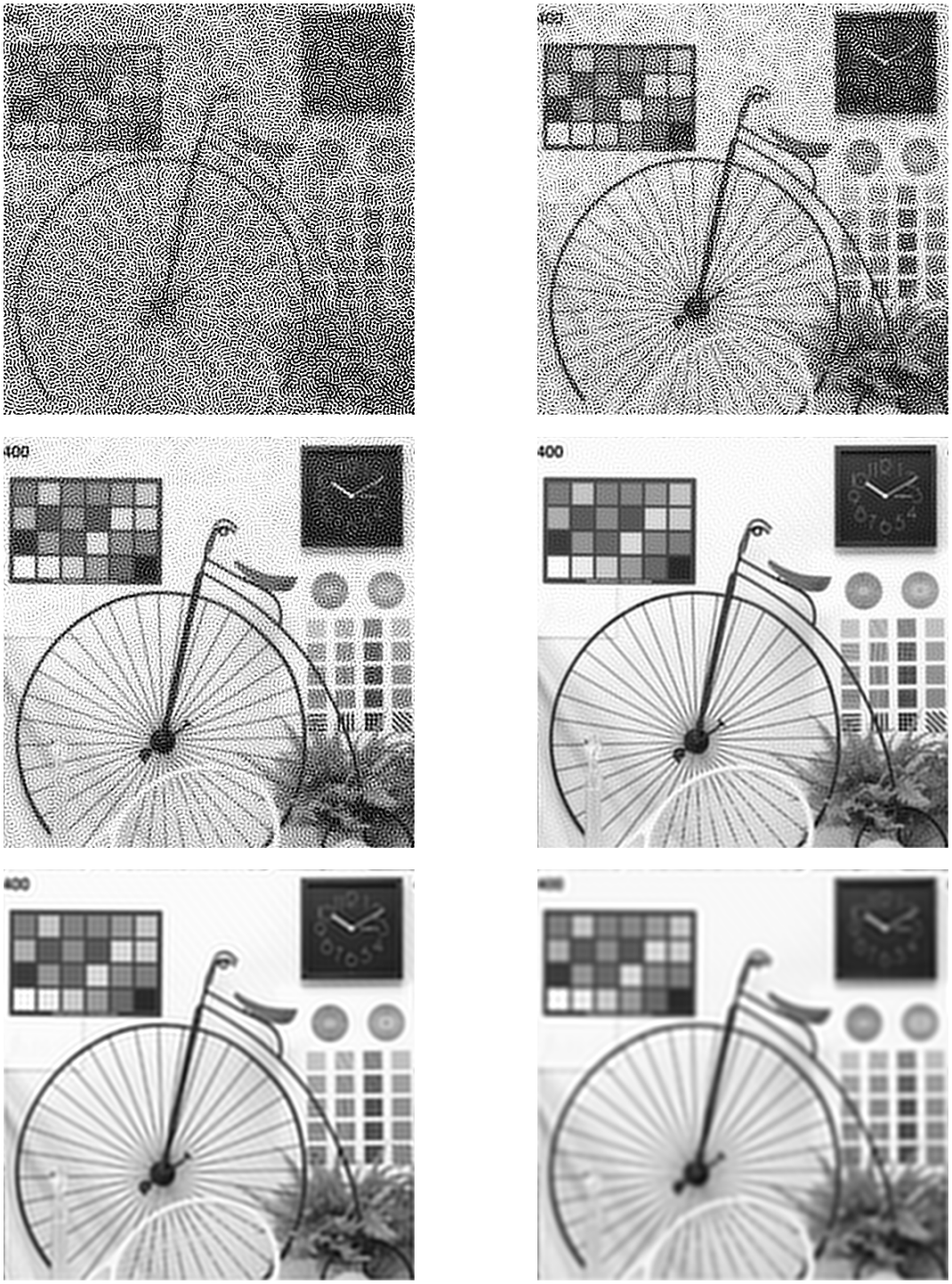


Figure 3.5: Solutions for different $\alpha$, from ultra-rough to oversmooth

# 4 Algorithms

After exploring the mathematical theory of linear inverse problems in some detail we now present a selection of derived algorithms. Spurred by the development of new imaging technologies, the problem has been in the focus of keen investigation ever since the 60's of the last century, calling the attention of many ambitious researchers. Over the years those have come up with a variety of approaches whose sheer number is symptomatic for the unrelenting interest dedicated to image deconvolution. Any attempt for completeness, therefore, is bound to fall short of its claim.

Orientation is facilitated by classifying the variety of algorithms on the basis of well-defined criteria. The most comprehensive and ambitious attempt to date can be found in the introductory chapter of [13], reflecting the state of the art of image restoration in 1989. Although somewhat out of data by now, to our knowledge it is still the best endeavor in this regard. (The distinction between stochastic and deterministic at the top node in his classification strikes us as unfortunate, though, being mostly a matter of perspective and therefore somewhat arbitrary.)

The algorithms, all of certain renown in their category, have been chosen so as to give a reasonably broad overview, covering both direct and iterative approaches. The comparably recent branch known as 'blind' deconvolution is represented by an EM-based algorithm. Diversity finally extends to the level of transform domains, where both Fourier and Wavelet approaches are considered.

## 4.1 Wiener Filter

The concept of the Wiener Filter is straightforward and naturally follows by developing the ideas outlined in the section on Bayesian MAP Parameter Estimation. By definition, the Wiener estimate $\hat{f}_W$ is the (affine-)linear function in the data $g$ of minimal mean (integrated) square error

$$\hat{f}_W = \arg\min_{\hat{f}\in\Omega} \mathrm{MISE}\left[\hat{f}\right] \qquad\qquad \Omega = \{\hat{f} \mid \hat{f}(g) = Rg + r\} \tag{4.1}$$

with $R$ a linear map and $r$ a constant. Note that the pristine image $f$ is modelled as a random variable. We shall see that due to the linearity constraint, it suffices that second order of the problem be known. For the case of Gaussian prior and data we already know that $\hat{f}_{\mathrm{mmse}}$ is (affine-)linear in $g$ and therefore

$$\hat{f}_W = \hat{f}_{\mathrm{mmse}} = \mathrm{E}\{f|g\} = \mu_f + \mathrm{Cov}(f,g)\,\mathrm{Var}[g]^{-1}(g - \mu_g) \tag{4.2}$$

according to (3.56) and (3.57), but the same is not true in general.

**Derivation** For the general case we are going to use the following lemma, given here without proof.

**Lemma.** *Let $x : \Omega \longrightarrow \mathbb{R}^n$ be a random variable with density $\rho(x)$. Then for $h(x;\ p) : \Omega \times P \longrightarrow \mathbb{R}^m$ with both $h(x;\ p)$, $\partial_p h(x;\ p) \in \mathcal{L}^1(\Omega)$ for each $p \in P$ we can interchange expectation and differentiation*

$$\partial_p \operatorname{E}\{h(x;\ p)\} = \partial_p \int h(x;\ p)\rho(x)dx = \int \partial_p h(x;\ p)\rho(x)dx = \operatorname{E}\{\partial_p h(x;\ p)\} \tag{4.3}$$

By postulation, the mean (integrated) square error functional (restricted on $\Omega$) has a minimum at $\hat{f} = \hat{f}_W$ and thus

$$\begin{aligned} 0 = \nabla_r \operatorname{MISE}\left[\hat{f}_W\right] &= \nabla_r \operatorname{E}\left\{(f - Rg - r)^H (f - Rg - r)\right\} \\ &= 2\,(r + R - \mu_f) \end{aligned} \tag{4.4}$$

will hold. (Here the lemma has been used.) The gradient vanishes for $r = \mu_f - R\mu_g$ and we obtain

$$\hat{f}_W(g) = R(g - \mu_f) + \mu_f \tag{4.5}$$

The identification of the remaining parameter is simplified by considering the equivalent mean-subtracted problem with $f' := f - \mu_f$ and $g' = g - \mu_g$. Again, we differentiate with respect to $R_{ij}$ and solve for the root of the partial derivative

$$\begin{aligned} 0 = \frac{\partial}{\partial R_{ij}} \operatorname{MISE}\left[\hat{f}_W\right] &= \frac{\partial}{\partial R_{ij}} \operatorname{E}\left\{(f' - Rg')^H (f' - Rg')\right\} \\ &= \frac{\partial}{\partial R_{ij}} \operatorname{E}\left\{\sum_k \left(f'_k - \sum_l R_{kl} g'_l\right)^2\right\} \\ &= \operatorname{E}\left\{2\left(f'_i - \sum_l R_{il} g'_l\right) \cdot (-g'_j)\right\} \\ &= 2\left(\sum_l R_{il} \operatorname{E}\left\{g' g'^H\right\}_{lj} - \operatorname{E}\left\{f' g'^H\right\}_{ij}\right) \end{aligned} \tag{4.6}$$

Hence $R \operatorname{Var}[g] = \operatorname{Cov}(f, g)$ or, equivalently, $R = \operatorname{Cov}(f, g) \operatorname{Var}[g]^{-1}$. Substituting into (4.5) finally yields

$$\hat{f}_W = \mu_f + \operatorname{Cov}(f, g) \operatorname{Var}[g]^{-1} (g - \mu_g) \tag{4.7}$$

which is identical to (4.2). We may then further specify our finding by stating that $\hat{f}_W$ is the optimal (affine-)linear approximation of $\hat{f}_{\text{mmse}}$ with equality holding if the distributions are Gaussian and thus completely specified by second order statistics.

**Fourier-Domain Representation and Implementation** For the case of interest in this paper — estimating the inverse when the forward mapping takes the form of circular convolution

with the impulse response of an LSI system — the Wiener Filter has a particularly simple and computationally efficient representation in the Fourier domain.

Consider the following scenario. Let

$$f \sim \mathcal{N}(\mu_f, \Phi_f) \qquad \nu | f = g - Af \sim \mathcal{N}(0, \Phi_\nu) \tag{4.8}$$

both wide-sense-stationary with $A, \Phi_f, \Phi_\nu \in \mathbb{R}^{mn \times mn}$ BCCB, $m$ and $n$ denoting height and width of the image respectively. Let further $F = F_m \otimes F_n$ the two-dimensional DFT matrix as in (2.24). Then, by substituting into (4.7) and taking Fourier Transform

$$\mathcal{F} = \mathcal{S}\mathcal{A}^H(\mathcal{A}\mathcal{S}\mathcal{A}^H + \mathcal{N})^{-1}(\mathcal{G} - \mathcal{A}\mathcal{M}) + \mathcal{M} \tag{4.9}$$

where

$$\mathcal{S} = F\Phi_f F^H \qquad \mathcal{F} = F\hat{f}_W \tag{4.10a}$$
$$\mathcal{N} = F\Phi_\nu F^H \qquad \mathcal{G} = Fg \tag{4.10b}$$
$$\mathcal{A} = FAF^H \qquad \mathcal{M} = F\mu_f \tag{4.10c}$$

According to the convolution theorem 2.2 all Fourier transformed matrices in (4.9) are diagonal

$$\begin{aligned} \mathcal{C}_f &= \operatorname{diag}(F\Phi_f e_1) \\ \mathcal{C}_\nu &= \operatorname{diag}(F\Phi_\nu e_1) \\ \mathcal{A} &= \operatorname{diag}(FAe_1) \end{aligned} \tag{4.11}$$

so that (4.9) factorizes into a set of independent scalar operations. Rewriting the equation for any one component, using superscript indices and the asterisk ($*$) denoting complex conjugation, we find that

$$\begin{aligned} \mathcal{F}^{(i)} &= \frac{\mathcal{S}^{(i,i)}\mathcal{A}^{(i,i)*}}{\mathcal{S}^{(i,i)}|\mathcal{A}^{(i,i)}|^2 + \mathcal{N}^{(i,i)}} \left(\mathcal{G}^{(i)} - \mathcal{A}^{(i,i)}\mathcal{M}^{(i)}\right) + \mathcal{M}^{(i)} \\ &= \frac{|\mathcal{A}^{(i,i)}|^2}{|\mathcal{A}^{(i,i)}|^2 + \frac{\mathcal{N}^{(i,i)}}{\mathcal{S}^{(i,i)}}} \left(\frac{\mathcal{G}^{(i)}}{\mathcal{A}^{(i,i)}} - \mathcal{M}^{(i)}\right) + \mathcal{M}^{(i)} \end{aligned} \tag{4.12}$$

for $1 \leq i \leq mn$ and $\mathcal{S}(i,i) \neq 0$. Written as a product of two terms, one acting as a regularization coefficient while the other takes charge of the actual inversion, (4.12) closely resembles the Tikhonov estimate. Note, however, that the ratio $\mathcal{N}^{(i,i)}/\mathcal{S}^{(i,i)}$ and hence the amount of regularization is allowed to vary as a function of frequency here. Translating (4.12) into MATLAB-code a primitive form of the Wiener-Filter could be implemented as follows

```
function f = wiener (A, g, mu, Cf, Cnu)

%% A    discrete blur operator (BCCB)
%% g    blurred image
%% mu   mean of the prior distribution
%% Cf   variance of the Gaussian prior (BCCB)
%% Cnu  variance of the Gaussion noise (BCCB)

mn  = size(g);

H  = fft2(reshape(A(:,1), mn));
S  = fft2(reshape(Cf(:,1), mn));
N  = fft2(reshape(Cnu(:,1), mn));
G  = fft2(g);
M  = fft2(mu);

R  = S .* conj(H) ./ (S .* abs(H).^2 + FCnu);
F  = R .* (G - H .* M) + M;
f  = real(ifft2(F));
```

A more sophisticated implementation addressing issues of numerical stability can be found in the MATLAB image processing toolbox. Succinctness and clarity of code have been given priority over these considerations, as it befits the conceptual framework elaborated here.

**Choice of Parameters** A open question of practical interest is the choice of $(\mu_f, \Phi_f)$ comprising the prior distribution. Being rarely in the position to make an informed conjecture about the 'true' image, the best we can do here is a zero-mean prior $\mu_f = 0$ penalizing solutions with high energy. As for the remaining parameter which contributes indirectly to the ratio $\mathcal{N}^{(i,i)}/\mathcal{S}^{(i,i)}$ in (4.12) and effectively controls the amount of regularization for individual subspaces, consider the following argumentation. Let $X$ be zero-mean WSS with

$$\operatorname{Cov}(X_i, X_j) = R_{XX}(i-j) \tag{4.13}$$

(Without loss of generality we consider the one-dimensional case for simplicity.) An estimate of this quantity is given by the sample auto-correlation

$$R_{XX} \approx c\,\overline{R}_{XX} \qquad\qquad \overline{R}_{XX} := \sum_k X_k X_{k+i} \tag{4.14}$$

where $c$ is a normalizing constant — usually $1/(\dim(X) - 1)$ — and $X$ is assumed to be periodic in each dimension. By letting $\Phi_X := \operatorname{Var}[X]$ and taking Fourier-Transform on both sides we obtain

$$F\Phi_X e_1 \propto F\overline{R}_{XX} = |FX|^2 \tag{4.15}$$

where the absolute value is to be understood componentwise. The last equality uses a simple corollary of 2.2 known as Wiener-Khinchin theorem, stating that the (unnormalized) auto-correlation and the power spectrum density of the process are transform pairs. Using (4.15),

we return to the pending issue of sensibly choosing the regularization parameter and suggest the following heuristic

$$\frac{\mathcal{N}^{(i,i)}}{\mathcal{S}^{(i,i)}} \approx \frac{|(F\nu)_i|^2}{|(Ff)_i|^2} \approx \frac{\sigma_\nu^2}{|(Fg)_i|^2} \tag{4.16}$$

where the last approximation may be argued by taking expectation in the numerator

$$\mathrm{E}\left\{|(F\nu)_i|^2\right\} = \left(\mathrm{E}\left\{F\nu(F\nu)^H\right\}\right)_{ii} = \sigma_\nu^2 \tag{4.17}$$

and substituting the observable degraded image for the unknown 'true' one in the denominator. Admittedly, this latter replacement is somewhat gross and should only be used if no a-priori information is available.

**Complexity** The computational complexity is dominated by the asymptotic behaviour of the Fast Fourier Transform (FFT), which is $O(n \log n)$. This efficiency is quite remarkable and the main reason why the class of Wiener Filters — with all its numerous offsprings and variants — is still the method of choice for real-time applications.

## 4.2 Expectation Maximization

Although not by itself a deconvolution algorithm, Expectation Maximization (EM) is a useful technique to identify the ML estimate under degenerate conditions. Here is what we mean by it.

Maximum Likelihood estimation seeks to maximize the conditional probability $f(X|\theta)$ for a given observation $X$ over the parameter set $\Theta$. Exploiting positivity of probabilities, the problem often is considered in terms of log-likelihood

$$\theta_{ML} = \arg\max_{\theta\in\Theta} \log f_{X|\theta}(X|\theta) \tag{4.18}$$

where $X$ and $\theta$ are data and model-parameter respectively and $f_{X|\theta}$ is typically derived from a well-investigated physical model. (Usually the distribution of the forward mapping is peaked around the deterministic solution with a certain spread or variance allowing for measurement errors and limited accuracy.) In some cases, however, we can only observe a quantity $Y$ related indirectly to the complete data $X$ by a many-to-one mapping

$$Y = h(X) \tag{4.19}$$

for a non-injective $h$. EM-iterations are useful whenever the explicit derivation of the modified pdf $f_{Y|\theta}$ is too complicated or simply not feasible. They may be shown to converge to a local maximum of the pdf, allowing the identification of the ML estimate under conditions as in (4.19). This convenience, however, comes at the price of an often painfully slow performance.

In the following we outline the general concept before presenting two particular applications of EM in subsequent sections of this chapter. For a more comprehensive derivation of the algorithm and its illuminating recast as maximization of cross-entropy, see [21], [13], [18] and the references therein.

If an explicit representation of the unknown density $f_{Y|\theta}$ is not available we can start from the following observation

$$f(X|Y,\theta) = \frac{f(X,Y|\theta)}{f(Y|\theta)} \tag{4.20}$$

which relates the unknown pdf of the observable data $Y$ to the known distribution of the unobservable quantity $X$. Since $h$ is a deterministic mapping the joint pdf $f_{X,Y} = f_X$ is simply the density of the complete data. By simplifying (4.20) in this sense and taking the logarithm after solving for $f_{Y|\theta}$ we obtain

$$\log f(Y|\theta) = \log f(X|\theta) - \log f(X|Y,\theta) \tag{4.21}$$

Now let $\theta_p \in \Theta$ be arbitrary but fixed. Integrating with respect to the probability measure $f_{X|Y,\theta_p}$ yields

$$\begin{aligned} \int \log f(Y|\theta)\, f(X|Y,\theta_p)\, dX &= \int \big(\log f(X|\theta) - \log f(X|Y,\theta)\big)\, f(X|Y,\theta_p)\, dX \\ \log f(Y|\theta) &= \underbrace{\mathrm{E}\left\{\log f(X|\theta) \mid Y,\theta_p\right\}}_{=:\, Q(\theta|\theta_p)} - \underbrace{\mathrm{E}\left\{\log f(X|Y,\theta) \mid Y,\theta_p\right\}}_{=:\, H(\theta|\theta_p)} \end{aligned} \tag{4.22}$$

since the left-hand-side does not depend on $X$. Note that the result is a log-likelihood

$$\log f(Y|\theta) = Q(\theta|\theta_p) - H(\theta|\theta_p) \tag{4.23}$$

which for any given observation $Y$ can be maximized over the parameter $\theta$. This would be the ordinary proceeding of ML-estimation if the right hand side of (4.23) did not depend on the previously fixed $\theta_p$. If we take this to be the current guess for the Maximum Likelihood estimate $\theta_{ML}$, the following iterative scheme is obvious enough. The very idea of EM, in fact, is to get successively better approximations by letting

$$\begin{aligned} \theta_{p+1} &= \arg\max_{\theta\in\Theta} \log f(Y|\theta) \\ &= \arg\max_{\theta\in\Theta} \big(Q(\theta|\theta_p) - H(\theta|\theta_p)\big) \end{aligned} \tag{4.24}$$

the maximizer of (4.23) and repeat (4.24) until convergence is achieved. It turns out that this first draft can be greatly simplified by the following observations. In particular, we shall see that

$$Q(\theta_{p+1}|\theta_p) \geq Q(\theta_p|\theta_p) \Rightarrow \log f(Y|\theta_{p+1}) \geq \log f(Y|\theta_p) \tag{4.25}$$

In other words, for the sequence of Log-likelihoods $\log f(Y|\theta_p), p = 0,1,2,...$ to be monotonously non-decreasing it is sufficient that the left-hand-side of (4.25) holds true. To prove the implication, consider the update $\Delta^{(p)}$ at step $p \to p+1$

$$\begin{aligned} \Delta^{(p)} &= \log f(Y|\theta_{p+1}) - \log f(Y|\theta_p) \\ &= \big(Q(\theta_{p+1}|\theta_p) - Q(\theta_p|\theta_p)\big) + \big(H(\theta_p|\theta_p) - H(\theta_{p+1}|\theta_p)\big) \end{aligned} \tag{4.26}$$

We show that the second difference on the right-hand-side is non-negative. For this end consider Jensen's inequality, stating that $\varphi(\mathrm{E}\{x\}) \leq \mathrm{E}\{\varphi(x)\}$ for a convex function $\varphi$ and

both $x, \varphi(x) \in \mathcal{L}^1$. Since $\log'' x = -x^{-2} < 0$ the inequality can be applied for $\varphi(x) = -\log x$ to find that

$$\begin{aligned} H(\theta_p|\theta_p) - H(\theta_{p+1}|\theta_p) &= \mathrm{E}\left\{\log f(X|Y,\theta_p) - \log f(X|Y,\theta_{p+1}) \mid Y,\theta_p\right\} \\ &= -\int \log\left(\frac{f(X|Y,\theta_{p+1})}{f(X|Y,\theta_p)}\right) \cdot f(X|Y,\theta_p)\, dX \\ &\geq -\log \int \frac{f(X|Y,\theta_{p+1})}{f(X|Y,\theta_p)} \cdot f(X|Y,\theta_p)\, dX \\ &= -\log 1 \\ &= 0 \end{aligned} \tag{4.27}$$

as claimed above. Hence improvement $\Delta^{(p)} \geq 0$ is guaranteed whenever the condition $Q(\theta_{p+1}|\theta_p) \geq Q(\theta_p|\theta_p)$ is satisfied, ensuring that we move toward a local maximum of the pdf $f(Y|\theta)$. This finding renders the update (4.24) a lot easier, since we only need to concentrate on maximizing $Q_p(\theta) := Q(\theta|\theta_p)$.

In pseudo-code the EM algorithm then reads as follows:

- **Set starting value**
  Let $\theta_0 \in \Theta$ arbitrary.
- **Iterate**
  1. **E-step**:
     Compute $Q_p(\theta) = \mathrm{E}\left\{\log f(X|\theta) \mid Y,\theta_p\right\}$
  2. **M-step**:
     Set $\theta_{p+1} = \arg\max_{\theta\in\Theta} Q_p(\theta)$

  while $\theta_{p+1} \neq \theta_p$ repeat steps 1-2 with $p = p+1$

If the pdf is multimodal, some care must be taken when choosing the starting value, since EM converges only to a local maximum.

**Exponential Family** Another important simplification can be achieved whenever the pdf $f(X|\theta)$ — or pmf in case of a discrete random variable — is a member of the so-called exponential family which have the representation

$$f(X|\theta) = \frac{b(X) \cdot \exp\left(c(\theta)^T t(X)\right)}{a(\theta)} \tag{4.28}$$

The majority of the commonly used distributions fall into that category, including Gaussian and Poissonian which shall be considered in some detail hereafter.

Factorizing the pdf in quantities that depend exclusively on either $X$ or $\theta$ greatly simplifies the proceeding, for in either of the alternating steps we only have to consider the components relevant to the respective stage. The E-step exclusively involves quantities depending on $X$

which in turn may be ignored during subsequent maximization over $\theta$ in the M-step. By substituting (4.28) into the defining equation we obtain

$$\begin{aligned} Q_p(\theta) &= \mathrm{E}\left\{\log f(X|\theta) \mid Y, \theta_p\right\} \\ &= \mathrm{E}\left\{\log b(X) \mid Y, \theta_p\right\} + c(\theta)^T \mathrm{E}\left\{t(X) \mid Y, \theta_p\right\} - \log a(\theta) \end{aligned} \tag{4.29}$$

Better still, the first term on the right-hand-side, being independent of $\theta$, is inconsequential for the M-step. In the E-step, thus, it is sufficient to compute

$$t_{p+1}(X) = \mathrm{E}\left\{t(X) \mid Y, \theta_p\right\} \tag{4.30}$$

while the M-step simply reduces to

$$\theta_{p+1} = \arg\max_{\theta \in \Theta} c(\theta)^T t_{p+1} - \log a(\theta) \tag{4.31}$$

We shall make use of this simplification throughout the following sections, effectively replacing the two steps in the diagram by (4.30) and (4.31).

## 4.3 Richardson-Lucy

This algorithm, originally developed by Richardson and later refined by Lucy, is one of the few to actually take into account a Poissonian distribution of the noise. It is being successfully used in the fields of astronomical imagery. Its popularity and renown stems to a great extend from its much acclaimed application to Hubble Space Telescope (HST) data, where to the common appraisal it did a terrific job in restoring the blurred images. Richardson-Lucy has become a prime example of how astute mathematical processing of the data — and thus comparably inexpensive software — can make up, at least within certain limits, for hardware deficiency as was the case with the initially flawed optics of the HST.

The Richardson-Lucy algorithm can be viewed as a first example of EM iterations that converge to the ML-estimate of the pristine image when the model is a linear mapping with Poisson-distributed intensities.

**Derivation** We shall need the following lemma which by induction readily extends to any finite number of variables. To resolve ambiguities, the notation $p(\mathrm{var} = \mathrm{val})$ is used occasionally in order to discriminate clearly between the random variable in question and its value.

**Lemma.** *Let $X \sim P_o(\lambda_x)$ and $Y \sim P_o(\lambda_y)$ be two independent Poissonian random variables. Then the sum $Z := X + Y$ is also Poisson with distribution $Z \sim P_o(\lambda_x + \lambda_y)$.*

**Proof:**

$$\begin{aligned} p(Z=N) &= \sum_{K=0}^{N} p(X=K)\cdot p(Y=N-K) \\ &= \sum_{K=0}^{N} e^{-\lambda_x}\frac{\lambda_x^K}{K!}\cdot e^{-\lambda_y}\frac{\lambda_y^{N-K}}{(N-K)!} \\ &= e^{-(\lambda_x+\lambda_y)}\frac{1}{N!}\sum_{K=0}^{N}\binom{N}{K}\lambda_x^K\,\lambda_y^{N-K} \\ &= e^{-(\lambda_x+\lambda_y)}\frac{(\lambda_x+\lambda_y)^N}{N!} \end{aligned} \tag{4.32}$$

■

We now introduce the complete data as follows. Let $X = \{X_{ij}\}_{i,j=1\ldots mn}$ an (unobservable) random variable with $X_{ij}$ the number of photons emitted at location $j$ in the original scene and detected at location $i$ by the recording system. Assume further that all components are independently Poisson with distribution $X_{ij} \sim P_o(A_{ij}\cdot f_j)$.

$$p(X|f) = \prod_{i,j=1}^{mn}\exp(-A_{ij}f_j)\frac{(A_{ij}f_j)^{X_{ij}}}{X_{ij}!} \tag{4.33}$$

$$= \exp(-1^T Af)\cdot\prod_{ij=1}^{mn}\frac{(A_{ij}f_j)^{X_{ij}}}{X_{ij}!} \tag{4.34}$$

Then the intensities of the blurred and pristine image are given by the row-sums and column-sums of $X$ respectively

$$g = X1 \in \mathbb{R}^{mn} \qquad\qquad f = 1^T X \in \mathbb{R}^{mn} \tag{4.35}$$

Note that the complete data $X$ includes both the observation $g$ and the parameter $f$ which for convenience only are assumed to have identical size. There is no need that $X$ be square, and the algorithm readily extends to a scenario of different resolution.

Distribution (4.33) is a member of the exponential family, factorizing into

$$p(X|f) = \frac{b(X)\exp\left(c(f)^T t(X)\right)}{a(f)} \tag{4.36}$$

with

$$t(X) = \operatorname{vec} X \qquad a(f) = \exp(1^T Af) \tag{4.37a}$$

$$b(X) = \Big(\prod_{i,j} X_{ij}!\Big)^{-1} \qquad c(f) = \operatorname{vec} C \quad C_{ij} = \log A_{ij}f_j \tag{4.37b}$$

According to (4.30) the E-step consists in computing $t_{p+1} = \mathrm{E}\left\{t(X)|g, f^{(p)}\right\} = \operatorname{vec}\mathrm{E}\left\{X|g, f^{(p)}\right\}$. Since all components $X_{ij}$ are assumed to be independent we have

$$\mathrm{E}\left\{X \mid g, f^{(p)}\right\}_{ij} = \mathrm{E}\left\{X_{ij} \mid g, f^{(p)}\right\} = \mathrm{E}\left\{X_{ij} \mid g_i, f^{(p)}\right\} \tag{4.38}$$

for $1 \leq i, j \leq mn$. To evaluate the conditional expectation, consider that the corresponding pmf is

$$p(X_{ij} = M | g_i = N, f^{(p)}) = \frac{p(g_i = M | X_{ij}, f^{(p)}) \cdot p(X_{ij} = M | f^{(p)})}{p(g_i = N | f^{(p)})} \tag{4.39}$$

according to Bayes rule. By applying the lemma (4.32) on the first term in the numerator we obtain

$$\begin{aligned} p(g_i = M \mid X_{ij}, f^{(p)}) &= p\big(\sum_{k \neq j} X_{ik} = N - M | f^{(p)}\big) \\ &= \exp\big(-\sum_{k \neq j} A_{ik} f_k^{(p)}\big) \cdot \frac{\big(\sum_{k \neq j} A_{ik} f_k^{(p)}\big)^{N-M}}{(N-M)!} \end{aligned} \tag{4.40}$$

Proceeding in the same way for the remaining terms in (4.39) yields an explicit representation for the conditional pmf

$$\begin{aligned} p(X_{ij} = M | g_i = N, f^{(p)}) &= \frac{e^{-\sum_{k \neq j} A_{ik} f_k^{(p)}} \cdot \big(\sum_{k \neq j} A_{ik} f_k^{(p)}\big)^{N-M} \cdot e^{-A_{ij} f_j^{(p)}} \cdot \big(A_{ij} f_j^{(p)}\big)^M \cdot N!}{(N-M)! \cdot M! \cdot e^{-\sum_k A_{ik} f_k^{(p)}} \cdot \big(\sum_k A_{ik} f_k^{(p)}\big)^N} \\ &= \binom{N}{M} \cdot \frac{\big((Af^{(p)})_i - A_{ij} f_j^{(p)}\big)^{N-M} \cdot \big(A_{ij} f_j^{(p)}\big)^M}{\big((Af^{(p)})_i\big)^N} \\ &= \binom{N}{M} \cdot q^M \cdot (1-q)^{N-M} \end{aligned}$$

with $q = A_{ij} f_j^{(p)} / (Af^{(p)})_i = \mathrm{E}\{X_{ij}\} / \mathrm{E}\{g_i\}$. Obviously this is the pmf of a binomial distribution, with $M$ the number of positive outcomes out of a total of $N$ Bernoulli-trials, each successful with probability $q$. According to a well-known rule the expectation of a binomial random variable $M \sim Bin(N, q)$ is simply given by $\mathrm{E}\{M\} = Nq$ or the number of trials multiplied with the probability of success. Thus

$$\begin{aligned} \mathrm{E}\left\{X_{ij} \mid g_i = N, f^{(p)}\right\} &= \sum_{M=0}^{N} M \cdot p(X_{ij} = M \mid g_i = N f^{(p)}) \\ &= N \cdot \frac{A_{ij} f_j^{(p)}}{(Af^{(p)})_i} \end{aligned} \tag{4.41}$$

Vertically stacking the components finally yields $t_{p+1} = \mathrm{vec}\,\mathrm{E}\left\{X \mid g, f^{(p)}\right\}$. We now take a closer look at the M-step of the algorithm which requires the maximization of

$$Q^{(p)}(f) = c(f)^T t_{p+1} - \log a(f) \tag{4.42}$$

Differentiating with respect to each component of $f$ and equating the partial derivatives to zero yields an explicit representation of $f^{(p+1)} = \arg\max Q^{(p)}(f)$. As necessary condition we

thus obtain

$$\begin{aligned} 0 = \frac{\partial}{\partial f_k} Q^{(p)}(f) &= \frac{\partial}{\partial f_k} \left[ \sum_{i,j=0}^{mn} \log A_{ij} f_j \cdot \mathrm{E}\left\{ X_{ij} \mid g_i, f^{(p)} \right\} - 1^T A f \right] \\ &= \sum_{i=0}^{mn} \left( \frac{1}{A_{ik} f_k} A_{ik} \cdot \mathrm{E}\left\{ X_{ik} \mid g_i, f^{(p)} \right\} \right) - \sum_{i=0}^{mn} A_{ik} \\ &= \frac{f_k^{(p)}}{f_k} \sum_{i=0}^{mn} \left( g_i \cdot \frac{A_{ik}}{(Af^{(p)})_i} \right) - \sum_{i=0}^{mn} A_{ik} \end{aligned} \tag{4.43}$$

Solving for $f_k$ finally yields the new estimate

$$f_k^{(p+1)} = \frac{f_k^{(p)}}{1^T A} \cdot \sum_{i=1}^{mn} g_i \cdot \frac{A_{ik}}{(Af^{(p)})_i} \tag{4.44}$$

for $1 \leq k \leq mn$. If we define dot and bar to be element-wise multiplication and division respectively, the update rule may be expressed compactly as

$$f^{(p+1)} = \frac{f^{(p)}}{1^T A} \cdot A^T \left( \frac{g}{Af^{(p)}} \right) \tag{4.45}$$

**Implementation** As before, we give the bare skeleton of an implementation in MATLAB-code, reduced to the essential. For additional feature see the image processing toolbox.

```
function [f, delta, k] = RL(A, g, f, iters)
%% A      discrete blur operator (BCCB)
%% g      blurred image
%% f      start value for estimate
%% iters  maximum number of steps

[m,n] = size(g);
h = fft2(A(:,1));
c = real(ifft2(conj(h) .* fft2(ones(m, n))));

k = 0;
delta = 1;
while k < iters && delta > 1e-3
    z = g ./ real(ifft2(h .* fft2(f)));
    fnew = f ./ c .* real(ifft2(conj(h) .* fft2(z)));
    delta = (fnew - f) ./ f;
    delta = max(delta(:));
    f = fnew;
    k = k + 1;
end
```

**Complexity and Convergence Rate** Assessing the convergence properties of EM-based algorithms under sufficiently general conditions proves to be extremely difficult. To our best knowledge, there are no reasonably tight bounds on the rate of convergence without reverting to assumptions more or less peculiar to a certain domain of application.

The few known endeavors in this respect [26], [4], both fairly recent publications, assume that the 'true' image is made up of Gaussian shaped light sources which restricts their findings more or less to astronomical imagery. True enough, this is where Richardson-Lucy is being primarily applied, but it does not make for our case. Lacking a formal quantification, then, we have to fall back on experimental observations. Compared to the direct approach of, say, a Wiener filter, EM-iterations are often painfully slow. [1] have proposed a numerical technique for accelerating the procedure which, by their claim, boasts an average speed-up factor of 40 in the long run (for restorations iterating more than 250 cycles). Being rather technical and not specific to the Richardson-Lucy algorithm, we omit its presentation as not pertinent to the conceptual framework elaborated here.

When it comes to convergence rates, it is important to bear in mind another aspect. As a matter of fact, ultimate convergence, very often, is not even desirable. Richardson-Lucy is an example of non-Bayesian ML estimation, which normally leads to data-overfitting and noise amplification, if no additional constraint is applied. This effect is less dramatic in an iterative approach, where a premature halt of the algorithm — that is, before convergence is achieved — can act as a form of regularization. The optimal moment for termination, however, proves difficult to determine by an objective criterion. The question when to stop the iterations, analogue to parameter rules in direct approaches, is much less well-investigated and has not yet been satisfactorily answered.

## 4.4 Blind Deconvolution

What makes EM so powerful is its ability to cope with 'hidden' information by introducing the notion of the unobservable complete data. Basically this is another stage of indirection in the model, allowing for a degenerate relation between the observable quantity and the pdf to be maximized. It stands to reason that this principle can be usefully applied to tackle the problem of blind deconvolution. Due to its reputed and factual difficulty, this variant is less well investigated than its non-blind counterpart — regrettably so, for the point-spread-function is rarely known in advance. Even with the optical device at hand, which is already a privileged situation, making this information available is not a trivial matter. Lately, however, blind deconvolution has received growing attention, reflected by publications such as [15], [16], [22] and others.

One of the more promising proposals in this field of research goes back to Katsaggelos [14, 13] who has shown that EM naturally expands to blind deconvolution by including the PSF among the parameters to be estimated. In this section we review the derivation of his algorithms or, more precisely, a variant that strikes us as particularly useful.

**Derivation** With the presently discussed algorithm we return to a Bayesian framework of parameter estimation where regularization takes the form of prior probability. Also, to keep the problem tractable in spite of the innate difficulty of blind deconvolution we stick with the

comparably simple Gaussian models for prior and noise

$$f \sim \mathcal{N}(0, \Lambda_f) \qquad\qquad g|f \sim \mathcal{N}(Af, \Lambda_\nu) \tag{4.46}$$

where $f, g \in R^{mn}$, as usual, denote pristine and blurred image respectively, both of size $m \times n$. In contrast to the models considered so far, however, both of the above densities are implicitely understood as conditioned upon the set of parameters

$$\theta = (\Lambda_f, \Lambda_\nu, A) \tag{4.47}$$

Again, it is assumed that all matrices involved are BCCB to facilitate computation later on by discrete Fourier-transforms. Somewhat surprisingly perhaps, $f$ does not show up as one of the parameters to be estimated, as one would expect of the quantity we are most interested in, after all. Instead, it constitutes the unobservable part of the complete data $x$, whose ML-estimate will be computed as a by-product during the E-step. Hence the observable information $g$ relates to the complete data by the non-injective mapping

$$g = (0^T \quad 1^T) \cdot x \qquad\qquad x = \begin{pmatrix} f \\ g \end{pmatrix} \tag{4.48}$$

which is simply the projection on the second $mn$ components of $x$. As for the distribution of the complete data, we find that since both $f \sim \mathcal{N}(0, \Lambda_f)$ and $g \sim \mathcal{N}(0, A\Lambda_f A^H + \Lambda_\nu)$ are zero-mean Gaussian, their joint pdf

$$p(x|\theta) = \frac{1}{\sqrt{(2\pi)^{2mn}|\Lambda_x|}} \cdot \exp(-\frac{1}{2} x^H \Lambda_x^{-1} x) \tag{4.49}$$

must also be Gaussian with mean $\mu_x = \mathrm{E}\{x\} = 0$. On the other hand, it clearly holds that

$$\begin{aligned} p(x|\theta) &= p(g|f,\theta) \cdot p(f|\theta) \\ &= \frac{1}{\sqrt{(2\pi)^{2mn}|\Lambda_\nu||\Lambda_f|}} \exp\left[-\frac{1}{2}\big((g-Af)^H \Lambda_\nu^{-1}(g-Af) + f^H \Lambda_f^{-1} f\big)\right] \end{aligned} \tag{4.50}$$

By equating (4.49) and (4.50) an explicit representation of $\Lambda_x$ can be derived. Note that the exponential is a positive definite quadratic form in $x$ which, by sorting for the quadratics in either of the components and a mixed term is given by

$$\begin{aligned} x^H \Lambda_x^{-1} x &= (g-Af)^H \Lambda_\nu^{-1}(g-Af) + f^H \Lambda_f^{-1} f \\ &= f^H B_{11} f + f^H (B_{12} + B_{21}^H) g + g^H B_{22} g \end{aligned} \tag{4.51}$$

with

$$\begin{aligned} B_{11} &= A^H \Lambda_\nu^{-1} A + \Lambda_f^{-1} \\ B_{12} = B_{21}^H &= -A^H \Lambda_\nu^{-1} \\ B_{22} &= A^H \Lambda_\nu^{-1} \end{aligned} \tag{4.52}$$

Identifying the $B_{ij}$ for $i, j \in \{1, 2\}$ with the blocks of the partitioned inverse variance matrix finally yields

$$\Lambda_x^{-1} = \begin{pmatrix} A^H \Lambda_\nu^{-1} A + \Lambda_f^{-1} & -A^H \Lambda_\nu^{-1} \\ -\Lambda_\nu^{-1} A & \Lambda_\nu^{-1} \end{pmatrix} \in \mathbb{R}^{2mn \times 2mn} \tag{4.53}$$

We can simplify the EM iteration by the shortcuts (4.30) and (4.31) for densities of the exponential family, since

$$p(x|\theta) = \frac{b(x) \cdot \exp(c(\theta)^T t(x))}{a(\theta)} \tag{4.54}$$

with

$$b(x) = 1 \qquad c(\theta) = -\frac{1}{2} \operatorname{vec} \Lambda_x^{-1} \tag{4.55a}$$

$$t(x) = \operatorname{vec} xx^H \qquad a(\theta) = (2\pi)^{mn} \sqrt{|\Lambda_\nu||\Lambda_f|} \tag{4.55b}$$

So, again, we have

$$Q_p(\theta) = c(\theta)^T t_{p+1} - \log a(\theta) \tag{4.56}$$

this time, however, with

$$\begin{aligned} c(\theta)^T t_{p+1} &= -\frac{1}{2} \sum_{i=1}^{2mn} \sum_{j=1}^{2mn} (\Lambda_x^{-1})_{ij} \cdot \mathrm{E}\left\{xx^H \mid g, \theta\right\}_{ji} \\ &= -\frac{1}{2} \sum_{i=1}^{2mn} \left(\Lambda_x^{-1} \mathrm{E}\left\{xx^H \mid g, \theta\right\}\right)_{ii} \\ &= -\frac{1}{2} \operatorname{trace}\left(\Lambda_x^{-1} \mathrm{E}\left\{xx^H \mid g, \theta\right\}\right) \end{aligned} \tag{4.57}$$

and

$$-\log a(\theta) = -mn \log 2\pi - \frac{1}{2} \log|\Lambda_\nu| - \frac{1}{2} \log|\Lambda_f| \tag{4.58}$$

Ignoring the constant term irrelevant for the maximization, the M-step during which parameters are actually updated, looks as follows

$$\begin{aligned} \theta_{p+1} &= \arg\max_\theta Q_p(\theta) \\ &= \arg\min_\theta \left\{\operatorname{trace}\left(\Lambda_x^{-1} \mathrm{E}\left\{xx^H \mid g, \theta\right\}\right) + \log|\Lambda_\nu| + \log|\Lambda_f|\right\} \end{aligned} \tag{4.59}$$

Now let $\mu_{f|g}^{(p)} = \mathrm{E}\left\{f \mid g, \theta_p\right\}$ denote the conditional mean of the pristine image which is also the approximation of the sought ML-estimate at step $p$. Considering the first term of the right hand side in (4.59) and using (4.53) we find that

$$\underbrace{\begin{pmatrix} A^H\Lambda_\nu^{-1}A + \Lambda_f^{-1} & -A^H\Lambda_\nu^{-1} \\ -\Lambda_\nu^{-1}A & \Lambda_\nu^{-1} \end{pmatrix}}_{\Lambda_x^{-1}} \cdot \underbrace{\begin{pmatrix} \Lambda_{f|g}^{(p)} + \mu_{f|g}^{(p)}\mu_{f|g}^{(p)H} & \mu_{f|g}^{(p)H} g \\ g\mu_{f|g}^{(p)H} & gg^H \end{pmatrix}}_{\mathrm{E}\left\{xx^H \mid g, \theta\right\}}$$

$$= \begin{pmatrix} (\Lambda_f^{-1} + A^H\Lambda_\nu^{-1}A)(\Lambda_{f|g}^{(p)} + \mu_{f|g}^{(p)}\mu_{f|g}^{(p)H}) - A^H\Lambda_\nu^{-1}g\mu_{f|g}^{(p)} & * \\ * & -\Lambda_\nu^{-1}(A\mu_{f|g}^{(p)H} g - gg^H) \end{pmatrix}$$

since we are only interested in the diagonal entries. Applying the trace operator and using the identity $\operatorname{trace}(Sxy^H) = y^H Sx$, the function to be minimized becomes

$$\begin{aligned} Q'_p(\theta) = &\operatorname{trace}\left((\Lambda_f^{-1} + A^H\Lambda_\nu^{-1}A)\cdot\Lambda_{f|g}^{(p)}\right) + \log|\Lambda_\nu| + \log|\Lambda_f| \\ &+ \mu_{f|g}^{(p)H}(\Lambda_f^{-1} + A^H\Lambda_\nu^{-1}A)\mu_{f|g}^{(p)} - 2g^H A\Lambda_\nu^{-1}\mu_{f|g}^{(p)} + g^H\Lambda_\nu^{-1}g \end{aligned} \tag{4.60}$$

The conditional mean is constructed according to theorem (3.57) using the current estimates

$$\begin{aligned} \mu_{f|g}^{(p)} &= \operatorname{Cov}(f,g)\operatorname{Var}[g]^{-1}g \\ &= \Lambda_f A^{(p)H}(A^{(p)}\Lambda_f^{(p)}A^{(p)H} + \Lambda_\nu^{(p)})^{-1}g \end{aligned} \tag{4.61}$$

As for the conditional variance, substituting into (3.58) yields

$$\begin{aligned} \Lambda_{f|g}^{(p)} &= \operatorname{Var}[f] - \operatorname{Cov}(f,g)\operatorname{Var}[g]^{-1}\operatorname{Cov}(g,f) \\ &= \Lambda_f^{(p)} - \Lambda_f A^{(p)H}(A^{(p)}\Lambda_f A^{(p)H} + \Lambda_\nu^{(p)})^{-1}A^{(p)}\Lambda_f^{(p)} \end{aligned} \tag{4.62}$$

As usual, blur, prior and noise have been assumed to be shift-invariant — or wide-sense-stationary, as far as random processes are concerned —, so that all matrices are BCCB and can be diagonalized relative to the same basis. For a vector $v$ define $\widehat{v} := Fv$ its Fourier-transform and similarly $\widehat{M} := FMF^H$ for a matrix $M$. Then

$$\widehat{A} = \operatorname{diag}(\alpha) \qquad \alpha = FAe_1 \tag{4.63a}$$

$$\widehat{\Lambda}_f = \operatorname{diag}(\lambda_f) \qquad \lambda_f = F\Lambda_f e_1 \tag{4.63b}$$

In the following, to avoid further complicating (4.64) by additional indices, we shall understand multiplication and division to be defined componentwise. Assuming AWGN noise with standard deviation $\sigma_\nu$ we have, in current notation

$$\widehat{\mu}_{f|g}^{(p)} = \frac{\alpha^{(p)*}\cdot\lambda_f^{(p)}}{|\alpha^{(p)}|^2\cdot\lambda_f^{(p)} + \sigma_\nu^{2(p)}}\cdot\widehat{g} \tag{4.64}$$

Note that this is the Fourier-domain representation of the Wiener filter already derived in (4.12). Applying a similar transformation to the conditional variance yields $\widehat{\Lambda}_{f|g}^{(p)} = \operatorname{diag}(\lambda_{f|g}^{(p)})$ with

$$\lambda_{f|g}^{(p)} = \frac{\lambda_f^{(p)}\cdot\sigma_\nu^{2(p)}}{|\alpha^{(p)}|^2\cdot\lambda_f^{(p)} + \sigma_\nu^{2(p)}} \tag{4.65}$$

Exploiting invariance of the determinant and trace operators under orthogonal transformation, (4.60) can be written as

$$\begin{aligned} Q'_p(\theta) = &\operatorname{trace}\left(\widehat{\Lambda}_f^{-1} + \widehat{A}^H\widehat{\Lambda}_\nu^{-1}\widehat{A})\cdot\widehat{\Lambda}_{f|g}^{(p)}\right) + \log|\widehat{\Lambda}_\nu| + \log|\widehat{\Lambda}_f| \\ &+ \widehat{\mu}_{f|g}^{(p)H}(\widehat{\Lambda}_f^{-1} + \widehat{A}^H\widehat{\Lambda}_\nu^{-1}\widehat{A})\widehat{\mu}_{f|g}^{(p)} - 2\Re\left(\widehat{g}^H\widehat{A}\widehat{\Lambda}_\nu^{-1}\widehat{\mu}_{f|g}^{(p)}\right) + \widehat{g}^H\widehat{\Lambda}_\nu^{-1}\widehat{g} \end{aligned} \tag{4.66}$$

Since all matrices are diagonal, we may write out (4.66) as a sum of scalar operations

$$Q'_p(\theta) = \sum_{i=1}^{mn} \left[ \left( \frac{1}{\lambda_f(i)} + \frac{|\alpha(i)|^2}{\sigma_\nu^{2(p)}} \right) \cdot \lambda_{f|g}^{(p)}(i) + \left( \frac{1}{\lambda_f(i)} + \frac{|\alpha(i)|^2}{\sigma_\nu^{2(p)}} \right) \cdot |\widehat{\mu}_{f|g}^{(p)}(i)|^2 - \frac{2\Re\left(\widehat{g}(i)^* \cdot \alpha(i) \cdot \widehat{\mu}_{f|g}^{(p)}(i)\right) + |\widehat{g}(i)|^2}{\sigma_\nu^{2(p)}} + \log \lambda_f^{(p)}(i) + \log \sigma_\nu^{2(p)} \right] \tag{4.67}$$

Sorting the terms and minor simplifications finally yield the modified cost function

$$Q'_p(\theta) = mn \cdot \log \sigma_\nu^{2(p)} + \sum_{i=1}^{mn} \left[ \frac{1}{\lambda_f(i)} \cdot \left( \lambda_{f|g}^{(p)}(i) + |\widehat{\mu}_{f|g}^{(p)}(i)|^2 \right) + \log \lambda_f^{(p)}(i) \right] + \frac{1}{\sigma_\nu^{2(p)}} \sum_{i=1}^{mn} \left[ |\alpha(i)|^2 \left( \lambda_{f|g}^{(p)}(i) + |\widehat{\mu}_{f|g}^{(p)}(i)|^2 \right) - 2\Re\left( \widehat{g}(i)^* \, \alpha(i) \, \widehat{\mu}_{f|g}^{(p)}(i) \right) + |\widehat{g}(i)|^2 \right] \tag{4.68}$$

which has to be minimized during the M-step in order to obtain the updated estimates

$$\theta_{p+1} = (\Lambda_f^{(p+1)}, \Lambda_\nu^{(p+1)}, A^{(p+1)}) = \arg\min Q'_p(\theta) \tag{4.69}$$

for the next iteration. Proceeding as usual, we set

$$\nabla_\theta Q'_p(\theta_{p+1}) = \begin{pmatrix} \nabla_\alpha Q'_p \\ \nabla_{\lambda_f} Q'_p \\ \nabla_{\sigma_\nu^2} Q'_p \end{pmatrix} (\theta_{p+1}) = 0 \tag{4.70}$$

By solving for the parameters of interest we get

$$\lambda_f^{(p+1)} = \lambda_{f|g}^{(p)} + |\hat{\mu}_{f|g}^{(p)}|^2 \qquad\qquad \alpha^{(p+1)} = \frac{\hat{g}^* \hat{\mu}_{f|g}^{(p)}}{\lambda_{f|g}^{(p)} + |\hat{\mu}_{f|g}^{(p)}|^2} \tag{4.71}$$

as new estimates of the 'true' image and the spectrum of the blur kernel. As to the noise-variance, we find that

$$\sigma_\nu^{2(p+1)} = \sum_{i=1}^{mn} \left[ |\alpha^{(p)}|^2 \left( \lambda_{f|g}^{(p)} + |\hat{\mu}_{f|g}^{(p)}| \right) + |\hat{g}|^2 - 2\Re\left( \hat{g}^* \alpha^{(p)} \hat{\mu}_{f|g}^{(p)} \right) \right] \tag{4.72}$$

(Again, operations are to be taken componentwise). Equations (4.71) and (4.72) constitute the body of EM iterations.

**Non-uniqueness of the PSF** Some comment is indicated concerning the uniqueness of the recovered PSF. From (4.71) is seems as if the algorithm was estimating amplitude *and* phase information of the PSF, the latter one represented by the imaginary part of the eigenvalues $\alpha$. However, this proves an unrealistic expectation the algorithm cannot live up to. Here is why. Clearly the observed image has a pdf proportional to

$$p(g|\theta) \propto \exp(-g^H \, \mathrm{Var}\,[g \,|\, \theta]^{-1} \, g) \qquad\qquad \mathrm{Var}\,[g \,|\, \theta] = A\Lambda_f A^H + \Lambda_\nu \tag{4.73}$$

but the quadratic in $g$ has the Fourier-domain representation

$$g^H(A\Lambda_f A^H + \Lambda_\nu)^{-1} g = \frac{|\widehat{g}|^2}{|\alpha|^2 \lambda_f + \sigma_\nu^2} \tag{4.74}$$

Note, then, that the spectrum contributes only as $|\alpha|^2$ to the pdf whose maximization we undertake. In other words, all eigenvalues of equal magnitude have the same likelihood. One way to resolve this ambiguity is by imposing additional constraints on the PSF. Usually a zero-phase is assumed, resulting in a symmetric PSF. By setting $\alpha \in \mathbb{R}^{mn}$ we also coerce symmetry of the matrix $A$, since $A = F^H \Delta F = F^H \Delta^* F = A^H$ with $\Delta = \mathrm{diag}\,(\alpha)$. Likewise, a normalization constraint $\alpha_1 = 1/\sqrt{mn}$ will have the weights $h_i$ satisfy $\sum_i h_i = 1$. Apart from preserving the signal's energy it can further help establish uniqueness of the recovered PSF. In practice, this might be realized by running the algorithm for some time, stop to normalize the PSF and then start a new cycle. Our experiments with this technique, however, were not encouraging.

**Convergence Rate** As to performance, we refer to the already lamented absence of convergence rates for EM-based methods under sufficiently general conditions. Although the algorithm works exclusively in the Fourier-domain and does not need to constantly switch between spatial and frequency representation, in all tests it has proved significantly slower than Richardson-Lucy.

**Implementation** As for the previous algorithms, we implement the basic steps in MATLAB-code, given in the listing below

```
function [f, alpha, delta, k] = blindEM(g, alpha, lambda, sigma2)

%% g blurred image
%% alpha   eigenvalues of convolution operator A (OTF)
%% lambda  eigenvalues of prior variance
%% sigma2  noise variance (scalar)
%% eta     eigenvalues of conditional variance f given g

mn = numel(g);
G  = fft2(g);
maxiters = 80;

for k = 1 : maxiters
  %% E-step: compute conditional mean and variance
  denom = abs(alpha).^2 .* lambda + sigma2;
  F = conj(alpha) .* lambda ./ denom .* G;
  eta = lambda * sigma2 ./ denom;

  %% save old values
  lambda_old = lambda;
  sigma2_old = sigma2;
  alpha_old = alpha;

  %% M-step: update estimates
  lambda = eta + abs(F).^2 / mn;
  alpha = G .* conj(F) ./ (lambda * mn);
  tmp = abs(alpha).^2 .* lambda;
  tmp = tmp + (abs(G).^2 - 2*real(conj(G) .* alpha .* F)) / mn;
  sigma2 = sum(tmp(:)) / mn;

  %% check for convergence
  delta(1) = max(abs(lambda(:) - lambda_old(:)) ./ abs(lambda(:)));
  delta(2) = max(abs((alpha(:) - alpha_old(:)) ./ abs(alpha(:)));
  delta(3) = max(abs(sigma2 - sigma2_old) / abs(sigma2_old));

  if max(delta) < 1e-3
    break;
  end
end

f = real(ifft2(F));
```

## 4.5 Neelamani (ForWaRD)

With the last algorithm selected for presentation, also the most recent of date, we may say to come full circle. Not only in the wider sense of a return to non-blind and direct approaches; the tie-up with the concept of linear filters is indeed almost literal. Seemingly rather traditional at first sight, the algorithm may be said to realize a genuine contribution by generalizing the concept to other transform domains. (Hence the fancy camel-case acronym, apparently a contraction of Fourier-Wavelet Regularized Deconvolution. In honour of its inventor [19] as much as for simplicity, however, we refer to it as Neelamani)

**Motivation and Idea** The dilemma of linear filters is the intricate mingling of noise and signal. If we choose to shrink the contributions of a particular subspace by a factor in $r_\alpha \ll 1$ we indiscriminately extenuate or suppress both noise and valid information. Likewise, if we set $r_\alpha \approx 1$ we are sure to retain good part of the signal — but also most of the unwished-for noise. This observation may be formalized as follows.

Recall that the mean (integrated) square error can be decomposed into the squared 2-norm of the bias plus the variances of the leaked, because insufficiently attenuated, noise. For an arbitrary filter function $r_\alpha \in [0,1]$ acting upon the inverted spectrum of the blur operator we get, by substituting into equation (3.1)

$$\mathrm{MISE}\left[\hat{f}\right] = \left\|(I - A_\alpha^\dagger A)f\right\|_2^2 + \mathrm{trace}\left(A_\alpha^\dagger \Phi_\nu A_\alpha^{\dagger H}\right) \tag{4.75}$$

This representation can be used to derive a lower bound on the MISE independent from the choice of filter coefficients. In fact, for any singular value $\sigma$ in the spectrum of $A$ with associated right singular vector $v$ it holds that

$$(1 - r_\alpha(\sigma))^2 \left|\langle v, f\rangle\right|^2 + r_\alpha(\sigma)^2 \frac{\mathrm{E}\left\{\left|\langle v, \nu\rangle\right|^2\right\}}{\sigma^2} \geq \frac{1}{2} \min\left\{\left|\langle v, f\rangle\right|^2, \frac{\mathrm{E}\left\{\left|\langle v, \nu\rangle\right|^2\right\}}{\sigma^2}\right\} \tag{4.76}$$

since $(1 - r_\alpha(\sigma))^2 + r_\alpha(\sigma)^2 \geq \frac{1}{2}$. By taking the sum over all non-zero $\sigma_i$ on both sides in (4.76) we obtain the inequality

$$\mathrm{MISE}\left[\hat{f}\right] \geq \frac{1}{2} \sum_{i=1}^{r} \min\left\{\left|\langle v_i, f\rangle\right|^2, \frac{\mathrm{E}\left\{\left|\langle v_i, \nu\rangle\right|^2\right\}}{\sigma_i^2}\right\} \tag{4.77}$$

Even for the best conceivable filter coefficients, the MISE of the regularized estimator is lower-bounded by this quantity. In a way, then, it defines the limits within which a sensible choice of $r_\alpha$ can help to improve the error performance. To make this margin as large as possible clearly we should strive to minimize the right-hand-side of (4.77). Ideally, of course, we would have noise and signal lie in different subspaces altogether, resulting in two disjunctive sets of coefficients. Unfortunately, this is not a realistic expectation. Note that in the case of AWGN noise $\nu \sim \mathcal{N}(0, \sigma_\nu^2 I)$ assumed without loss of generality, for any orthonormal basis $V = (v_1, ..., v_n)$ we have $\mathrm{E}\left\{\left|\langle v_i, \nu\rangle\right|^2\right\} = \mathrm{E}\left\{V^H \nu\nu^H V\right\}_{ii} = \sigma_\nu^2$ and thus a lower bound of

$$\frac{1}{2} \sum_{i=1}^{r} \min\left\{\left|\langle v_i, f\rangle\right|^2, \frac{\sigma_\nu^2}{\sigma_i^2}\right\} \tag{4.78}$$

As far as the noise is concerned, therefore, the particular choice of $V$ does not make any difference. The same is not true, however, for the other term inside the min-operator. For the signal, typically featuring a high degree of correlation and structure, the choice of an appropriate basis does matter indeed. In order to keep (4.78) low, it is desirable that the image have a representation as economic as possible, and be rendered by only few coefficients $\langle v_i, f\rangle \neq 0$. Whether a basis $V$ is adequate or suitable in this sense obviously depends on the class of images considered. It will be good for one and less so for others; none will be optimal for all.

The Fourier-Domain with its sinusoidal kernels turns out to be best for images featuring gradual rather than abrupt transitions in light-intensities. In fact, the decay in the Fourier coefficients may be shown to be directly related to the smoothness of $f$. On the other hand, images with sharp contrast and singularities are not rendered economically, resulting in a comparably large number of non-zero coefficients. Here is where the wavelets come into play. In fact, the ForWaRD algorithm is based on the very observation that wavelet transform domains are particularly apt for the representation of piecewise but not globally smooth signals. For a formal description of this class in terms of Sobolev and Besov spaces see the paper already cited [19] and the references therein.

**Wavelet Transform** Wavelet-theory by itself is rich and complex material. We only give the essential idea, sticking to the strictly indispensable. Let $\phi$ and $\psi$ be two prototype function (low-pass scaling and mother wavelet), chosen such that the set of dilated and shifted versions

$$\phi_{j,l}(t) = 2^{j/2}\,\phi(2^j t - l) \qquad\qquad \psi_{j,l} = 2^{j/2}\,\psi(2^j t - l) \tag{4.79}$$

with parameters $j, l \in \mathbb{Z}$ comprise an orthonormal basis. Then an arbitrary (one-dimensional) signal $f$ can be approximated up to a finite resolution by an appropriate linear combination

$$f^J(t) = \sum_{l=0}^{N_0} s_l\,\phi_l(t) + \sum_{j=0}^{J}\sum_{l=0}^{N_j} w_{j,l}\,\psi_{j,l}(t) \tag{4.80}$$

with $s_l := \langle \phi_l, f\rangle$ and $w_{j,l} := \langle \psi_{j,l}, f\rangle$ the projections onto the corresponding basis-functions. (The approximation $f^J$ may be shown to converge to $f$ in $\mathcal{L}^2$-norm, as the resolution gets finer). Just like the Fourier-transform the concept readily extends to higher dimensions and sampled discrete-time signals. Here $\langle \phi_l, f\rangle$ and $\langle \psi_{j,l}, f\rangle$ basically represent convolution with low and high- or band-pass filters respectively. An important family of kernels satisfying orthonormality and vanishing moment constraints is given by the Daubechies coefficient sets; they were used for the actual implementation.

**Concept and Pseudo-Code Implementation** With its concept of alternate filtering in Fourier and wavelet domains — combining and exploiting the specific advantages of either — the Neelamani approach is essentially hybrid.

To altogether abandon the Fourier-domain is not practical, as it is the only way to efficiently handle the actual deconvolution. Also, the inferiority in representing images of sharp contrast is balanced by the fact that colored noise can be rendered very economically. (Note that, once inverted, the noise is no longer white but strongly correlated). The wavelet-domain, on the

other hand, can be used in a second step to further attenuate the artefacts due to noise leaked from the first stage. The choice of filter functions — whether Tikhonov/Ridge Regression, Wiener or simple thresholding — is arbitrary and at the discretion of the implementor; may he proceed as he thinks fit. Following the recommendations in the aforementioned paper, a bit of all has been employed (see below for details).

Since MATLAB does not natively support discrete wavelet transform (DWT) we only give a schematic description using informal pseudo-code. A reference implementation by the inventors (MATLAB front-end based on C++ routines by MEX-interface) is hosted by the Rice University in Houston, Texas and available for download at `www.dsp.rice.edu/software`.

STEP 1: FOURIER DOMAIN FILTERING

- Regularized operator inversion using Ridge-Regression

$$r_\alpha(\sigma_i) = \frac{\sigma_i^2}{\sigma_i^2 + \alpha}$$

with $\alpha$ set judiciously (moderate amount of regularization)

STEP 2: WAVELET DOMAIN FILTERING

1. **Compute 'pilot' estimate**
   - (Redundant) DWT
     - coefficient set: Daubechies 6
     - fixed number of decomposition levels
   - Hard thresholding (as in TSVD) with

$$r_{\alpha^{(j)}}(x) = \begin{cases} 1 & |x| > \alpha^{(j)} \\ 0 & \text{otherwise} \end{cases} \qquad x \in \{s_l, w_{j,l}\}$$

     and $\alpha^{(j)}$ an estimate of the noise level at scale $j$.
   - Inverse DWT
2. **Compute final estimate**
   - (Redundant) DWT
     - coefficient set: Daubechies 2
     - fixed number of decomposition levels
   - Wiener Filtering with

$$r_{\alpha^{(j)}}(x) = \frac{|x|^2}{|x|^2 + \alpha^{(j)}} \qquad x \in \{s_l, w_{j,l}\}$$

     and $\alpha^{(j)}$ an estimate of the inverse SNR at scale $j$.
   - Inverse DWT

An important issue is that of balancing the amount of Fourier- and wavelet-filtering controlled by the regularization parameters, and hence the weight of the two stages relative to each other. This question is dealt with in the paper [19] to some detail. It should be consulted for information on how to estimate the optimal parameter configurations for a given setting.

**Complexity** The complexity of the algorithm is dominated by the two transforms, actually of order $O(n \log n)$. If the redundant variant of the wavelet transform is used, asymptotic behaviour is worse. However, for most images of customary size it proves sufficient to have a fixed number of decomposition levels which limits computational and storage cost, otherwise considerable. Even so, performance is swift compared with the iterative algorithms previously discussed.

# 5 Hardware Adaptation

As many of the best-performing algorithms to date do not fall into the blind category, their effective use entails the supply of a-priori information about the optical device and the recording system. Such contextual data, in general, is rarely available. To have the relevant hardware at our disposition, thus, is a privilege to take advantage of. See table 5.1 for a detailed lists of components.

In this chapter we discuss different ways of making such a-priori information available. The first part is concerned with quantifying the impulse response or point-spread-function of the microscope. Readout noise of the CCD-camera will be dealt with in the second part.

## 5.1 Microscope PSF

There are different ways to go about estimating the PSF, which could be classified according to the following scheme

- **Empirical approaches** attempt to *actually measure* the point-spread-function, approximating the Dirac-impulse by a light-source of as little extension as possible. The idea is straightforward, though possibly not the most challenging from a mathematical point of view. However, pertaining to the field of photometry rather than data analysis, this method is usually quite laborious and expensive in terms of material and equipment.
- **Algebraic approaches**, most notably represented by [6] and [2], are characterized by a high degree of mathematical sophistication. While they get along without so much as the notion of probability, they also fall short of correcting the fundamental ill-posedness of the problem. An important drawback of this determinism, hence, is their vulnerability in the presence of noise.
- **Probabilistic approaches** Many of the so-called blind deconvolution algorithms derived from probabilistic models compute, if only as a necessary by-product in the course of restoration, an estimate of the PSF. Examples of this class are [22, 23] and the EM-based algorithm discussed in the previous chapter. However, these methods tend to be

| **Component** | **Manufacturer or Specification** |
|---|---|
| Microscope | Zeiss Axio Imager M1, dry |
| Lens | Zeiss Achroplan 63x, 0,95 numerical aperture |
| CCD-camera | JAI Pulnix TMC-1402 CL (1392x1040 and 800x600) |
| Lighting | LED |

Table 5.1: Used Optical Hardware (Microscope plus CCD-camera)

lacking precision and rarely yield anything more fancy than a Gaussian bell-curve of slightly bigger or narrower spread.

- **Heuristic approaches** Given this difficulty of quantifying the PSF exclusively from the data, it is no surprise that parametric models are still a valuable resource. Based on the laws of optical physics, they provide a heuristic to predict the PSF under certain conditions.

In the following we discuss one representative of the latter class in some detail.

**Gibson-Lanni Model** For microscopy, one of the more sophisticated models at hand is the one developed by Gibson and Lanni [7]. Due to its high degree of specialization, caution is indicated when applying it to a scenario other than the one it was originally intended for. For two reasons, it is impossible to transfer offhand the findings of the paper to our situation:

1. The Gibson model is designed for fluorescence microscopy, as opposed to the light-transmitting confocal microscope in use at the Fraunhofer Institute of Integrated Circuits (IIS) where the tests have been conducted.
2. Contrary to the problem considered throughout this work, the Gibson paper and the model derived therein are concerned with methods of three-dimensional reconstruction, arising in confocal laser scanning microscopy (CLSM) and optical serial sectioning microscopy (OSM), all of which require that multiple samples of the specimen be available.

It is important to bear in mind these differences which compromise a one-to-one applicability to the present case. We say this in justification of the unorthodox use we are going to make of it. Rather than taking all slots literally, we use it as a heuristic to parameterize a subset of point-spread-functions and reduce the amount of unknowns to a — hopefully — manageable number. Then, starting with a set of given hardware parameters, a local optimization will be conducted to approximate the best fitting PSF.

The Gibson model describes the PSF as a radial symmetric intensity distribution varying as a function of defocus $\Delta z$. We do not undertake to derive the formulae, as a conscientious discussion is beyond the author's competence. The reader interested in the background information is referred to the paper already cited above and the numerous references therein.

The calculation involves an impressive amount of parameters, whose number is further increased — in fact nearly doubled — by distinguishing between actual and ideal values for many of them. Parameters denoting the values in the design system under ideal conditions are decorated with an asterisk. A explanatory list is given below (see also figure 5.1 for a schematic description)

For its essential part, the Gibson model may be reduced to the following equations. Let $k = 2\pi/\lambda$ the wave number and

$$\begin{aligned} W(\rho, \Delta z) = {} & t_s \sqrt{n_s^2 - (A\rho)^2} + t_g \sqrt{n_g^2 - (A\rho)^2} \\ & - t_g^* \sqrt{n_g^* - (A\rho)^2} - t_i^* \sqrt{n_i^{*2} - (A\rho^2)} \\ & + \left[ \Delta z + n_i \left( \frac{t_g^*}{n_g} + \frac{t_i}{n_i^*} - \frac{t_s}{n_s} - \frac{t_g}{n_g} \right) \right] \sqrt{n_i^2 - (A\rho)^2} \end{aligned} \tag{5.1}$$

| Parameter | Meaning |
|---|---|
| $A$ | Numerical Aperture |
| $M$ | Magnification |
| $t_i$ | Depth of the immersion medium |
| $n_i$ | Refractive index of the immersion medium |
| $t_g$ | Depth of the coverslip |
| $n_g$ | Refractive index of the coverslip |
| $t_s$ | Depth of the ROI in the specimen |
| $n_s$ | Refractive index of the specimen |
| $\lambda$ | Wavelength of the light |

Table 5.2: Parameters in the Gibson-Lanni Model

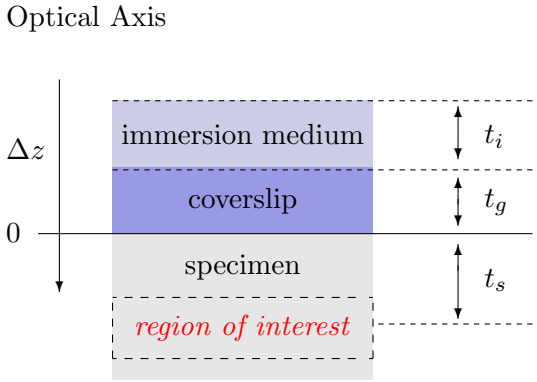



Figure 5.1: Longitudinal cut through the different strata of the optical path

the optical path difference (OPD) provoking phase aberrations. Then the intensity emanating from a point-light-source in the region of interest of the specimen and sensed by a detector at distance (radius) $\rho$ in the image plane is given by the following integral, derived from Kirchhoffs diffraction formula

$$I(r, \Delta z) \propto \left| \int_0^1 J_0\left(kr\sqrt{M^2 - A^2}\,\rho\right) \cdot e^{ikW(\rho,\Delta z)} \cdot \rho \, d\rho \right|^2 \tag{5.2}$$

where

$$J_\alpha(x) = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{\imath(\alpha t - x \sin t)}\, dt \tag{5.3}$$

the Bessel function of the first kind.

**Implementation** Our implementation had to overcome several obstacles related to the numerical evaluation of the integral (5.2), the most important one being insufficient performance.

Although the adaptive Simpson quadrature `quad.m` that ships with MATLAB is suitable for most purposes, it turned out to be prohibitively slow in this particular case. For some reason that we have not been able to fully track down, the routine either produced unacceptably inaccurate results or — with smaller tolerance — took around 40 seconds of computation time for one PSF (63x63 pixels, requiring approximately 45 integral evaluations), which was judged impractical for the optimization procedure with several hundreds to thousands of function calls. All calculations were performed on an Intel Pentium 4 with 3 GHz and 512 MB memory. The same problem arose when using GNU Octaves `quad.m`, which serves as a front end to the FORTRAN integration package Quadpack, worsened furthermore by the function's inability to handle complex integrals.

To altogether abandon the convenient scripting languages, however, would have meant to forfeit a precious instrumentarium for subsequent optimization. As a viable expedient, finally, it was chosen to implement the actual quadrature in plain C or C++, using the MEX and OCT interfaces and conduct the optimization in MATLAB/Octave as originally intended. By eliminating this performance bottleneck, the computation time for one PSF could be successfully reduced to a fraction of a second.

Figure 5.2 shows some intensity distributions as defined by (5.2), computed by numerical evaluation of the intergral for different values of $\Delta z$ and otherwise arbitrarily chosen but constant parameters.

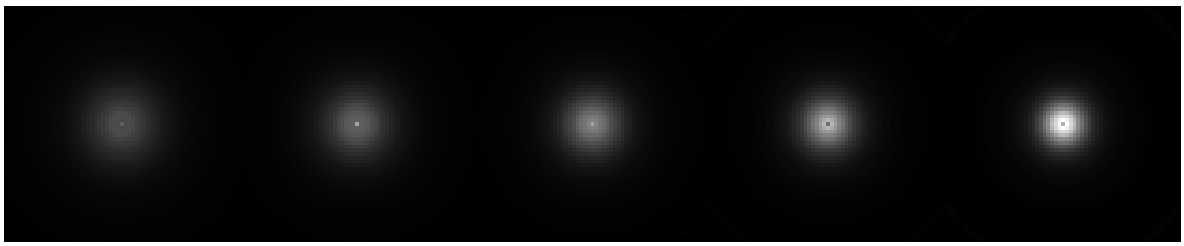


Figure 5.2: PSFs of varying defocus as predicted by the Gibson-Lanni model

**Execution** To assess how well a model-generated PSF approximates the actual impulse response of the microscope, the following approach was adopted. Given a pair of images, one of them in-focus (as much so as one can realistically hope for by manual calibration), the other deliberately out-of-focus but otherwise displaying exactly the same region of interest in the specimen, it is clear that the latter represents a degraded version of the former. Now let $f$ and $g$ denote sharp and blurred image respectively. To quantify the goodness of fit of a particular PSF, say $h$, we would convolve the focused image with the generated PSF and compare the result with the 'naturally' blurred out-of focus image. As cost function the square error

$$\phi(h) = \|f * h - g\|_2^2 \tag{5.4}$$

was chosen, with $(*)$ denoting convolution. Figure 5.1 shows the image pair used in this way, each of size 300x300 pixels. The start values for the iterations are given in the table 5.3 below, where 'ideal' and 'actual' parameters are assumed to be identical.

| Parameter | Start Value |
|---|---|
| $A$ | 0.95 |
| $M$ | 63.0 |
| $t_i$ | $0.19 \times 10^{-3}$ |
| $n_i$ | 1.0 |
| $n_i^*$ | 1.0 |
| $t_g$ | $0.17 \times 10^{-3}$ |
| $t_g^*$ | $0.17 \times 10^{-3}$ |
| $n_g$ | 1.515 |
| $n_g^*$ | 1.515 |
| $t_s$ | $20.21 \times 10^{-9}$ |
| $n_s$ | 1.46 |
| $\lambda$ | $520.0 \times 10^{-9}$ |
| $\Delta z$ | 0 |

Table 5.3: Start values for local optimization

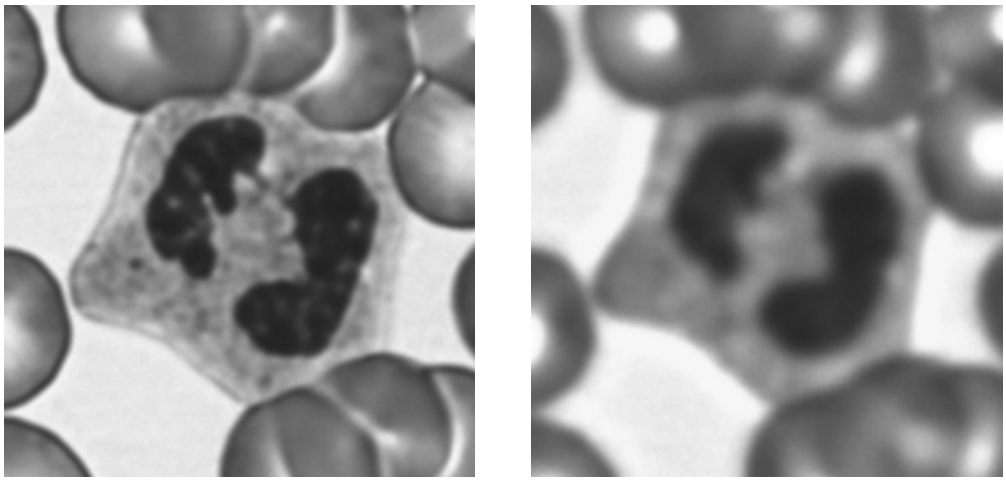


Figure 5.3: Image pair, focused and defocused, used for optimization

Local optimization was performed using MATLABs `fminsearch.m`, based on the Nelder-Mead simplex algorithm. The iterates finally converged to the local minimizer of (5.4) whose values are given in table 5.4.

| Parameter | Final Value |
|---|---|
| $A$ | 0.97 |
| $M$ | 30.856 |
| $t_i$ | $0.328 \times 10^{-3}$ |
| $n_i$ | 1.015 |
| $n_i^*$ | 1.010 |
| $t_g$ | $0.168 \times 10^{-3}$ |
| $t_g^*$ | $0.215 \times 10^{-3}$ |
| $n_g$ | 1.264 |
| $n_g^*$ | 1.248 |
| $t_s$ | $1.78 \times 10^{-5}$ |
| $n_s$ | 1.589 |
| $\lambda$ | $586 \times 10^{-9}$ |
| $\Delta z$ | $2.177 \times 10^{-3}$ |

Table 5.4: Local minimizer of the MSE cost function

Some of the values appear reasonable at first sight. A more careful examination, however, reveals that they are either trivial or nonsense. A defocus of more than one millimeter, in particular, is somewhat gross. Having mentioned, with utmost candour, the most palpable improbability, one will kindly exempt us from commenting any further on the results. We repeat, though, that from the very outset we used the model as a heuristic to parameterize a subset of intensity distributions rather than interpret all slots strictly on the grounds of their physical meaning.

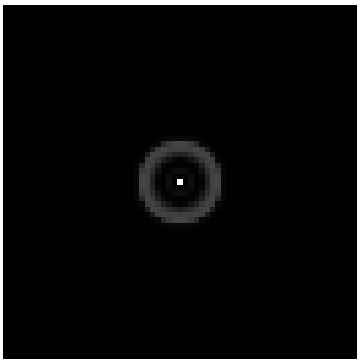


Figure 5.4: Best fitting PSF (logscale intensities)

To be sure, we also tried searching on a global scale for a minimum of (5.4). For this purpose simulated annealing, as provided by the function `samin.oct` in the package Octave Forge, was used — to no avail, though. To make the problem manageable and increase the odds of a successful termination, the dimensionality of the search space was drastically reduced by fixing up to 6 parameters. Even so, the routine failed to converge after $10^4$ evaluations and more than 24h of computation time.

## 5.2 CCD-Camera Noise

Consistent with the assumption of Gaussianity — and hence a distribution completely specified by the first two moments —, we shall seek to establish a second order statistic of the noise random process. Estimating the covariances, in particular, represents a major challenge here.

**Idea** Given an ensemble of, say, 100 darkframe images of size 800x600 each, an obvious way to go about a statistical analysis would be to construct the sample covariances from the data. (Even if the frames are captured with small temporal distance from each other, we shall assume them to be independent and identically distributed, thus neglecting a possible correlation along the temporal axis.) While this approach would permit, in principle, a verification of wide-sense-stationarity — one of the more axiomatic assumptions our noise model is based upon — the statistical significance of such an experiment is severely compromised by the scarcity of samples. We briefly elaborate this point. If we let $k$ denote the number of available observations, then the sample covariance matrix can be represented as a sum of as many rank-one-updates

$$\overline{C} = \frac{1}{k-1} \sum_{i=1}^{k} (\nu_i - \overline{\nu})(\nu_i - \overline{\nu})^H \tag{5.5}$$

where the sample mean $\overline{\nu}$ of the random process is defined as

$$\overline{\nu} = \frac{1}{k} \sum_{i=1}^{k} \nu_i \tag{5.6}$$

From (5.5) it is evident that $\operatorname{rank}(\overline{C}) \leq k$ is bounded by the number of contributing samples. For the true variance matrix, however, we generally have $\operatorname{rank}(\Lambda_\nu) = mn$, where $mn$ is the size of one observation, and $\operatorname{rank}(\Lambda_\nu) < mn$ if and only if two components are perfectly correlated by a virtual determinism such that

$$\rho = \frac{\operatorname{Cov}\left(\nu^{(x,y)}, \nu^{(\xi,\eta)}\right)}{\sqrt{\operatorname{Var}\left[\nu^{(x,y)}\right] \operatorname{Var}\left[\nu^{(\xi,\eta)}\right]}} \in \{-1, 1\} \tag{5.7}$$

for some $(x, y), (\xi, \eta) \in \{1, ..., m\} \times \{1, ..., n\}$.

Now, to yield a meaningful approximation to the real covariances, we should require the number of observations contributing in (5.5) to exceed the size of one observation. Given the actual resolution of 800x600 pixels, however, we will inevitably have $k \ll mn = 480000$ if the problem is to remain tractable. In other words, (5.5) is poorly determined by the few observations that happen to be at hand. Whatever the outcome of such an experiment, without sufficient backing from the data to sustain and substantiate it, our findings will remain challengeable at best.

Lacking an adequate amount of samples to either refute or verify wide-sense-stationarity with sufficient authority, we may just as well take it for granted and make it work for us. It stands to reason that incorporating a-priori knowledge into the estimator makes for more

accurate results with significantly fewer observations. In fact, we will show that the WSS-assumption can be used to improve the general purpose covariance estimate (5.5) by a substantial factor.

In order to address the aforementioned issues arising whenever

$$\dim(\nu) \gg k \tag{5.8}$$

with $k$ the number of available iid samples, we propose two covariance estimators applicable to large-scale problems where the random process is known to be wide-sense-stationary. We begin with a formal proof of unbiasedness before we present an efficient MATLAB implementation using Fast Fourier Transform (FFT) in a subsequent paragraph. Finally, the theoretical findings are confirmed and corroborated by numerical simulation, showing that for WSS-processes the proposed method indeed largely outperforms the general purpose estimator (5.5).

### WSS-covariance Estimators

**WSS Covariance Estimator (1).** *Let $\nu \in \mathbb{R}^{m\times n}$ be a wide-sense stationary random process with*

$$\forall x, y : \mathrm{E}\left\{\nu^{(x,y)}\right\} = \mu_\nu \tag{5.9a}$$

$$\mathrm{E}\left\{\nu^{(x,y)} \cdot \nu^{(\xi,\eta)}\right\} = R_{\nu\nu}(x-\xi, y-\eta) \tag{5.9b}$$

*Let further $\nu_1, ..., \nu_k$ be a sequence of $k$ independent and identically distributed observations. Define*

$$S_{ij} = \left\{ \begin{pmatrix} x \\ \xi \\ y \\ \eta \end{pmatrix} \in \{1,...,m\}^2 \times \{1,...,n\}^2 \;\middle|\; \begin{pmatrix} x-\xi \\ y-\eta \end{pmatrix} = \begin{pmatrix} i \\ j \end{pmatrix} \right\} \tag{5.10}$$

*to be the set of index-pairs representing components located at (two-dimensional) distance $(i, j)$ from each other. Finally let*

$$\overline{\nu} = \frac{1}{mnk} \sum_l \sum_{x,y} \nu_l^{(x,y)} \tag{5.11}$$

*denote the sample mean formed by averaging over the components of all observations. Then*

$$\overline{R}_{ij} = \frac{1}{k \cdot |S_{ij}|} \sum_{\substack{(x,\xi,y,\eta)\in S_{ij} \\ 1\le l\le k}} \sum \nu_l^{(x,y)} \nu_l^{(\xi,\eta)} \tag{5.12}$$

*and*

$$\overline{C}_{ij} = \overline{R}_{ij} - \frac{k\overline{\nu}^2 + 1/(m^2n^2)\sum_{x,y}|S_{xy}|\overline{R}_{xy}}{k+1} \tag{5.13}$$

*are unbiased estimators for $R_{\nu\nu}(i, j)$ and the corresponding covariance $C_{\nu\nu}(i, j) = R_{\nu\nu}(i, j) - \mu_\nu^2$ respectively.*

**Proof:** Due to the linearity of the expectation operator, it clearly holds that $\mathrm{E}\left\{\overline{R}_{ij}\right\} = R_{\nu\nu}(i,j)$. To see the second statement, we take a closer look at

$$\mathrm{E}\left\{\overline{\nu}^2\right\} = \frac{1}{m^2n^2k^2}\,\mathrm{E}\left\{\sum_{s,x,y}\sum_{t,\xi,\eta}\nu_s^{(x,y)}\cdot\nu_t^{(\xi,\eta)}\right\} \tag{5.14}$$

Since the samples are independent and therefore mutually uncorrelated we find that

$$\mathrm{E}\left\{\nu_s^{(x,y)}\,\nu_t^{(\xi,\eta)}\right\} = \begin{cases} \mu_\nu^2 & s \neq t \\ R_{\nu\nu}(x-\xi, y-\eta) & s = t \end{cases} \tag{5.15}$$

The former case occurs $m^2n^2(k^2-1)$ times in the sum of (5.14), the latter the remaining $m^2n^2k$ times. Thus we have

$$\begin{aligned} \mathrm{E}\left\{\overline{\nu}^2\right\} &= \frac{k-1}{k}\mu_\nu^2 + \frac{1}{m^2n^2k}\sum_{x,y,\xi,\eta}\sum R_{\nu\nu}(x-\xi, y-\eta) \\ &= \mu_\nu^2 + \frac{1}{m^2n^2k}\sum_{x,y,\xi,\eta}\sum C_{\nu\nu}(x-\xi, y-\eta) \end{aligned} \tag{5.16}$$

where $C_{\nu\nu} = R_{\nu\nu} - \mu_\nu^2$ is the (auto-)covariance function of the process. Now consider that $S$ defines a partition on $I^2$ where $I = \{1,...,m\}\times\{1,...,n\}$ is the set of two-dimensional indices. In other words, each index-pair is element of exactly one equivalence class $S_{ij}$. In particular, it holds that $\sum_{ij}|S_{ij}| = m^2n^2$. Since $\mathrm{Cov}\left(\nu^{(x,y)}, \nu^{(\xi,\eta)}\right) = C_{\nu\nu}(i,j)$ for all $(x,y,\xi,\eta) \in S_{ij}$ summing over all index-pairs in (5.16) is equivalent with the sum over the sets $S_{ij}$ weighted with their cardinality. Hence for a provisional covariance estimator $\widehat{C}_{ij} = \overline{R}_{ij} - \overline{\nu}^2$ we obtain

$$\begin{aligned} \mathrm{E}\left\{\widehat{C}_{ij}\right\} &= R_{\nu\nu}(i,j) - \mu_\nu^2 - \frac{1}{m^2n^2k}\sum_{x,y}|S_{xy}|\,C_{\nu\nu}(x,y) \\ &= C_{\nu\nu}(i,j) - \frac{1}{m^2n^2k}\sum_{x,y}|S_{xy}|\,C_{\nu\nu}(x,y) \end{aligned} \tag{5.17}$$

If we stack the components of $\widehat{C}$ and $C_{\nu\nu}$ vertically and denote by $|S|$ the vector formed of the cardinalities $|S_{ij}|$ we may write (5.17) as

$$\widehat{C} = (I+U)\cdot C_{\nu\nu} \tag{5.18}$$

where

$$U = \frac{1}{m^2n^2k}\begin{pmatrix} \cdots & |S|^T & \cdots \\ & \vdots & \\ \cdots & |S|^T & \cdots \end{pmatrix} \tag{5.19}$$

is a rank-one matrix with replicate rows. Evidently, $\widehat{C}$ has bias $UC_{\nu\nu}$. Note also that an unbiased estimator is implicitly given by the following system of linear equations

$$\overline{C} = (I+U)^{-1}\widehat{C} \tag{5.20}$$

which is impractical for a concrete implementation due to its huge dimension of order $O(m^2n^2)$. However, it turns out that (5.20) can be solved efficiently with much less effort from the following observations. Let $d = \dim(\widehat{C})$ denote the number of equations in (5.20) and, purely for notational convenience, $S' = 1/(m^2n^2k)|S|$ an appropriately scaled instance of the vector holding the cardinalities of the equivalence classes. Then from

$$\overline{C}_{ij} + \langle S', \overline{C}\rangle = \widehat{C}_{ij} \tag{5.21}$$

we get, by summing over all $i, j$

$$\begin{aligned} 0 &= \langle 1, \widehat{C}\rangle - \langle 1, \overline{C}\rangle - d \cdot \langle S', \overline{C}\rangle \\ &= \langle 1, \widehat{C}\rangle - \langle 1 + d\, S', \overline{C}\rangle \end{aligned} \tag{5.22}$$

On the other hand we find, likewise from (5.21), that $\overline{C} = \widehat{C} - \alpha 1$ with $\alpha = \langle S', \overline{C}\rangle$. Substituting into (5.22) yields

$$\begin{aligned} 0 &= \langle 1, \widehat{C}\rangle - \langle 1 + d\, S', \widehat{C} - \alpha 1\rangle \\ &= d \cdot \Big(\alpha - \langle S', \widehat{C}\rangle + \alpha\langle S', 1\rangle\Big) \end{aligned} \tag{5.23}$$

By solving for $\alpha$ we finally get

$$\alpha = \frac{\langle S', \widehat{C}\rangle}{\langle S', 1\rangle + 1} = \frac{\langle |S|, \overline{R} - \overline{\nu}^2\rangle}{m^2n^2(k+1)} = \frac{1/(m^2n^2)\langle |S|, \overline{R}\rangle - \overline{\nu}^2}{k+1} \tag{5.24}$$

where the identity $\langle |S|, 1\rangle = m^2n^2$ has been used. Hence the unbiased estimator has the explicit representation

$$\begin{aligned} \overline{C}_{ij} &= \overline{R}_{ij} - \overline{\nu}^2 - \alpha \\ &= \overline{R}_{ij} - \frac{k\overline{\nu}^2 + 1/(m^2n^2)\sum_{x,y}|S_{xy}|\,\overline{R}_{xy}}{k+1} \end{aligned} \tag{5.25}$$

as claimed. ■

Being, essentially, the unnormalized auto-correlation of the random process with retroactive mean-subtraction, the above estimator can be efficiently implemented using FFT. We give a reference implementation in MATLAB-code which should speak for itself. It is self-contained except for the outsourcing of the i/o part into the separate function `load_sample` for the sake of readability. Anticipating a variant that is covered in a later paragraph, the function takes a boolean parameter specifying whether or not the variance matrix has symmetric Toeplitz-blocks.

```
function [mu, C] = wss_estim1(k, m, n, STB)
%% unbiased wss covariance estimator (1)

%% k      number of iid samples to evaluate
%% m x n  size of one observation
%% STB    symmetric block Toeplitz variant

mu = 0;
R = zeros(2*m-1, 2*n-1);

for l = 1 : k
   s = load_sample(l, [m, n]);
   mu = mu + sum(s(:));
   s(2*m-1, 2*n-1) = 0;
   R = R + real(ifft2(abs(fft2(s)).^2));
end

mu = mu / (m*n*k);

%% enforce axial symmetry for STB variant
if STB
   R(2:end,2:end) = (R(2:end, 2:end) + R(2:end, end:-1:2)) / 2;
   R(2:end,2:end) = (R(2:end, 2:end) + R(end:-1:2, 2:end)) / 2;
end

%% normalize and solve implicit equation
S = [m:-1:1,1:m-1].'* [n:-1:1,1:n-1];
T = sum(R(:)) / (m^2*n^2*k);
C = R ./ (k*S) - (k*mu^2 + T)/(k+1);
```

We note as an aside that due to a well-known property of the Fourier-Transform, it holds that $\sum_{xy}|S_{xy}|\overline{R}_{xy} = 1/k\sum_l \|\nu_l\|^2$, where $\|\cdot\|$ is either the 2-norm or the Frobenius norm depending on whether we choose to regard $\nu_l$ as a column-vector or a matrix. However, since we need to calculate the FFT anyway, this identity, while looking more straightforward, does not reduce the overall-complexity of the algorithm.

For comparison, we give another covariance estimator, likewise unbiased. In contrast to the one developed above, normalization is done 'in place', resulting essentially in the auto correlation of the mean subtracted process.

**WSS Covariance Estimator (2).** *Suppose the conditions as described in* (5.9a) *hold with, again, $k$ iid sample observations $\nu_1, ..., \nu_k$. Define*

$$\overline{\nu}_{\bullet} = \frac{1}{m\,n\,(k-1)} \sum_{s \neq l} \sum_{x,y} \nu_s^{(x,y)} \tag{5.26}$$

*to be the leave-out-one sample mean, formed by averaging over all observations but one. Let further*

$$\widehat{C}_{ij} = \frac{1}{k \cdot |S_{ij}|} \sum_{\substack{(x,\xi,y,\eta) \in S_{ij} \\ 1 \leq l \leq k}} \sum (\nu_l^{(x,y)} - \overline{\nu}_{\bullet})(\nu_l^{(\xi,\eta)} - \overline{\nu}_{\bullet}) \tag{5.27}$$

*Then*

$$\overline{C}_{ij} = \widehat{C}_{ij} - \frac{\sum_{xy} |S_{xy}| \widehat{C}_{xy}}{m^2 n^2 k} \tag{5.28}$$

*is un unbiased estimator for the auto-covariance $C_{\nu\nu}(i,j)$.*

**Proof:** Let $i, j$ be fixed. For each $(x, \xi, y, \eta) \in S_{ij}$ we have

$$\mathrm{E}\left\{\nu_l^{(x,y)} \overline{\nu}_{\bullet}\right\} = \frac{\mathrm{E}\left\{\nu_l^{(x,y)} \sum_{s \neq l} \sum_{\xi,\eta} \nu_s^{(\xi,\eta)}\right\}}{mn(k-1)} = \mu_\nu^2 \tag{5.29}$$

due to the independency of the observations. On the other hand,

$$\mathrm{E}\left\{\overline{\nu}_{\bullet} \overline{\nu}_{\bullet}\right\} = \mu_\nu^2 + \frac{\sum_{xy} |S_{xy}| C_{\nu\nu}(x,y)}{m^2 n^2 (k-1)} \tag{5.30}$$

Therefore

$$\begin{aligned} \mathrm{E}\left\{\widehat{C}_{ij}\right\} &= \frac{1}{k|S_{ij}|} \mathrm{E}\left\{(\nu_l^{(x,y)} - \overline{\nu}_{\bullet})(\nu_l^{(\xi,\eta)} - \overline{\nu}_{\bullet})\right\} \\ &= C_{\nu\nu}(i,j) + \mu_\nu^2 - 2\mu_\nu^2 + \mu_\nu^2 + \frac{\sum_{xy} |S_{xy}| C_{\nu\nu}(x,y)}{m^2 n^2 (k-1)} \end{aligned} \tag{5.31}$$

or, equivalently, looking at the whole system and using matrix vector notation

$$\mathrm{E}\left\{\widehat{C}\right\} = (I + U) \cdot C_{\nu\nu} \tag{5.32}$$

with

$$U = \frac{1}{m^2 n^2 (k-1)} \begin{pmatrix} \dots & |S|^T & \dots \\ & \vdots & \\ \dots & |S|^T & \dots \end{pmatrix} \tag{5.33}$$

From (5.20) and following we know that the solution of $\overline{C} = (I + U)^{-1} \widehat{C}$ is given by

$$\overline{C} = \widehat{C} - \frac{\langle |S|, \widehat{C} \rangle}{\langle |S|, 1 \rangle + m^2 n^2 (k-1)} = \widehat{C} - \frac{\langle |S|, \widehat{C} \rangle}{m^2 n^2 k} \tag{5.34}$$

Then, clearly, we have

$$\mathrm{E}\left\{\overline{C}\right\} = (I + U)^{-1} \mathrm{E}\left\{\widehat{C}\right\} = C_{\nu\nu} \tag{5.35}$$

■

Again, we provide a reference implementation as a MATLAB function with the same signature as above. We repeat that the actual reading of the samples, for being alien to the problem of interest, has been omitted in the listing below.

```
function [mu, C] = wss_estim2(k, m, n, STB)
%% unbiased wss covariance estimator (2)

%% k      number of iid samples to evaluate
%% m x n  size of one observation
%% STB    symmetric block Toeplitz variant

%% calculate leave-out-one sample means
mu = zeros(k, 1);
for l = 1 : k
  s = load_sample(l, [m, n]);
  mu(l) = sum(s(:));
end
mu = (sum(mu) - mu) / ((k-1)*m*n);

%% calculate sample correlation via FFT
C = zeros(2*m-1, 2*n-1);
for l = 1 : k
  s = load_sample(l, [m, n]);
  s = s - mu(l);
  s(2*m-1, 2*n-1) = 0;
  C = C + real(ifft2(abs(fft2(s)).^2));
end

%% sample mean
mu = sum(mu) / k;

%% enforce axial symmetry for STB variant
if STB
  C(2:end,2:end) = (C(2:end, 2:end) + C(2:end, end:-1:2)) / 2;
  C(2:end,2:end) = (C(2:end, 2:end) + C(end:-1:2, 2:end)) / 2;
end

%% normalize and solve implicit equation
S = [m:-1:1,1:m-1].' * [n:-1:1,1:n-1];
C = C ./ (k*S) - sum(C(:))/(k^2*m^2*n^2);
```

**STB-Variant** From its very definition it is clear that the unnormalized auto-correlation for any random process is point-symmetric about the origin, such that

$$R_{\nu\nu}(i,j) = R_{\nu\nu}(-i,-j) \tag{5.36}$$

holds. Stating that covariance matrices are symmetric, then, is a rather trivial remark. Slightly more interesting is the case where this property supervenes in combination with wide-sense-stationarity of the process to induce a symmetric block Toeplitz structure with Toeplitz blocks (SBTTB)

$$\Lambda_\nu = \begin{pmatrix} C_1 & C_2 & & \dots & C_n \\ C_2^T & C_1 & C_2 & & \vdots \\ \vdots & C_2^T & C_1 & \ddots & \vdots \\ \vdots & & \ddots & \ddots & C_2 \\ C_n^T & & & C_2^T & C_1 \end{pmatrix} \tag{5.37}$$

Due to the symmetry of $\Lambda\nu$ we certainly have $C_1 = C_1^T$, but the same need not be true for the remaining blocks. Further simplification of the model, thus, might start right here. In fact, from wide-sense-stationarity it is not a far step to enforcing symmetry of all $C_i$, resulting in a covariance matrix that is SBTSTB. Note that this corresponds to the axial symmetry of the auto-correlation function $R_{\nu\nu}(i,j) = R_{\nu\nu}(-i,j)$, which in combination with (5.36) also coerces symmetry along the other axis. This assumption can be argued on the grounds that the Euclidean distance is the same in either case. In a way, then, we abide with the logic of wide-sense-stationary, albeit with a different concept of distance. One of the convenient side-effects, by the way, is to reduce the number of defining parameters in (5.37) to $mn$ or, equivalently, the size of one observation.

In this paragraph we present a variant of the above estimator suitable for WSS-processes where

$$\mathrm{E}\left\{\nu^{(x,y)} \cdot \nu^{(\xi,\eta)}\right\} = R_{\nu\nu}(|x-\xi|, |y-\eta|) \tag{5.38}$$

All that needs to be done, in fact, is change the definition of the equivalence classes (5.10) into

$$S_{ij}^{(\mathrm{STB})} = \left\{ \begin{pmatrix} x \\ \xi \\ y \\ \eta \end{pmatrix} \in \{1,...,m\}^2 \times \{1,...,n\}^2 \;\middle|\; \begin{pmatrix} |x-\xi| \\ |y-\eta| \end{pmatrix} = \begin{pmatrix} i \\ j \end{pmatrix} \right\} \tag{5.39}$$

the remaining formulae retain their validity. Note that the sets' cardinalities vary as a function of to the model type

$$|S_{ij}| = |m-i||n-j| \qquad |S_{ij}^{(\mathrm{STB})}| = \begin{cases} |S_{ij}| \cdot 2^{(2-\delta(i,0)-\delta(0,j))} & i,j \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{5.40}$$

with $\delta(i,j)$ the Kronecker delta. To keep the implementation as simple as possible, it has been chosen to always return a full matrix of estimated covariances, with entries corresponding to lags $-(m-1),...,(m-1)$ in the first and $-(n-1),...,(n-1)$ in the second dimension, regardless of redundancy.

**Error Performance** Though a rigorous analysis of the mean-square-error performance is somewhat tedious and has not yet been accomplished, it is important to highlight some aspects concerning the accuracy of the derived WSS-estimators. Lacking, for the time being, a precise quantification of the error, we propose the following heuristic model, to be tested, subsequently, by numerical simulation. For each $p = (x, \xi, y, \eta, l) \in S_{ij} \times \{1, ..., k\}$ with fixed $i, j$ define $X_p := (\nu_l^{(x,y)} - \overline{\nu})(\nu_l^{(\xi,\eta)} - \overline{\nu})$. Then, clearly

$$C_{ij} \approx \frac{1}{k\,|S_{ij}|} \sum_p X_p = \overline{X} \tag{5.41}$$

The sample mean over a population $X$ comprised of $k$ iid observations is known to have the variance

$$\mathrm{Var}\left[\overline{X}\right] = \frac{\sigma_X^2}{k} \tag{5.42}$$

wiht $\sigma_X = \sqrt{\mathrm{Var}\left[X\right]}$ and hence a standard deviation proportional to $k^{-1/2}$. Now, if the random process is Gaussian, all $X_p$ are identically distributed, but — in general — they are not independent. This motivates the following bounds on the standard deviation of the WSS-estimates

$$\frac{\sigma_X}{\sqrt{k\,|S_{ij}|}} \leq \sqrt{\mathrm{Var}\left[\overline{C}_{ij}\right]} \leq \frac{\sigma_X}{\sqrt{k}} \tag{5.43}$$

since only $k$ out of the total of $k|S_{ij}|$ samples are truly independent. In particular, the inequalities (5.43) suggest a convergence rate of $k^{-1/2}$. So while the WSS-estimates have the same asymptotic behaviour as the general purpose estimator (5.5), for any fixed number of samples we can hope to be up to $\sqrt{|S_{ij}|}$ times more accurate, depending on the actual amount of correlation among the components.

In support of such reasoning we are going to present results obtained by numerical simulation. To this effect, the two WSS-covariance estimators (5.13) and (5.28) have been tested on synthetically created data obeying a known distribution. In order to randomly generate the covariance matrix of a wide-sense-stationary process, the following lines of MATLAB code were used with `m = n = 32`

```
%% create SBTSTB variance matrix of size (mn)x(mn)
%% with randomly generated entries

Lambda = zeros(m*n);
T = toeplitz(1:n);
for l = 1 : n
   Lambda = Lambda + ...
   kron(sparse(T==k), toeplitz(rand(m, 1)-0.5));
end

%% make sure that Lambda is positive definite
Lambda = Lambda + (0.1+rand() - min(eig(Lambda))) * speye(m*n);
```

Positive definiteness was ensured by adding an appropriate multiple of the identity matrix. Then a normally distributed sequence, comprised of 500 independent samples, was generated as follows

```
%% calculate standard deviation sigma and define mean mu
sigma = chol(Lambda)';
mu = rand();

%% generate and store samples
for l = 1 : k
   s = sigma * randn(m*n,1) + mu;
   store_sample(s, l);
end
```

By applying the previously discussed estimators for different values of $k$ and subsequent reconstruction of the covariance matrices from their return values we can hope to obtain an eloquent assessment of the error performance. Results are displayed in figure 5.5 (dependency on variation in the number of samples $k$) and 5.7, showing the absolute error as function of the spatial coordinates $i, j$ for fixed $k$.

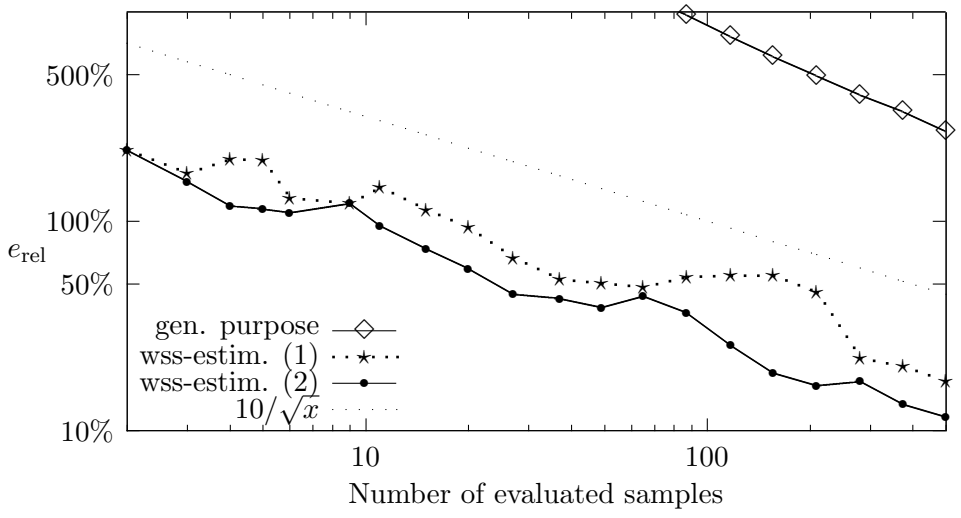



Figure 5.5: Relative Error of Covariance Estimators (in %)

The graphs represent the normalized difference between estimated and actual variance matrices $\left\|\Lambda_\nu - \overline{\Lambda_\nu}\right\|_2 \;/\; \|\Lambda_\nu\|_2$ as a function of evaluated samples. Note that $\nu_1, ..., \nu_{500} \in \mathbb{R}^{32\times 32}$

is an iid Gaussian sequence with $\mathrm{Var}[\nu] = \Lambda_\nu \in \mathbb{R}^{1024\times1024}$ the randomly generated SBT-STB matrix (Symmetric Block Toeplitz with symmetric Toeplitz Blocks). The curves have an average slope of $-1/2$ on the logscale plot, thus confirming the presumed proportionality in (5.43). In perfect accordance with the prognostic also, the three trajectories run roughly parallel to each other, where the smaller intercept suggests that the two WSS-estimators are more accurate by a constant factor.

The upper diagram in figure 5.7 shows a cross-section through the spatial coordinates $i, j$ of the estimate after evaluating the total of 500 samples, showing the absolute error as a function of two-dimensional lag. As expected, it is roughly proportional to $\sqrt{|S_{ij}|}$. To make this behaviour more apparent, compare with the surface-plot below.

Summarizing the above, we conclude that numerical simulation both confirms the superiority of the proposed WSS-covariance estimators over (5.5) and gives strong reason to believe that the bounds (5.43) on the standard deviation formulated *ad hoc* and without proof are correct.

**Application to the Real Data** According to figure 5.5 the second WSS-estimator (5.28) boasts the best overall performance. For the analysis of the 100 darkframe images, recorded by a commercially available standard CCD camera with a resolution of 800 by 600 pixels, this was our method of choice. The results are documented in figure 5.6 showing a perfectly uncorrelated 'white' noise with a peak only at lag $(0, 0)$ corresponding to the variance $\sigma_\nu^2$ and approximately zero otherwise.

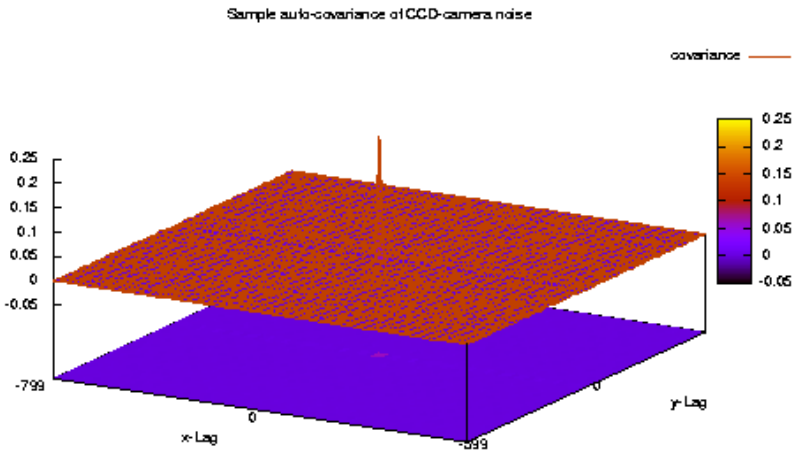



Figure 5.6: Autocorrelation of CDD-camera Noise

This is a dull result, no doubt about it, especially after making an effort to derive a reasonably good estimator. To go at such lengths only to find, in the end, that the noise is as innocently white as can be, was maybe not worth it. A (cold) comfort, then, is the full validation of the model assumption. Just as in the past, we continue to set $\Lambda_\nu = \sigma_\nu^2 I$, but — and this makes a difference after all — henceforth with clear conscience.

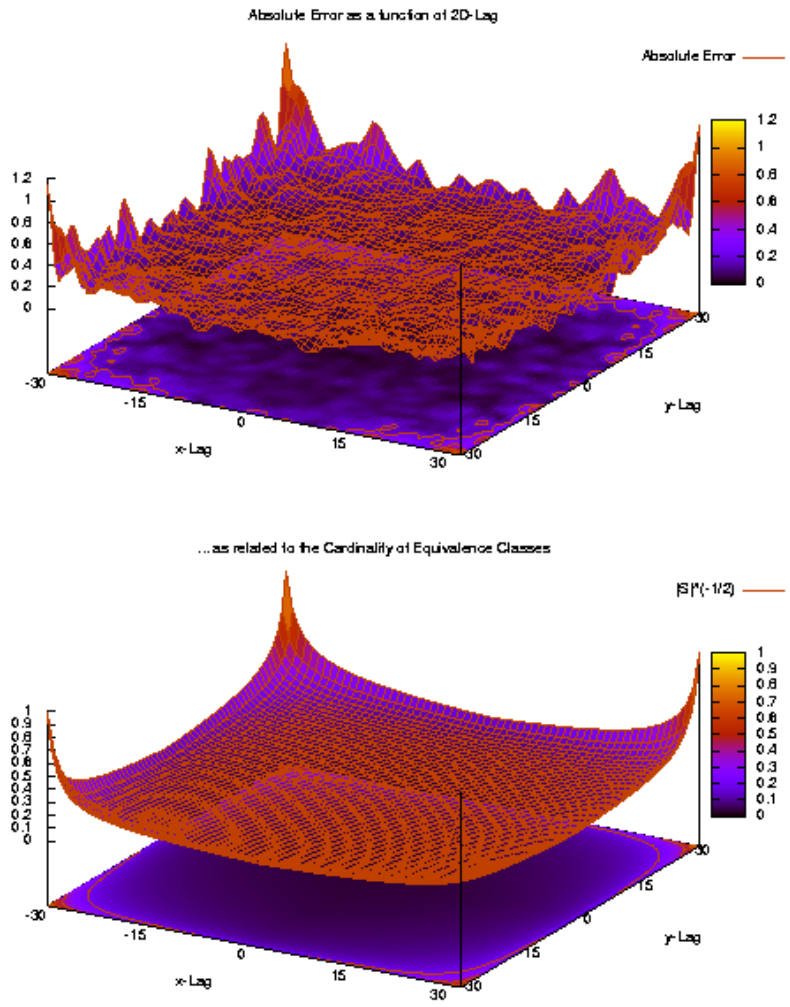



Figure 5.7: Absolute error as a function of 2D-lag and predicted standard deviation

# 6 Evaluation

In this chapter we present and evaluate the results obtained by the implemented algorithms Wiener Filter, Richardson-Lucy, Blind Expectation Maximization and Neelamani, in the following referred to as WF, RL, BEM and NM respectively.

## 6.1 Synthetic Data

**Experiment Setup** In order to measure the quality of restoration by an objective criterion, a first sequence of tests were performed on synthetic data. For this purpose we convolved an authentic and reasonably sharp picture showing blood cells — in the following referred to as $f$ — with the Gaussian OTF

$$h_\theta(r) = \exp\left(-\theta\, r^{\frac{5}{3}}\right) \qquad \theta = 0.005 \tag{6.1}$$

where $r = \sqrt{u^2 + v^2}$ and $(u, v)$ the two-dimensional frequency.

Robustness being an important aspect for the evaluation, different levels of white Gaussian noise were simulated by adding an appropriate random vector $\nu_\sigma$ with variance $\sigma^2$. Conditions range from ideal (i.e noise-free) to worst-case, the latter one represented by a standard deviation $\sigma = 5$ for intensities in $[0, 255]$ or, equivalently, a signal-to-noise-ratio of

$$\mathrm{SNR} = 10 \log_{10} \frac{\|f\|_2^2}{\|\nu\|_2^2} = 20 \log_{10} \frac{\|f\|_2}{\|\nu\|_2} \approx 32.5 \text{ dB} \tag{6.2}$$

The synthetic test-image for a given noise-level $\sigma$ thus is obtained as

$$g^{(\sigma)} = f * h + \nu_\sigma \tag{6.3}$$

For succinctness and to avoid inflating this chapter unnecessarily with dull material of little significance, we restrict ourselves to the extremal cases

$$\sigma = \begin{cases} 0 & \text{blurred} \\ 5 & \text{blurred and noisy} \end{cases} \tag{6.4}$$

The corresponding test-images $g^{(0)}$ and $g^{(5)}$, 512x512 pixels in size, are shown in figure 6.1.

**Error Metric** Each of the four computed estimates $\hat{f} \in \{\hat{f}_{WF}, \hat{f}_{RL}, \hat{f}_{BEM}, \hat{f}_{NM}\}$ was evaluated according to the following metric

$$d(f, \hat{f}) = \left\| \hat{f}(g^{(\sigma)}) - f \right\|_2^2 / \dim(f) \tag{6.5}$$

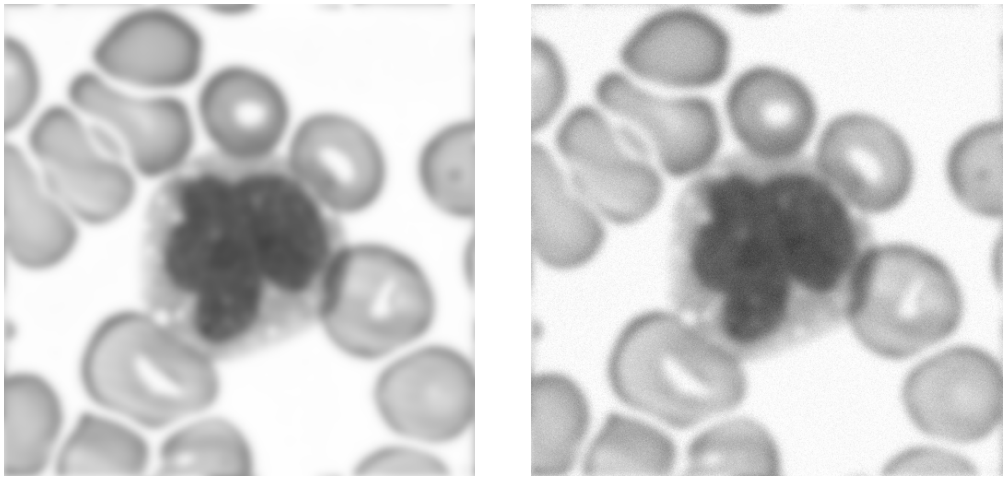


Figure 6.1: Synthetically blurred test-image, noise-free (left) and corrupted (right)

With a term that strikes us as unfortunate, this quantity is sometimes referred to as *mean square error* (MSE), 'mean' denoting average over the two spatial dimensions of the image. To avoid confusion, we stick with the stochastic acceptation of 'mean' as expectation or first moment of a random variable and prefer to speak of *average square error* (ASE) instead. However, what exactly shows up in the denominator — $\|f\|_2^2$ would be another candidate for that matter — is of little consequence so long as we evaluate all competitors on the same set of data.

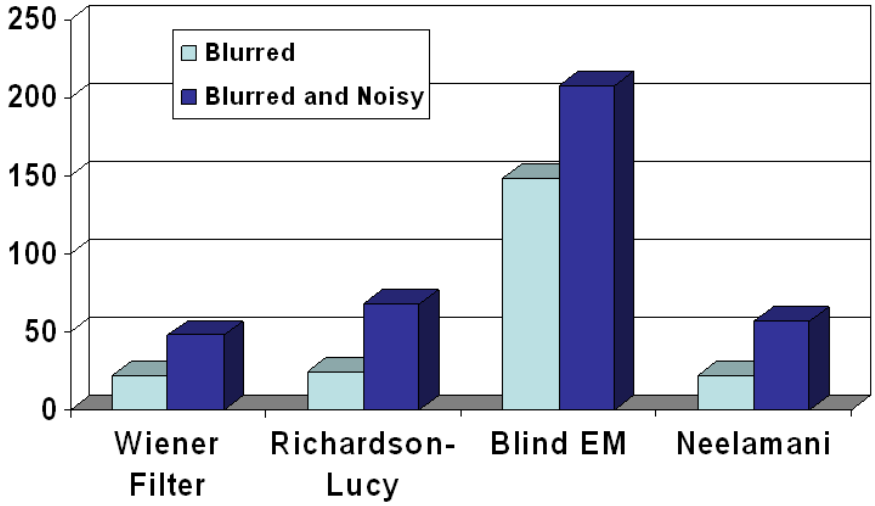



Figure 6.2: ASE Performance (diagram)

| Test-image | WF | RL | BEM | NE |
|---|---|---|---|---|
| Blurred | 22.2589 | 24.3831 | 148.6752 | 21.9385 |
| Blurred& Noisy | 47.903 | 68.1796 | 207.6493 | 56.551 |

Table 6.1: ASE Performance (figures)

**Performance** Figure 6.2 shows the ASE performance for each of the evaluated algorithms. The exact figures may be looked up in table 6.1, while the restored images are given in the appendix, arranged synoptically for convenient assessment.

With an average square error up to seven times as big as the rest, the blind EM algorithm is not competitive. Taking into account the difficulty of estimating the pristine image without precise knowledge of forward-mapping, this result can hardly surprise. As for the three remaining ones, performance under ideal conditions is very similar and does not exhibit marked differences. Though Richardson-Lucy seemingly falls behind with a slightly greater ASE, this is most likely due to premature halting of the algorithm after 50 iterations, which corresponds to approximately 40 seconds processing time on a Pentium 4 with 3Ghz. Convergence, it has already been said, is rather slow and only desirable in the absence of noise. Running more one minute for a greyscale-image of size 512x512 the blind EM algorithm is also the costliest among the four which makes its poor performance all the more disappointing. In this regard the two direct approaches, and in particular the Wiener Filter, are difficult to get ahead of. Relating cost to efficiency, this latter one is indeed unrivalled with roughly 3 seconds of computation time.

All algorithms prove to be fairly robust to noise. Considering the ill-posedness of the inverse problem this is a remarkable result in itself, owed to the regularization techniques discussed in chapter three. It is clear that in a noisy environment loss of information occurs that no algorithm will ever be able to make up for. Complete restoration, then, is not a realistic expectation. But it is the hallmark of a robust algorithm to allow for such loss and to ensure that deterioration is gradual rather than abrupt. In this regard all four algorithms can be said to be well-behaved.

The difference between the two direct approaches — Wiener Filter and Neelamani — was not as pronounced as expected or even completely inexistent in most of the tests that were conducted with microscopic images. The specific advantage of Neelamani over the Wiener Filter being the ability to cope with singularities and sharp contrasts in the image, this potential benefit did not show to advantage for the smooth images used in the experiment.

## 6.2 Real-World Data

Experiments with mock data as in the previous section are legitimate and necessary, allowing an objective quantification of the performance under controlled conditions. The true touchstone for any implemented method, however, will always be its application to the real-world data it was conceived for in the first place. For two reasons, though, evaluation is more difficult. To begin with, there is no pristine image at hand to evaluate against. Any judgement, then, will have to rely exclusively on visual impression, lacking authoritative figures to sustain it. Moreover, the blur-kernel being unknown, a lucky guess often is crucial for success.

In the following we present another sequence of tests conducted with two authentic images featuring different degrees of out-of-focus blur (shown in figure 6.3). One is extremely poor in detail — an intrinsic property worsened by the effect of blur; the other represents the pristine object $f$ from which the test-images of the previous section were generated.

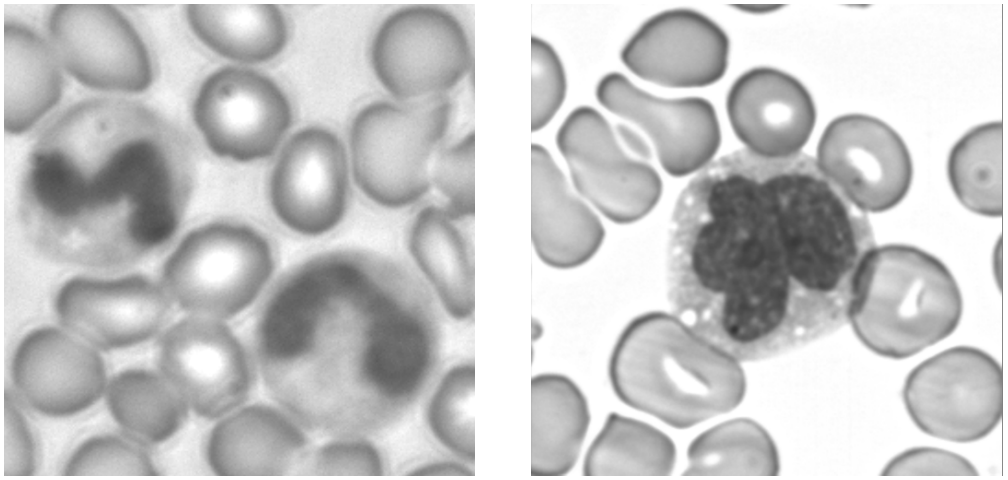


Figure 6.3: Authentic test-images with different degree of out-of-focus blur

For the left image featuring a decent amount of blur a Bessel OTF was assumed with

$$h(r) = \begin{cases} 1 & r = \sqrt{u^2 + v^2} = 0 \\ \dfrac{J_1(0, \theta\, r)}{\theta\, r} & \text{otherwise} \end{cases} \qquad \theta = 0.04 \tag{6.6}$$

with $J_1$ the Bessel function of the first kind. For the nearly sharp image on the right-hand-side we chose the Gaussian OTF already given in (6.1) with parameter $\theta = 0.001$ modelling very modest out-of-focus blur.

The results for the left-hand-side image (see figure A.6 in the appendix) are sobering and cannot convince. For hard we have tried, substantial improvement was not observable. We may take credit for our sincerity by showing the limits of what is feasible. (Why conceal it, after all? The algorithms, it will have been noted, are not ours; we sure won't take the blame for a possible failure). Slightly better are the results for the right-hand-side image, where details could be recovered to a certain, limited, extent. (See figure A.8 in the appendix) Explaining the unequal outcome with a more or less well-fitting PSF respectively, this points out, once again, the crucial importance of a-priori information.

# 7 Conclusion

The purpose of this thesis was to provide some insight into image deconvolution and to review the more important approaches that have crystallized in this area of research over the years. The evaluation, finally, was to investigate their aptitude for application in microscopy, exploring individual strengths and limitations of each. The good news, certainly, is that effective restoration does not necessarily have to be expensive. Rather than mathematical sophistication it is the adequate parameters which turned out to be crucial for success. This result reflects an intuitive truth which we have insistently tried to bring across. In fact, where the solution is not well-determined by the data, an accordingly important role devolves upon the incorporation of a-priori knowledge. While the question of how to enforce regularization constraints is satisfactorily answered in theory, the actual acquisition of such knowledge — the model problem — very often remains challenging in practice. Our contribution, here, was to devise a novel covariance estimator for large-scale WSS random-processes which tackles some of the pratical limitations observed in conventional approaches. This way a more accurate modelling of CCD-camera noise could be achieved.

Clearly, much has evolved since the time where ill-posed or inverse problems were considered mere anomalies of no particular interest. Ever since, starting essentially with Tikhonov, this area of research has witnessed a significant increase in publications. Today, a vast literature is available both off- and online. A Google search for the keyword 'inverse problem' currently lists more than 5 million items — an eloquent number, if this is to be any indication. (This is not just rhetoric. To be sure, we have not checked all sites, but we will not deny owing much to them, either; see the bibliography for some links.)

As always, the linear kind represents a particularly amenable subset within the much more comprehensive class of inverse problems, being easy to handle as much in theory as in practice. Due to their relative simplicity, they are also particularly well-investigated. From the student's perspective, then, there are two sides of the medal. Competing with decades of high-profile research, substantial contributions and scientific progress are comparably difficult to realize. On the other hand, beyond the hunt for patents and glorious publications, it is rewarding for its universality and practical relevance in many real-world-applications.Doubtlessly, image deconvolution and related technologies like computer tomography have had their share in promoting advance, but most of the math is not specific to this field of application. Whether it happens to be an image whose estimate is sought has little importance after all. Extending the scope to non-linear problems it could be the interior of the earth measured by means of seismic waves, the organs of a patient absorbing X-rays in computer tomography and many things more.

In other words, it pays to systematically investigate the abstract concepts behind a particular algorithm, and dedicate an adequate amount of time and space elaborating them. We hope, in this regard, that the third chapter on robust parameter estimation is more than just a lengthy digression. Our purpose, there, was to embed image deconvolution within its

broader mathematical context. If this should have transcended the immediate use case and eventually permit a generalization to other fields of applications, we think it was worthwhile.

# A Appendix

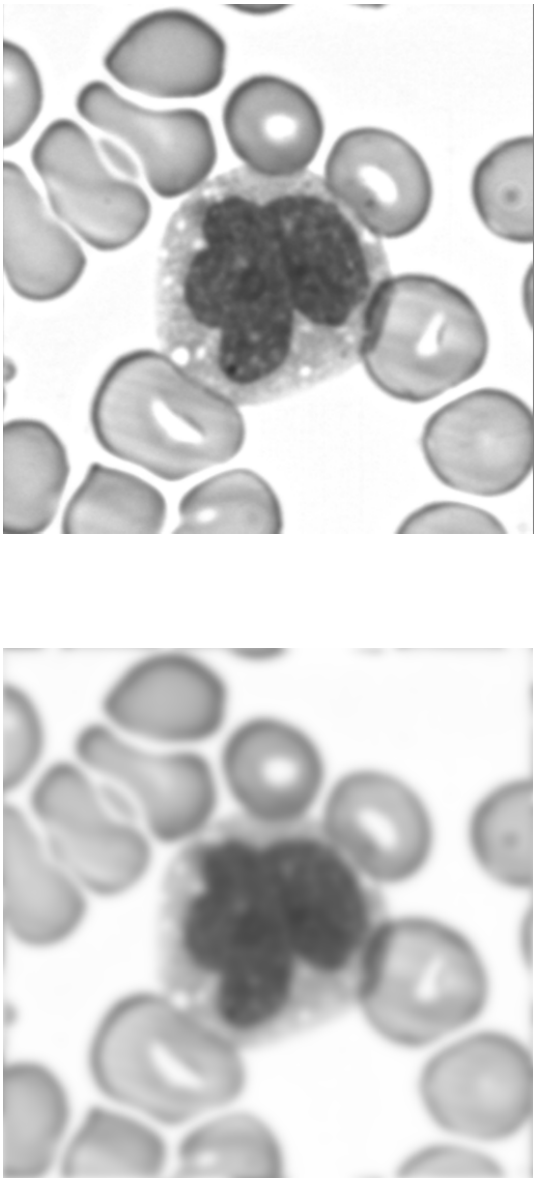
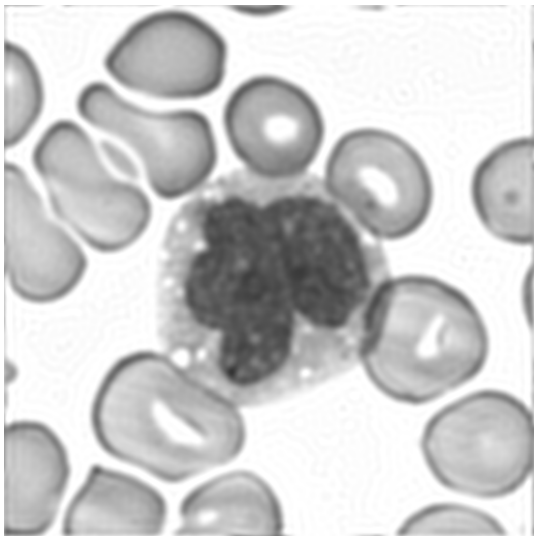
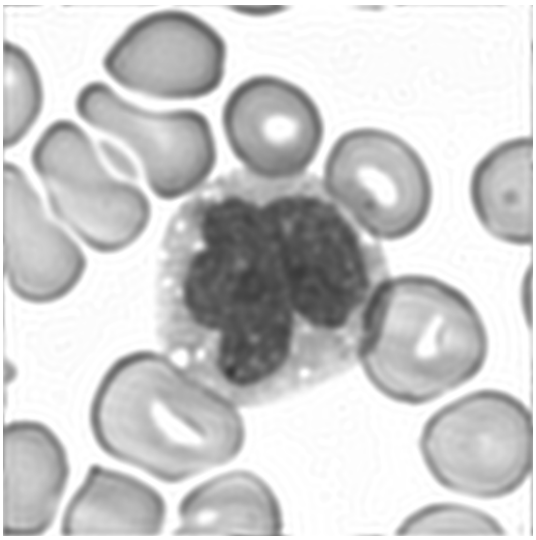


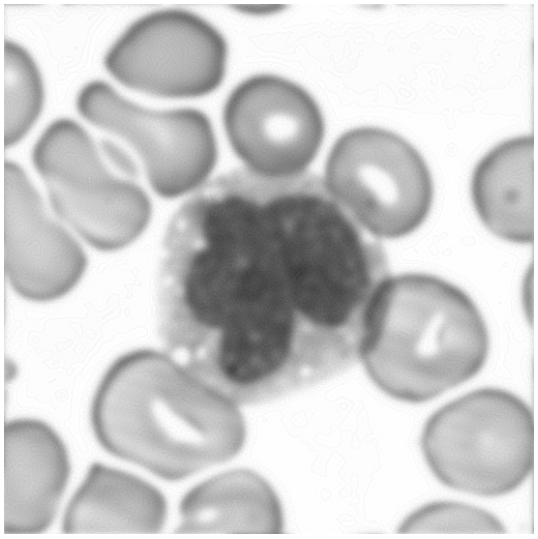
Figure A.1: Above: pristine image, below: blurred



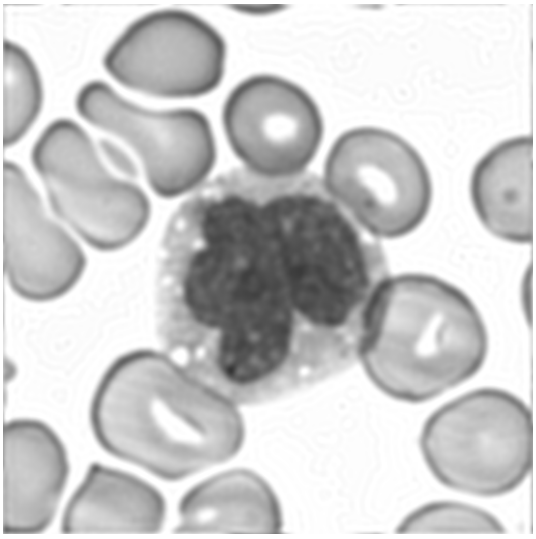
Wiener Filter (WF)
ASE = 22.2589



Richardson-Lucy (RL)
ASE = 24.3831



Blind Expectation Maximization (BEM)
ASE = 148.6752



Neelamani (NM)
ASE = 21.9385

Figure A.2: Estimates for mock data with noise level $\sigma = 0$

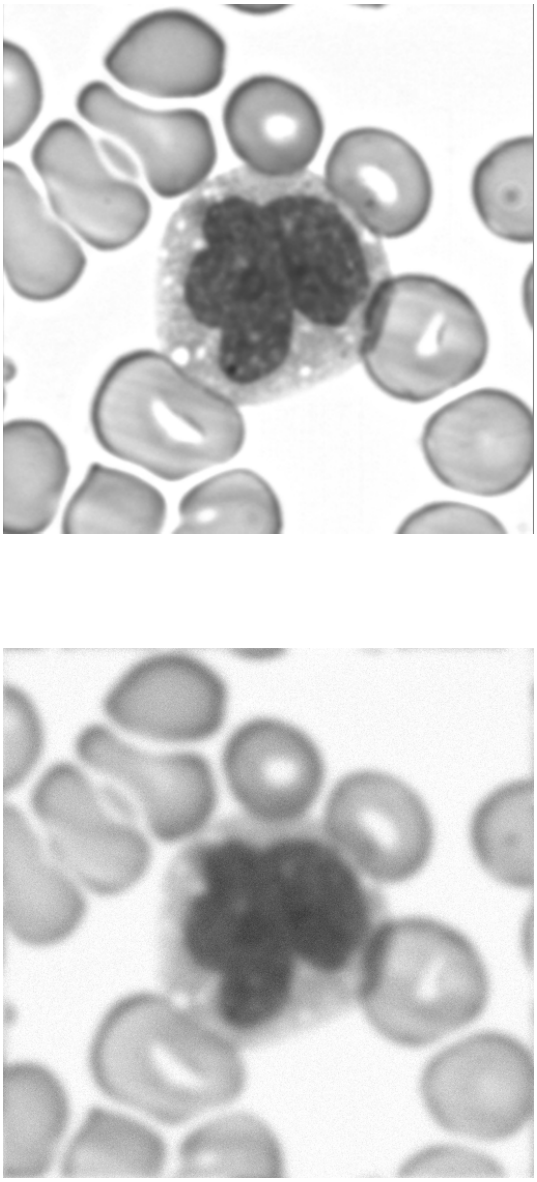
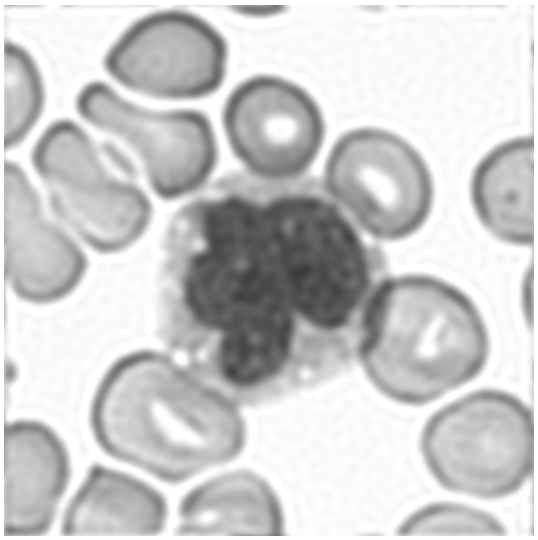
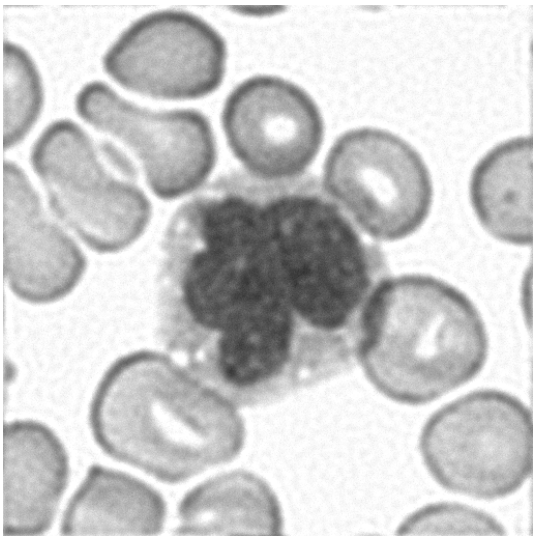


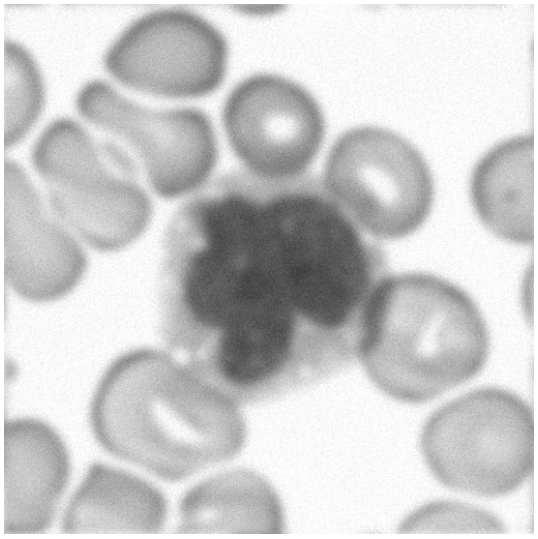
Figure A.3: Above: pristine image, below: blurred and noisy



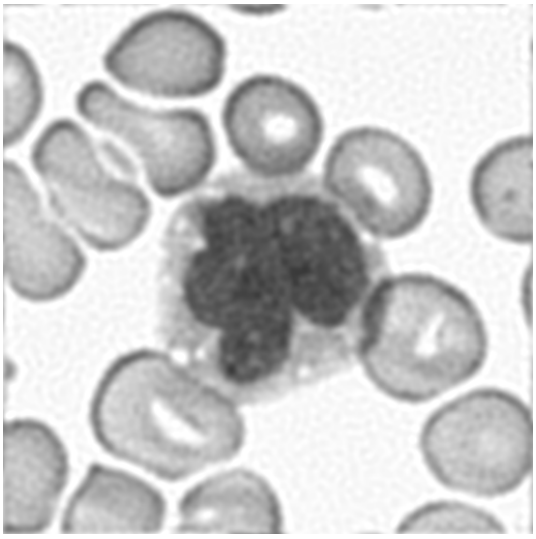
Wiener Filter (WF)
ASE = 47.903



Richardson-Lucy (RL)
ASE = 68.1796



Blind Expectation Maximization (BEM)
ASE = 207.6493



Neelamani (NM)
ASE = 56.551

Figure A.4: Estimates for mock data with noise level $\sigma = 5$

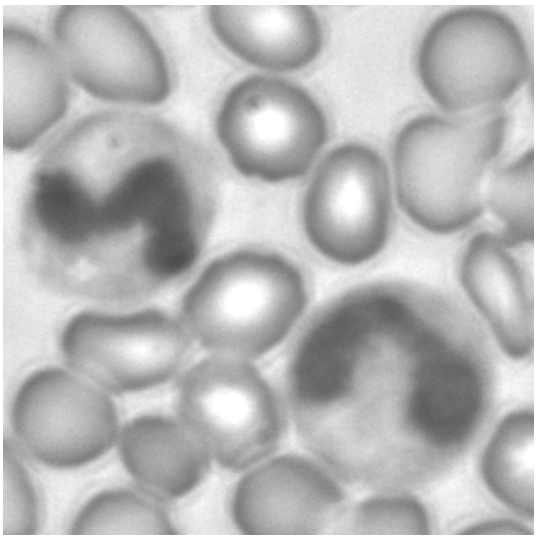
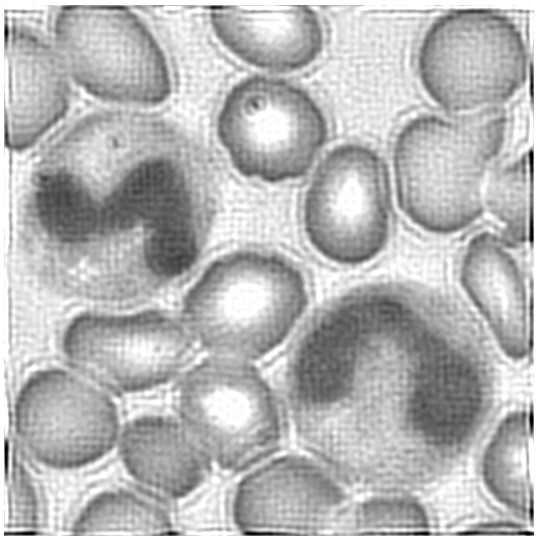
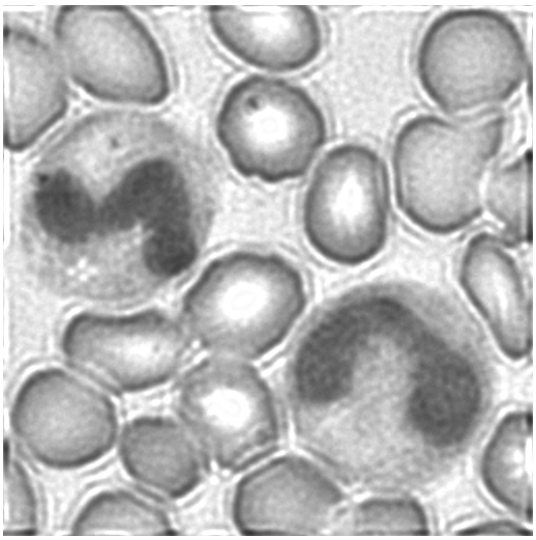


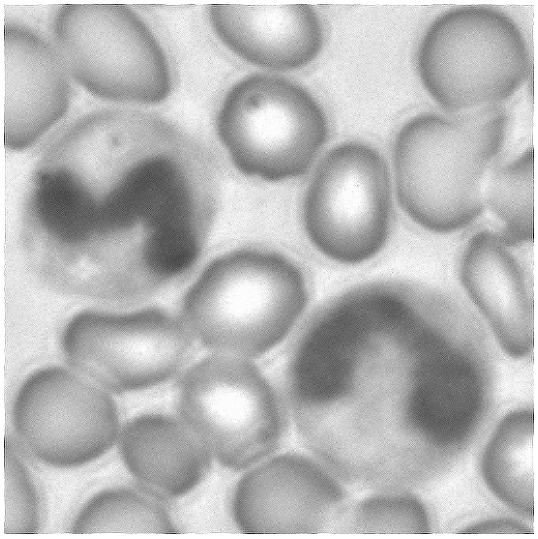
Figure A.5: Recorded image (strongly degraded)



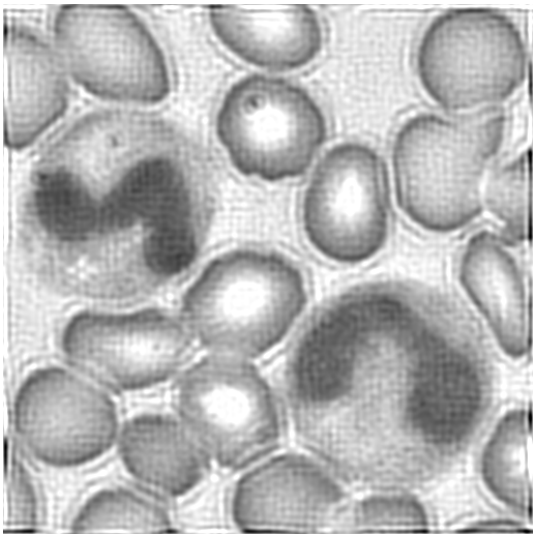
Wiener Filter (WF)



Richardson-Lucy (RL)



Blind Expectation Maximization (BEM)



Neelamani (NM)

Figure A.6: Estimates for real-world data (1)

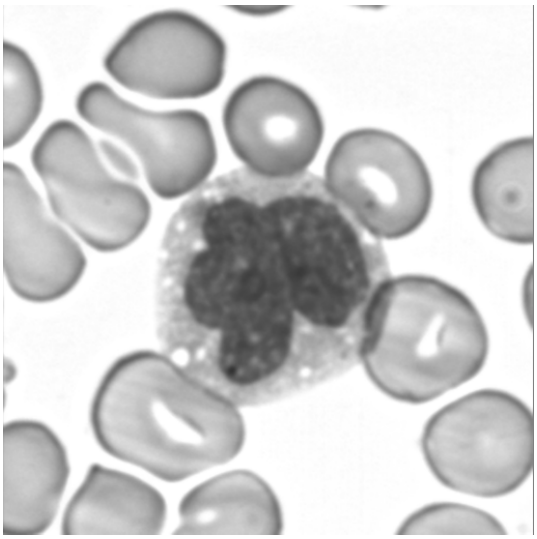
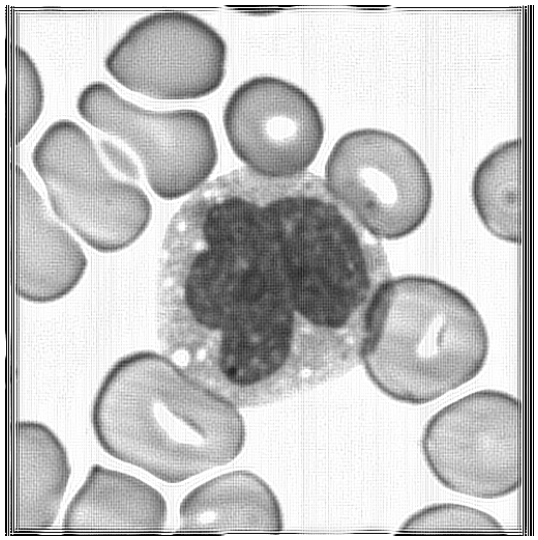
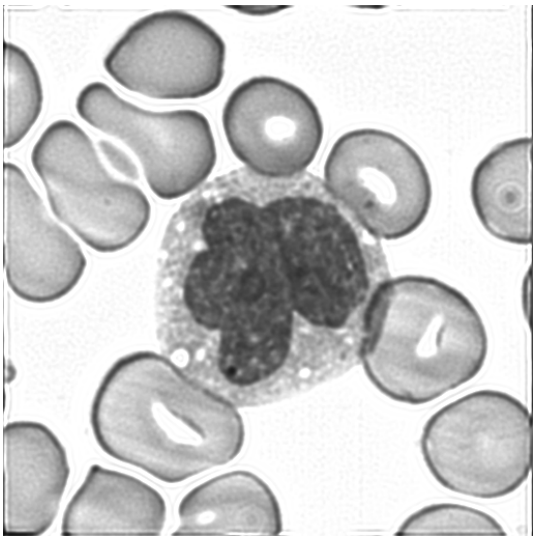


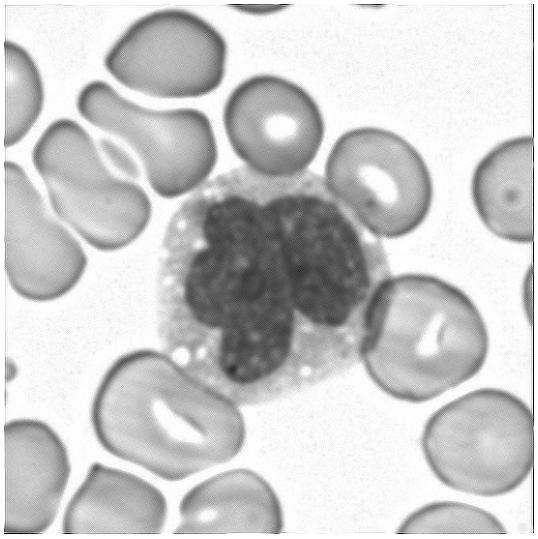
Figure A.7: Recorded image (small amount of blur)



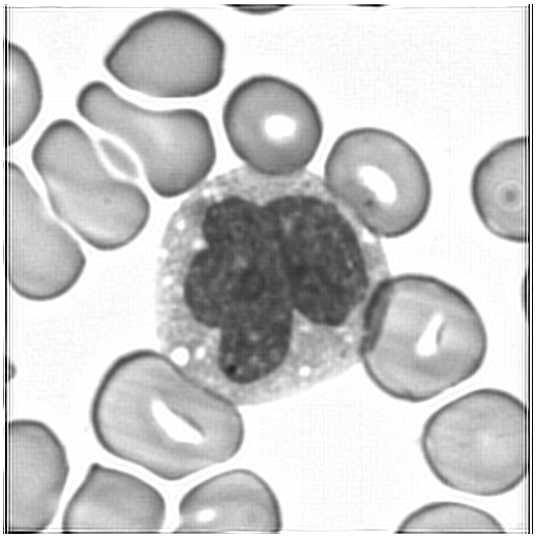
Wiener Filter (WF)



Richardson-Lucy (RL)



Blind Expectation Maximization (BEM)



Neelamani (NM)

Figure A.8: Estimates for real-world data (2)

# List of Figures

# List of Tables

# Bibliography


[1] D. S. C. Biggs and M. Andrew, *Acceleration of iterative image restoration algorithms*, Applied Optics, 36 (1997), pp. 1766–1775.

[2] P. J. Bones, C. R. Parker, B. L. Satherley, and R. W. Watson, *Deconvolution and phase retrieval with use of zero-sheets*, Journal of the Optical Society of America, 12 (1995), pp. 1842–1857.

[3] A. Bttcher, B. Hofmann, U. Tautenhahn, and M. Yamamoto, *Convergence rates for Tikhonov regularization from different kinds of smoothness conditions*, Applicable Analysis, (2006), pp. 555–578.

[4] M. Cignoni, *Star formation rate in the solar neighborhood*, PhD thesis, Universit degli Studi di Pisa, 2006.

[5] H. W. Engl and R. S. Anderssen, *The role of Linear Functionals in Improving Convergence Rates for Parameter Identification via Tikhonov Regularization*, Tech. Rep. 427, University of Linz, December 1990.

[6] D. C. Ghiglia, L. A. Romero, and G. A. Mastin, *Systematic approach to two-dimensional blind deconvolution by zero-sheet separation*, Journal of the Optical Society of America, 10 (1993), pp. 1024–1036.

[7] S. F. Gibson and F. Lanni, *Experimental test of an analytical model of aberration in an oil-immersion objective lens used in three-dimensional light microscopy*, Journal of the Optical Society of America, 8 (1991), pp. 1601–1613.

[8] G. H. Golub and U. von Matt, *Generalized Cross-Validation for Large Scale Problems*, Tech. Rep. TR-96-28, ETH, SCSC, September 1996.

[9] E. Haber, *Numerical Strategies for the Solution of Inverse Problems*, PhD thesis, University of British Columbia, 1997.

[10] P. C. Hansen, *The L-Curve and its Use in the Numerical Treatment of Inverse Problems*, in Computational Inverse Problems in Electrocardiology, no. 5 in Advances in Computational Bioengineering, WIT Press, Southampton, 2001, pp. 119–142.

[11] B. Hofmann, *Approximate source conditions in Tikhonov-Phillips regularization and consequences for inverse problems with multiplication operators*, Mathematical Methods in the Applied Sciences, (2006), pp. 351–371.

[12] E. T. Jaynes, *Information Theory and Statistical Mechanics*, The Physical Review, 106 (1957), pp. 620–630. `http://bayes.wustl.edu/etj/articles/theory.1.pdf`.

[13] A. K. Katsaggelos, *Digital Image Restoration*, Springer Verlag, Berlin, Heidelberg, 2nd ed., 1991. Springer Series in Information Sciences, No. 23.

[14] A. K. Katsaggelos and K. T. Lay, *Maximum Likelihood Blur Identification and Image Restoration Using the EM Algorithm*, IEEE Transactions on Signal Processing, 39 (1991), pp. 729–733.

[15] D. Kundur, *Blind Deconvolution of Still Images using Recursive Inverse Filtering*, May 1995. http://www.ece.tamu.edu/~deepa/pdf/masc95.pdf.

[16] D. Kundur and D. Hatzinakos, *Blind Image Deconvolution*, IEEE Signal Processing Magazine, 13 (1996), pp. 43–64.

[17] D. MacKay, *Information Theory, Inference, and Learning Algorithms*, Cambridge Press, 2003.

[18] T. K. Moon, *The Expectation-Maximization Algorithm*, IEEE Signal Processing Magazine, (1996), pp. 47–60.

[19] R. Neelamani, H. Choi, and R. Baraniuk, *ForWaRD: Fourier-Wavelet Regularized Deconvolution for Ill-Conditioned Systems*, IEEE Transactions on Signal Processing, 52 (2004), pp. 418–433.
http://www.dsp.rice.edu/publications/pub/neelshdecon.pdf.

[20] H. J. Nussbaumer, *Fast Fourier Transform and Convolutional Algorithms*, Springer Verlag, Berlin, Heidelberg, 2nd ed., 1990. Springer Series in Information Sciences, No. 2.

[21] T. Oliphant. Lecture notes from a course on Inverse Problems at Brigham Young University, 2006. http://www.et.byu.edu/groups/ece771web/.

[22] S. Reeves and R. Mersereau, *Blur identification by the method of generalized cross-validation*, IEEE Transactions on Image Processing, 1 (1992), pp. 301–311.

[23] A. E. Savakis and J. Trussell, *Blur Identifikation by Residual Spectral Matching*, IEEE Transactions on Image Processing, 2 (1993), pp. 141–151.

[24] C. E. Shannon, *A Mathematical Theory of Information*, The Bell System Technical Journal, 27 (1948), pp. 379–423.
http://cm.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf.

[25] A. Tarantola, *Inverse Problem Theory and Methods for Model Parameter Estimation*, SIAM, 1st ed., 2005.

[26] R. Vio, J. Bardsley, and W. Wamsteker, *Least-Squares methods with Poissonian noise: an analysis and a comparison with the Richardson-Lucy algorithm*, Astronomy and Astrophysics, (2005).
http://web.math.umt.edu/bardsley/papers/Lspoisson04.pdf.

[27] C. R. Vogel, *Computational Methods for Inverse Problems*, SIAM, 2002.

[28] G. Wahba, *Practical Approximate Solutions to Linear Operator Equations when the Data are noisy*, SIAM Journal of Numerical Analysis, 14 (1977), pp. 651–667.