# Chapter 2

# Regularization

ch,reg

## Contents

s,reg,intro

## 2.1 Introduction (s,reg,intro)

The previous chapter on image restoration described some basic methods for **regularization** of **ill-posed inverse problems**. This chapter describes several regularization methods, including implementation details.

The subject of regularization dates back at least to the early work of Phillips [1], Tikhonov [2] and Miller [3]. Survey papers on the topic include [4, 5]. There are also several related books, including [6, 7] and [8, §5.1]. Software tools are also available, *e.g.*, [9].

The sections in this chapter address different aspects of regularization, and are largely independent.

s,reg,spline

## 2.2 Splines and nonparametric function estimation (s,reg,spline)

The desirability of regularization can be illustrated by considering the following simple problem, known as **nonparametric regression** or **nonparametric function estimation**. Suppose we measure the value of a function (or signal) $f(t)$ at several distinct points $t_1, \dots, t_{n_\mathrm{d}}$ with measurement error:

e,spline,yi

$$y_i = f(t_i) + \varepsilon_i, \qquad i = 1, \dots, n_\mathrm{d}, \tag{2.2.1}$$

where the measurement noise is independent and, for simplicity, normally distributed: $\varepsilon_i \sim \mathsf{N}(0, \sigma^2)$. We would like to estimate the function $f(\cdot)$ from the measurements $\boldsymbol{y} = (y_1, \dots, y_{n_\mathrm{d}})$.

### 2.2.1 ML estimation / interpolation

For gaussian measurement errors, maximum-likelihood estimation of $f$ corresponds to the following minimization problem:

$$\hat{f} = \arg\min_f \sum_{i=1}^{n_\mathrm{d}} \frac{1}{2} |y_i - f(t_i)|^2 .$$

However, there is an infinite collection of choices $\hat{f}$ that fit the data exactly, *i.e.*, for which $y_i = \hat{f}(t_i), \ \forall i$. So the ML criterion does not specify a unique estimate. This 1D example is a classic under-determined problem.

In many cases, we expect $f$ to be a smooth function. So one method for choosing among the many ML estimates is to select the $\hat{f}$ that has minimal roughness. A reasonable roughness measure is the energy of one of its derivatives [1, 10, 11]:

e,spline,fh,interp,cost

$$\hat{f} = \arg\min_f \int \left| f^{(m)}(t) \right|^2 \mathrm{d}t \tag{2.2.2}$$

e,spline,yi=f(ti)

$$\text{s.t.} y_i = f(t_i), \quad i = 1, \dots, n_\mathrm{d}, \tag{2.2.3}$$

where $f^{(m)}$ denotes the $m$th derivative of $f$. The questions then become: (i) how does one compute $\hat{f}$, (ii) what are the properties of $\hat{f}$, (iii) how should we choose $m$, and (iv) are there better measures than (2.2.2)?

The Euler-Lagrange equation for the variational problem (2.2.2) is [12–14]

$$\hat{f}^{(2m)}(t) = \sum_{i=1}^{n_\mathrm{d}} \lambda_i \, \delta(t - t_i),$$

where $\delta(\cdot)$ denotes the Dirac impulse. The $\lambda_i$ values are Lagrange multipliers that one must choose to satisfy the constraints (2.2.3). Integrating this equation $2m$ times yields the following expression for $\hat{f}$:

e,spline,fh,plus

$$\hat{f}(t) = \sum_{k=0}^{2m-1} c_k \frac{1}{k!} t^k + \sum_{i=1}^{n_\mathrm{d}} \lambda_i \frac{1}{(2m-1)!} \left[ t - t_i \right]_+^{2m-1}, \tag{2.2.4}$$

where $[t]_+$ equals $t$ if $t > 0$ and is otherwise zero. The $c_k$ values denote $2m$ free coefficients that one must select based on the desired boundary conditions, *i.e.*, the desired behavior of $\hat{f}$ for $t < t_1$ and $t > t_{n_\mathrm{d}}$. The usual choice is to require that $\hat{f}^{(n)}(t) = 0$ for all $t < t_1$ for $n \geq m$, which implies that $c_m = c_{m+1} = \cdots = c_{2m-1} = 0$. In addition, requiring $\hat{f}^{(n)}(t) = 0$ for all $t > t_{n_\mathrm{d}}$ for $n \geq m$ implies that $0 = \sum_{i=1}^{n_\mathrm{d}} \lambda_i t_i^k$, for $k = 0, \dots, m-1$. Therefore, we can determine $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_{n_\mathrm{d}})$ and $\boldsymbol{c} = (c_0, \dots, c_{m-1})$ by solving the following $(n_\mathrm{d} + m) \times (n_\mathrm{d} + m)$ system of equations

e,spline,interp,sys

$$\begin{bmatrix} \boldsymbol{A} & \boldsymbol{C} \\ \boldsymbol{T} & \boldsymbol{0}_{m \times m} \end{bmatrix} \begin{bmatrix} \boldsymbol{\lambda} \\ \boldsymbol{c} \end{bmatrix} = \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{0}_{m \times 1} \end{bmatrix}, \tag{2.2.5}$$

where $\boldsymbol{0}_{m \times n}$ denotes the $m \times n$ array zeros, $\boldsymbol{A}$ is the lower triangular, $n_\mathrm{d} \times n_\mathrm{d}$ matrix with elements $A_{il} = \frac{1}{(2m-1)!} \left[ t_i - t_l \right]_+^{2m-1}$, $\boldsymbol{C}$ is the $n_\mathrm{d} \times m$ matrix with elements $C_{ik} = \frac{1}{k!} t_i^k$, and $\boldsymbol{T}$ is the $m \times n_\mathrm{d}$ matrix with elements $T_{ki} = t_i^k$, for $k = 0, \dots, m-1$. Applying the transpose of the bracketed matrix to both sides yields

e,spline,interp,sys,sq

$$\begin{bmatrix} \boldsymbol{A}'\boldsymbol{A} + \boldsymbol{T}'\boldsymbol{T} & \boldsymbol{A}'\boldsymbol{C} \\ \boldsymbol{C}'\boldsymbol{A} & \boldsymbol{C}'\boldsymbol{C} \end{bmatrix} \begin{bmatrix} \boldsymbol{\lambda} \\ \boldsymbol{c} \end{bmatrix} = \begin{bmatrix} \boldsymbol{A}'\boldsymbol{y} \\ \boldsymbol{C}'\boldsymbol{y} \end{bmatrix}. \tag{2.2.6}$$

Using the block inverse formula (26.1.11), the solution is:

$$\begin{bmatrix} \hat{\boldsymbol{\lambda}} \\ \hat{\boldsymbol{c}} \end{bmatrix} = \begin{bmatrix} \left[ \boldsymbol{A}'\mathcal{P}_{\boldsymbol{C}}^{\perp}\boldsymbol{A} + \boldsymbol{T}'\boldsymbol{T} \right]^{-1} & -\boldsymbol{A}'\boldsymbol{A}\boldsymbol{C}\boldsymbol{\Delta}^{-1} \\ -\boldsymbol{\Delta}^{-1}\boldsymbol{C}\boldsymbol{A}\boldsymbol{A}' & \boldsymbol{\Delta}^{-1} \end{bmatrix},$$

where $\mathcal{P}_{\boldsymbol{C}}^{\perp} = \boldsymbol{I} - \boldsymbol{C} \left[ \boldsymbol{C}'\boldsymbol{C} \right]^{-1} \boldsymbol{C}'$ and the **Schur complement** is $\boldsymbol{\Delta} = \boldsymbol{C}'\boldsymbol{C} - \boldsymbol{C}'\boldsymbol{A} \left[ \boldsymbol{A}'\boldsymbol{A} + \boldsymbol{T}'\boldsymbol{T} \right]^{-1} \boldsymbol{A}'\boldsymbol{C}$. However, this simple approach is poorly conditioned and not recommended for implementation.

The solution $\hat{f}$ in (2.2.4) is called a **spline** of degree $2m-1$; it is a piece-wise polynomial with a **knot** at each $t_i$. In between the knots, $\hat{f}$ is a polynomial of degree $2m-1$. At each knot, $\hat{f}$ and its first $2m-2$ derivatives are continuous. The usual choice is $m=2$, in which case $\hat{f}$ is called the **cubic spline interpolator**. In this case, one can derive the solution $\hat{f}$ without applying the calculus of variations [11, Ch. 2]. A simple derivation based on Fourier transforms is also available [15].

Mat `spapi`

Fig. 2.2.1 illustrates spline interpolators for $m=0,1,2$, for an example with noisy samples where $\sigma=1$ and $n_\mathrm{d}=80$. As this example shows, the cubic spline interpolant, which is one of many possible "ML estimates," oscillates excessively for noisy data. Even though we penalized roughness in (2.2.2), the requirement in (2.2.3) that the estimate $\hat{f}$ interpolate the data *exactly* causes wild oscillations because it is fitting the noise.

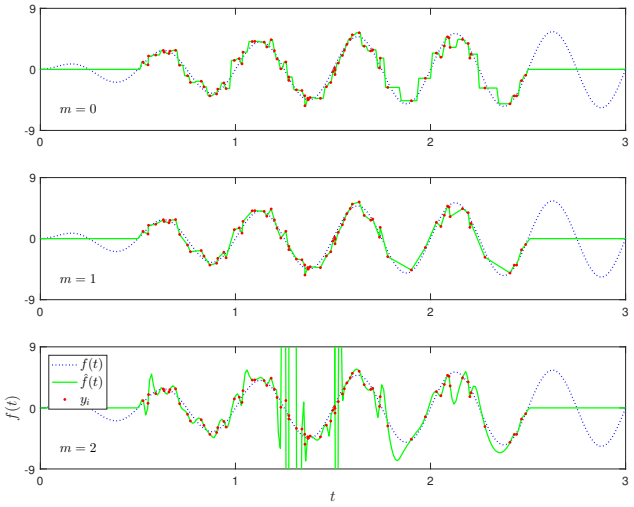



Figure 2.2.1: Spline interpolation of noisy data for $m=0$ (nearest neighbor), $m=1$ (linear interpolation), and $m=2$ (cubic spline). The noisy samples $\{y_i\}$ are the red points, the true function $f(t)$ is the dashed blue curve, and the interpolator $\hat{f}(t)$ is the solid green curve. fig_spline_interp

An alternative to (2.2.2) is to replace the $\mathcal{L}_2$ norm of $f^{(m)}$ with the $\mathcal{L}_1$ norm, which is related to its **total variation** (**TV**) when $m=1$ [16].

## 2.2.2 B-splines

Although the form of the solution (2.2.4) arises naturally from the Euler-Lagrange equation, the system of equations (2.2.5) is unstable for large $m$ due to the nature of the unbounded one-sided polynomials $[t]_+^{2m-1}$. Fortunately, there are alternative bases for the space of spline functions. In particular, any spline of the form (2.2.4) can be written

$$\hat{f}(t) = \sum_{k=1}^{n_\mathrm{d}} \alpha_k b_k(t), \tag{2.2.7}$$ e,spline,basis,interp

on the interval $[t_1, t_{n_\mathrm{d}}]$. Each $b_k$ is a **B-spline**, a spline of degree $2m-1$ that is supported on the *finite* interval $[t_{j-m}, t_{j+m}]$ with knots at each of the $t_i$ values in that interval. Because of this finite and local support, there are stable methods for computing the B-spline interpolation coefficients [14].

For equally spaced knots, *i.e.*, $t_i - t_{i-1} = \Delta$, a **B-spline of degree** $n$ is simply the convolution of $n+1$ rect functions:

$$b_k(t) = \left( \underbrace{\mathrm{rect}\left(\frac{\cdot}{\Delta}\right) * \cdots * \mathrm{rect}\left(\frac{\cdot}{\Delta}\right)}_{n+1 \text{ times}} \right)(t - k\Delta)\,.$$

For example, for $n = 2m-1$ with $m=1$, each B-spline is a triangle function, resulting in linear interpolation.

Note that we began this discussion without any assumptions about polynomials or splines. We chose the cost function in (2.2.2), a measure of the bending energy of a thin rod, and the *solution* turned out to be a spline. And then it was found that splines can be expressed in the form (2.2.7). This suggests that splines are inherently natural tools for problems with smoothness constraints. Indeed, the series representation (2.2.7) is used even in problems with more complicated models than (2.2.1) where the variational solution may be intractable.

### 2.2.3 PL estimation / smoothing

For problems with noisy data, a preferable alternative to interpolation is to relax the requirement that $\hat{f}$ fit the noisy data exactly, and instead find an estimate that compromises between data fit and smoothness. A natural way to compromise between such conflicting goals is to minimize a cost function that is a weighted sum of two terms, such as the following **penalized least-squares** criterion:

$$\hat{f} = \arg\min_f \frac{1}{n_\mathrm{d}} \sum_{i=1}^{n_\mathrm{d}} \frac{1}{2} |y_i - f(t_i)|^2 + \beta \int \frac{1}{2} \left| f^{(m)}(t) \right|^2 \mathrm{d}t, \tag{2.2.8}$$

e,spline,fh,pls

where $\beta$ is a **regularization parameter** (or **smoothing parameter** or **hyper-parameter**) that controls the trade-off between data fit and roughness. This type of penalized-LS estimator is known as **nonparametric regression** or **nonparametric function estimation**, because we have not assumed any parametric model for $f$. The generalization to 2D is known as **surface recovery** or **surface interpolation** in computer vision, *e.g.*, [17]. See [18] for related $\ell_1$ versions of **trend filtering**.

Again it follows from the Euler-Lagrange equations that the unique minimizer $\hat{f}$ is a spline of degree $2m-1$. In the usual case where $m = 2$, this method is called **cubic spline smoothing**. Again the form (2.2.7) is applicable, and there is a simple linear relationship between the coefficients of that spline and the data $\boldsymbol{y}$ [10]. In particular, the roughness penalty (2.2.2) is a quadratic function of the spline coefficients, *i.e.*,

$$\int \left| \hat{f}^{(m)}(t) \right|^2 = \|\boldsymbol{C}\boldsymbol{\alpha}\|^2, \tag{2.2.9}$$

e,spline,deriv,norm

for some matrix $\boldsymbol{C}$ with $n_\mathrm{d}$ columns and approximately $n_\mathrm{d}$ rows, where $\boldsymbol{\alpha}$ denotes the B-spline coefficients in (2.2.7).

As a concrete example, if $m = 1$, then the basis functions in (2.2.7) are 1st-degree splines. In the unit-spaced case with $t_i = i$, the basis functions are $b_k(t) = \mathrm{tri}(t-k) = \mathrm{rect}(t) * \mathrm{rect}(t-k)$ which has the following derivative:

$$\frac{\mathrm{d}}{\mathrm{d}t} b_k(t) = \mathrm{rect}(t-k+1/2) - \mathrm{rect}(t-k-1/2).$$

So $\hat{f}^{(1)}(t) = \sum_{k=1}^{n_\mathrm{d}} \alpha_k \left[\mathrm{rect}(t-k+1/2) - \mathrm{rect}(t-k-1/2)\right]$ and it follows from a small calculation that (2.2.9) holds with $\boldsymbol{C}$ defined to be the following $(n_\mathrm{d}+1) \times n_\mathrm{d}$ differencing matrix (*cf.* (1.8.7)):

$$\boldsymbol{C} = \begin{bmatrix} 1 & 0 & 0 & 0 & \dots & 0 \\ -1 & 1 & 0 & 0 & \dots & 0 \\ 0 & -1 & 1 & 0 & \dots & 0 \\ 0 & 0 & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & -1 & 1 & 0 \\ 0 & \dots & 0 & 0 & -1 & 1 \\ 0 & \dots & 0 & 0 & 0 & 1 \end{bmatrix}.$$

The penalty Hessian $\boldsymbol{C}'\boldsymbol{C}$ is tri-diagonal with elements $\{-1, 2, -1\}$. Because in this case of unit-spaced knots we have $f(t_i) = \alpha_i$, we can rewrite (2.2.8) as follows:

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \frac{1}{n_\mathrm{d}} \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{\alpha}\|^2 + \beta \frac{1}{2} \boldsymbol{\alpha}' \boldsymbol{C}' \boldsymbol{C} \boldsymbol{\alpha}. \tag{2.2.10}$$

e,spline,smooth,valf

The solution is

$$\hat{\boldsymbol{\alpha}} = \left[\boldsymbol{I} + n_\mathrm{d} \beta \boldsymbol{C}'\boldsymbol{C}\right]^{-1} \boldsymbol{y}.$$

There are fast algorithms for solving such banded systems of equations, even for the more complicated case of nonuniform knot spacing and/or $m > 1$ [13, 19].

Mat `spaps`

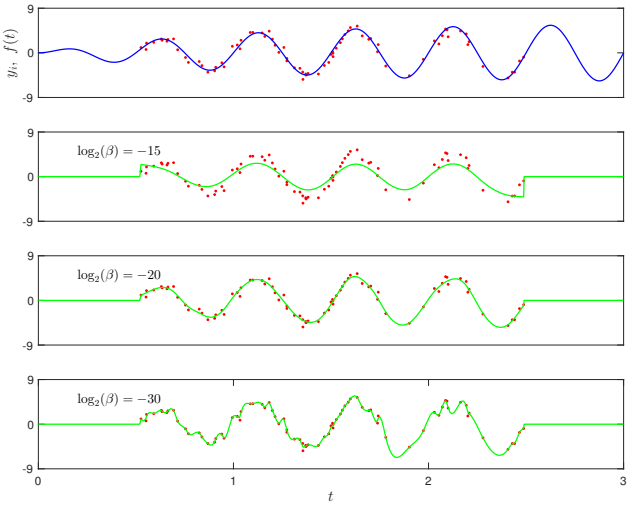
Fig. 2.2.2 illustrates cubic spline smoothing for a range of values of the regularization parameter $\beta$. As $\beta \to 0$, the estimator will approach the spline interpolator. As $\beta \to \infty$, the estimator will approach the best-fit line (for $m = 2$). Automatic methods for selecting $\beta$ have been studied extensively [10].

### 2.2.4 Parametric function estimation

For models that are more complicated than (2.2.1), the variational solution can be intractable so one may need to use parametric approach instead of the nonparametric approach in (2.2.8). Motivated by (2.2.7), a natural approach is to parameterize $f$ *at the outset* using a linear combination of basis functions:

$$f(t) \approx \sum_{j=1}^{n_\mathrm{p}} x_j \, b_j(t), \tag{2.2.11}$$

e,spline,basis




fig_spline_smooth

Figure 2.2.2: Cubic spline smoothing of noisy data for various regularization parameter values.

where $b_j(t)$ denotes the $j$th basis function (chosen by the algorithm designer). Now the problem is to estimate the unknown coefficients $\boldsymbol{x} = (x_1, \ldots, x_{n_\mathrm{p}})$ from the data $\boldsymbol{y}$. To relate the data $\boldsymbol{y}$ to the coefficients $\boldsymbol{x}$, note that

e,reg,spline,A

$$\mathsf{E}[y_i] = f(t_i) \approx \sum_{j=1}^{n_\mathrm{p}} x_j b_j(t_i) = [\boldsymbol{A}\boldsymbol{x}]_i \,, \tag{2.2.12}$$

where $a_{ij} = b_j(t_i)$. So we have the ordinary linear model

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{\varepsilon},$$

with corresponding ML or LS estimate

e,spline,ls

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|^2 = [\boldsymbol{A}'\boldsymbol{A}]^{-1}\boldsymbol{A}'\boldsymbol{y}. \tag{2.2.13}$$

When the number of parameters is small, *i.e.*, $n_\mathrm{p} \ll n_\mathrm{d}$, usually the LS estimate is stable for reasonable choices of basis functions. However, if $n_\mathrm{p}$ is too small, then the approximation (2.2.11) will be poor. So for an accurate approximation to $f$, we would like to increase $n_\mathrm{p}$. But when $n_\mathrm{p} \approx n_\mathrm{d}$, the LS estimate becomes unstable, and if $n_\mathrm{p} > n_\mathrm{d}$ then the problem is under determined. Choosing a **model order** like $n_\mathrm{p}$ is another extensively studied problem [20–26].

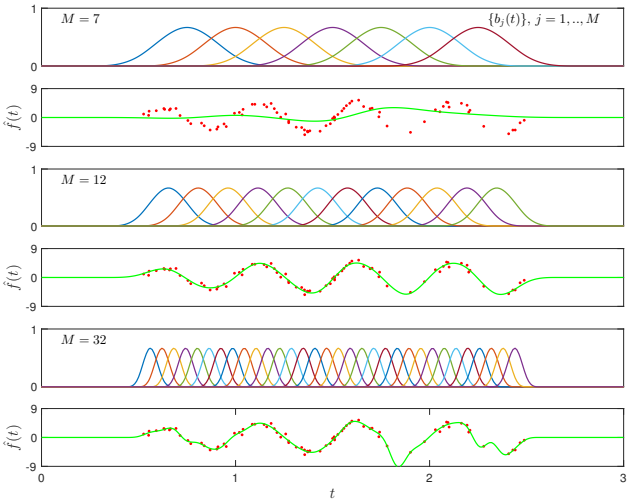



Figure 2.2.3: Cubic B-spline regression of noisy data. Noisy samples $y_i$ shown in red dots, interpolator $\hat{f}(t)$ in green, for several values of M $= n_\mathrm{p}$.

fig_spline_regress

Fig. 2.2.3 illustrates B-spline regression for various values of $n_\mathrm{p}$. For large $n_\mathrm{p}$, the estimate becomes oscillatory, much like the spline interpolator.

### 2.2.5 Penalized B-spline fits with fine parameterization

Now we present a final alternative that is the most analogous to what is done in image reconstruction. To ensure a reasonable approximation to $f$, we want to use many narrow basis functions (*e.g.*, small pixels), so we want $n_\mathrm{p}$ to be large, *i.e.*, $n_\mathrm{p} \approx n_\mathrm{d}$. And for computational convenience, usually we want to use equally spaced basis functions, even if the data is in some sense nonuniformly spaced. But to control noise, we include a regularization term in the cost function rather than using the unregularized choice (2.2.13). Motivated by (2.2.10), we use a penalized least-squares cost function of the following form:

e,spline,xh,pls

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|^2 + \beta \frac{1}{2} \|\boldsymbol{C}\boldsymbol{x}\|^2 , \tag{2.2.14}$$

where $\boldsymbol{A}$ is defined in 2.2.12 and $\boldsymbol{C}$ is one of the 1D finite-differencing matrices defined in §1.8.1, typically (1.8.4). This form is closely related to (2.2.8), but (2.2.14) generalizes more easily to situations with more complicated noise models, physical models, and regularization methods.

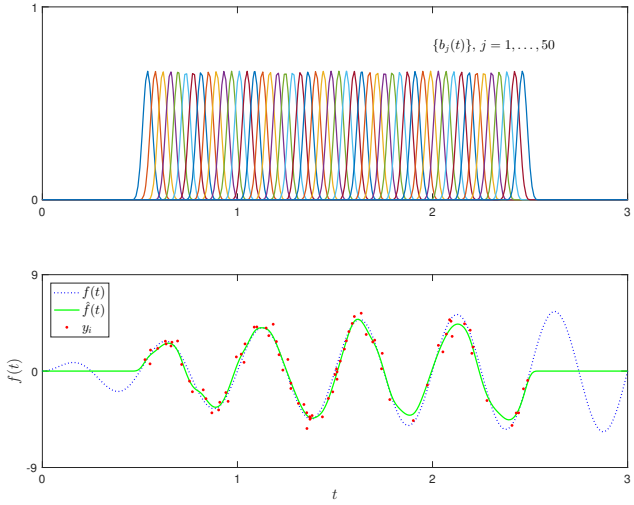



Figure 2.2.4: Penalized least-squares cubic B-spline smoothing of noisy data. The noisy samples $\{y_i\}$ are the red points, the true function $f(t)$ is the dashed blue curve, and the interpolator $\hat{f}(t)$ is the solid green curve.

fig_spline_pls

Fig. 2.2.4 illustrates a penalized LS B-spline fit for a large value of $n_\mathrm{p}$ and a reasonable value of $\beta$, chosen by trial and error. For large $n_\mathrm{p}$, the estimate $\hat{f}$ is indistinguishable from the spline smoothing estimate.

### 2.2.6 Splines with uniform sampling (s,reg,spline,unif)

s,reg,spline,unif

The properties of nonparametric regression are easily understood in signal processing terms by considering the case where the sample points are uniformly spaced.

Suppose we measure the value of a function (or signal) $f(t)$ at $N$ points over the unit interval:

e,spline,unif,yn

$$y_n = f(n/N) + \varepsilon_n, \qquad n = 0, \ldots, N-1, \tag{2.2.15}$$

where $\varepsilon_n$ has zero mean and variance $\mathsf{Var}\{\varepsilon_n\} = N\sigma^2$.

We would like to estimate the function $f(\cdot)$ from the measurements $\{y_n : n = 0, \ldots, N-1\}$. A parametric approach to this problem would assume that $f$ is linear or has some other simple parametric form, and would estimate the parameters that describe $f$ using criteria like maximum-likelihood or least-squares.

A nonparametric approach is to find a compromise between fit to the data and smoothness of the estimated function, as quantified by the following cost function and estimator:

$$\hat{f} = \arg\min_{f} \frac{1}{N} \sum_{n=0}^{N-1} \frac{1}{2} |y_n - f(n/N)|^2 + \beta \int_0^1 \frac{1}{2} \left| \frac{d^m}{dt^m} f(t) \right|^2 \mathrm{d}t \,.$$

The adjustable parameters in such an approach are $\beta$ and $m$.

When the samples are uniformly spaced, we can find the solution for $\hat{f}$ analytically using a Fourier series expansion of $f$ over the interval $[0, 1]$:

e,spline,unif,series

$$f(t) = \sum_{k=-\infty}^{\infty} c_k \, \mathrm{e}^{\imath 2\pi k t} \,. \tag{2.2.16}$$

(This choice imposes periodic boundary conditions.) The derivatives of $f(t)$ are thus

$$\frac{d^m}{dt^m} f(t) = \sum_{k=-\infty}^{\infty} c_k \, (\imath 2\pi k)^m \, \mathrm{e}^{\imath 2\pi k t} \,,$$

so Parseval's theorem expresses the roughness penalty in the frequency domain:

$$\int_0^1 \left| \frac{d^m}{dt^m} f(t) \right|^2 \mathrm{d}t = \sum_{k=-\infty}^{\infty} |c_k|^2 \, (2\pi k)^{2m}.$$

Thus, in terms of the $c_k$ values the cost function becomes:

$$\Psi(\boldsymbol{c}) = \frac{1}{N} \sum_{n=0}^{N-1} \frac{1}{2} \left| y_n - \sum_{k=-\infty}^{\infty} c_k \, \mathrm{e}^{\imath \frac{2\pi}{N} kn} \right|^2 + \beta \sum_{k=-\infty}^{\infty} \frac{1}{2} |c_k|^2 \, (2\pi k)^{2m}.$$

Because $\mathrm{e}^{\imath \frac{2\pi}{N} kn}$ is $N$-periodic in $k$, there is redundancy in the Fourier series expansion (2.2.16) for this problem. Because the penalty function increases as $|k|^2$, to minimize $\Psi$ we must use the set of $c_k$ values with the smallest possible $k$ values, *i.e.*, the set $-N/2 \le k < N/2$ (for $N$ even). In terms of these $c_k$ values the cost function becomes:

$$\Psi(\boldsymbol{c}) = \frac{1}{N} \sum_{n=0}^{N-1} \frac{1}{2} \left| y_n - \sum_{k=-N/2}^{N/2-1} c_k \, \mathrm{e}^{\imath \frac{2\pi}{N} kn} \right|^2 + \beta \sum_{k=-N/2}^{N/2-1} \frac{1}{2} |c_k|^2 \, (2\pi k)^{2m}.$$

To minimize, we equate the partial derivatives of $\Psi$ to zero (*cf.* §28.2):

$$0 = \frac{\partial}{\partial c_k} \Psi(\boldsymbol{c}) = \frac{1}{N} \sum_{n=0}^{N-1} \left( - \mathrm{e}^{-\imath \frac{2\pi}{N} kn} \right) \left( y_n - \sum_{l=-N/2}^{N/2-1} c_l \, \mathrm{e}^{\imath \frac{2\pi}{N} ln} \right) + \beta \, c_k \, (2\pi k)^{2m} \,,$$

so

$$\begin{aligned} Y_k \triangleq \frac{1}{N} \sum_{n=0}^{N-1} y_n \, \mathrm{e}^{-\imath \frac{2\pi}{N} kn} &= \sum_{l=-N/2}^{N/2-1} c_l \, \frac{1}{N} \sum_{n=0}^{N-1} \mathrm{e}^{\imath \frac{2\pi}{N} (l-k)n} + \beta \, c_k \, (2\pi k)^{2m} \\ &= \sum_{l=-N/2}^{N/2-1} c_l \, \delta[(k-l) \bmod N] + \beta \, c_k \, (2\pi k)^{2m} = c_k + \beta \, c_k \, (2\pi k)^{2m} \,, \end{aligned}$$

where $Y_k$ denotes the $N$-point DFT of $y_n$. Thus the optimal Fourier coefficients are

$$\hat{c}_k = \frac{1}{1 + \beta (2\pi k)^{2m}} Y_k.$$

Thus, we can find the $c_k$ values by windowing the DFT of the signal samples with the Butterworth-like filter having frequency response $\frac{1}{1+\beta(2\pi k)^{2m}}$.

So for the simple model of equally spaced samples (2.2.15), spline smoothing is equivalent to Butterworth filtering. However, the principles that underly spline smoothing generalize to nonuniform sample spacing and to problems with more complicated forward models than (2.2.15).

s,reg,spline,summ

### 2.2.7 Summary (s,reg,spline,summ)

Although the 1D spline smoothing problem is much simpler than typical image reconstruction problems, it illustrates many of the challenges faced in inverse problems, including ill-posedness, object parameterization, and regularization. This section focused on cases where the unknown function $f(t)$ is thought to be smooth. In cases where $f(t)$ is piecewise smooth, we might prefer to replace the quadratic roughness measure in (2.2.2) with a $\mathcal{L}_1$ norm: $\int \left| f^{(m)}(t) \right| \mathrm{d}t$. This is equivalent to assuming that the $m$th derivative of $f$ is **sparse**. For $m = 1$ this roughness measure is called **total variation** (**TV**). (See §2.4.)

## 2.3 Regularization implementations (s,reg,irt)

s,reg,irt

As discussed in §1.10.2, this *book* focuses on regularizers having the general form (1.10.10), repeated here:

$$\mathsf{R}(\boldsymbol{x}) = \sum_{k=1}^{K} \psi_k([\boldsymbol{C}\boldsymbol{x}]_k), \tag{2.3.1}$$

e,reg,irt,Rx

where $[\boldsymbol{C}\boldsymbol{x}]_k = \sum_{j=1}^{n_\mathrm{p}} c_{kj} x_j$. The matrix $\boldsymbol{C}$ is $K \times n_\mathrm{p}$ where $\boldsymbol{x} \in \mathbb{C}^{n_\mathrm{p}}$ or $\boldsymbol{x} \in \mathbb{R}^{n_\mathrm{p}}$. This form is sufficiently general to represent most, but not all, penalty functions (and log priors) that have been described in the literature. §2.7 describes choices for the **potential functions** $\psi_k$ in more detail. Typical choices for $\boldsymbol{C}$ include **finite differences**, as described in §1.8.1 and §1.10, or **wavelet transforms**, as mentioned in §1.12.3.

Most iterative optimization algorithms need to evaluate either $\mathsf{R}(\boldsymbol{x})$ or its gradient, or both, where the gradient was given in (1.10.13) and is also repeated here for convenience:

$$\nabla \mathsf{R}(\boldsymbol{x}) = \sum_{k=1}^{K} \nabla \psi_k(\boldsymbol{c}_k' \boldsymbol{x}) = \sum_{k=1}^{K} \boldsymbol{c}_k \dot{\psi}_k([\boldsymbol{C}\boldsymbol{x}]_k) \tag{2.3.2}$$

e,reg,irt,R,cgrad,dpot

$$= \sum_{k=1}^{K} \boldsymbol{c}_k \, \omega_k([\boldsymbol{C}\boldsymbol{x}]_k)[\boldsymbol{C}\boldsymbol{x}]_k = \boldsymbol{C}' \, \boldsymbol{D}(\boldsymbol{x}) \, \boldsymbol{C}\boldsymbol{x}, \tag{2.3.3}$$

e,reg,irt,R,cgrad,wpot

where $\boldsymbol{c}_k' = \boldsymbol{e}_k' \boldsymbol{C}$ denotes the $k$th row of $\boldsymbol{C}$ and we define the following $K \times K$ diagonal weighting matrix:

$$\boldsymbol{D}(\boldsymbol{x}) \triangleq \mathsf{diag}\{\omega_k([\boldsymbol{C}\boldsymbol{x}]_k)\}\,. \tag{2.3.4}$$

e,reg,irt,D

The **potential weighting function** $\omega_k(z) \triangleq \frac{\dot{\psi}_k(z)}{z}$ was introduced in (1.10.12) and we assume it is nonnegative and finite whenever we use $\omega_k$.

There are many ways to implement in software the operations required for regularization. This section describes some of the options that are available in the *Michigan Image Reconstruction Toolbox*. For simplicity we focus on 2D regularization but the principles generalize readily. We focus on 1st-order finite differences but the principles generalize to 2nd-order differences and other linear combinations. We focus on methods for computing the gradient (2.3.3) because that is usually more essential for implementation than the cost function (2.3.1) itself. In particular, we focus primarily on methods for computing *all* elements of $\nabla \mathsf{R}(\boldsymbol{x})$ simultaneously, as required for most gradient-based algorithms.

### 2.3.1 Basic matrix implementation

s,reg,irt,basic

A direct implementation of the gradient (2.3.2) uses the following steps.

- Use some type of matrix-vector multiplication to compute $\boldsymbol{d} = \boldsymbol{C}\boldsymbol{x}$.
- Compute the vector $\boldsymbol{g}$ defined by $g_k = \dot{\psi}_k(d_k),\ k = 1, \dots, K$.
- Use some type of matrix-vector multiplication to compute $\nabla \mathsf{R}(\boldsymbol{x}) = \boldsymbol{C}' \boldsymbol{g} = \sum_{k=1}^{K} \boldsymbol{c}_k g_k$.

The matrix $\boldsymbol{C}$ is usually extremely sparse in image reconstruction problems. Specifically, if we use 1st-order differences as described in (1.10.1), then each row of $\boldsymbol{C}$ has at most two nonzero elements (out of $n_\mathrm{p}$). Therefore, one natural way to store $\boldsymbol{C}$ is as a **sparse matrix**, meaning a data structure that stores only the nonzero values and the locations of those values in a list. However, there are even more efficient methods for computing $\boldsymbol{C}\boldsymbol{x}$ that exploit the structure of $\boldsymbol{C}$. See §2.14 for more details.

### 2.3.2 General 1st-order roughness penalty in 2D

To illustrate the practical challenges in implementing the procedure described in §2.3.1, consider the case of 2D regularization based on 1st-order differences. For a $M \times N$ image $f[m,n]$, a general roughness penalty has the form

$$\mathsf{R}(f) = \sum_{l=1}^{L} \beta_l \, \mathsf{R}_l(f) \tag{2.3.5}$$

e,reg,irt,Rf,L

$$\mathsf{R}_l(f) = \sum_{m=\max(m_l,0)}^{M-1+\min(m_l,0)} \sum_{n=\max(n_l,0)}^{N-1+\min(n_l,0)} \psi_{m,n,l}(f[m,n] - f[m-m_l, n-n_l]) \tag{2.3.6}$$

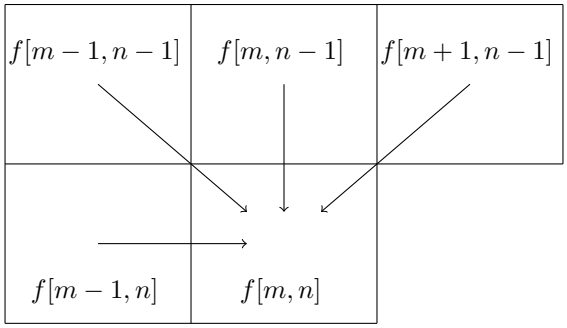
e,reg,irt,Rf_l

for some integers $m_l \in \{-(M-1), \dots, M-1\}$ and $n_l \in \{-(N-1), \dots, N-1\}$ chosen by the algorithm designer.

Each $(m_l, n_l)$ pair denotes the coordinate offset to a neighbor. For example, the simple case in (1.10.1) corresponds to $L = 2$, $(m_1, n_1) = (1, 0)$, $(m_2, n_2) = (0, 1)$, and $\beta_1 = \beta_2$. When we include diagonal neighbors, then $L = 4$ and the $(m_l, n_l)$ pairs are

$$\{(1,0),\ (0,1),\ (-1,1),\ (1,1)\}\,, \tag{2.3.7}$$

e,reg,irt,2d,hvd




fig,reg,penal2,4

Figure 2.3.1: Four neighbors used for 2D regularization with 1st-order differences.

as illustrated in Fig. 2.3.1.

We allow a different regularization parameter $\beta_l$ for each neighbor offset, because often we give less weight to neighbors that are more distant, *e.g.*, by choosing

$$\beta_l \propto \frac{1}{\sqrt{m_l^2 + n_l^2}}.$$

In particular, often we give a weight of $1/\sqrt{2}$ (or $1/2$, see §5.2) for the diagonal neighbors relative to the horizontal and vertical neighbors. A reasonably general form is

$$\beta_l \propto \left( \frac{1}{\sqrt{m_l^2 + n_l^2}} \right)^p,$$

for some power $p$; typically $p = 0$, $p = 1$ or $p = 2$.

MIRT *The regularization functions* `Reg1` *and* `Rweights` *have an option* `distance_power` *for selecting the power* $p$. *The default is* $p = 1$ *for historical reasons but* $p = 2$ *may be preferable in terms of resolution properties per* §5.2.

For generality, we allow $\psi_{m,n,l}$ in (2.3.7) to depend on spatial location, because space-varying regularization can be useful in some applications, and possibly even to apply different amounts of regularization in the various directions, *e.g.*, [27, 28]. Often the generality needed is to have

e,reg,irt,pot,rmn_l

$$\psi_{m,n,l}(z) = r_l[m,n]\, \psi(z), \tag{2.3.8}$$

where the possibly space-varying regularization coefficients $\{r_l[m,n] : l = 1, \dots, L\}$ must be designed somehow, *e.g.*, as described in Chapter 5 and [27–29].

MIRT *The regularization functions* `Reg1` *and* `Rweights` *have an option* `'user_wt'` *for providing a* $M \times N \times L$ *array specifying the* $\{r_l[m,n]\}$ *values.*

We can express (2.3.6) in matrix-vector notation as

$$\mathsf{R}_l(\boldsymbol{x}) = \sum_k \psi_k([\boldsymbol{C}_l \boldsymbol{x}]_k)$$

provided we define appropriately the matrices $\boldsymbol{C}_l$ for $l = 1, \dots, L$. Each row of $\boldsymbol{C}_l$ corresponds to one term in the summation (2.3.6), because each difference of nearby pixel values, $f[m,n] - f[m - m_l, n - n_l]$, is a simple linear combination of the $f[m,n]$ values. The natural choice for $\boldsymbol{C}_l$ would have size $(M - |m_l|)(N - |n_l|) \times MN$, because this is the number of terms in the sum (2.3.6). However, it can be more convenient for implementation to choose $\boldsymbol{C}_l$ to have size $MN \times MN$, allowing $\boldsymbol{C}_l$ to have a few rows that are entirely zero. Such zero rows do not change the value of the penalty function. Or, instead of being entirely zero, those rows may have entries that correspond to other end conditions.

Recalling (1.4.16), we can identify the term $f[m,n] - f[m - m_l, n - n_l]$ with the $k$th row of $\boldsymbol{C}_l$, where $k = 1 + m + nM$. With this natural ordering, the elements of $\boldsymbol{C}_l$ are as follows:

e,reg,irt,Cl

$$[\boldsymbol{C}_l]_{kj} = \begin{cases} 1, & k = j = 1 + m + nM, \ m \in \mathcal{S}(m_l, M), \ n \in \mathcal{S}(n_l, N) \\ -1, & k = 1 + m + nM \\ & j = 1 + m - m_l + (n - n_l)M \\ 0, & \text{otherwise,} \end{cases}, \ m \in \mathcal{S}(m_l, M), \ n \in \mathcal{S}(n_l, N) \tag{2.3.9}$$

where we define the support set

$$\mathcal{S}(n, N) \triangleq \{\max(n, 0), \ \dots, \ N - 1 + \min(n, 0)\}.$$

Each row of $\boldsymbol{C}_l$ has (at most) a single "$-1$" entry and one "1" entry, and all other elements are zero. Thus $\boldsymbol{C}_l$ is a very **sparse matrix**. One can verify that if $\boldsymbol{x}$ denotes the lexicographic representation of $f[m,n]$ per (1.4.14), then

$$[\boldsymbol{C}_l \boldsymbol{x}]_k \Big|_{k=1+m+nM} = \begin{cases} f[m,n] - f[m-m_l, n-n_l], & m \in \mathcal{S}(m_l, M),\ n \in \mathcal{S}(n_l, N) \\ 0, & \text{otherwise.} \end{cases}$$

Having defined these $\boldsymbol{C}_l$ matrices, we can write the 2D penalty function (2.3.5) in the general form (2.3.1) by defining the following $LMN \times MN$ matrix that generalizes (1.10.8):

$$\boldsymbol{C} = \begin{bmatrix} \boldsymbol{C}_1 \\ \vdots \\ \boldsymbol{C}_L \end{bmatrix}. \tag{2.3.10}$$ e,reg,irt,C,stack

Thus, in the usual 2D case (2.3.5), we have $K = LMN$ in (2.3.1).

MIRT *The function* `Cdiff1` *generates such* $\boldsymbol{C}_l$ *objects using several methods, including* MATLAB *indexing, or a* **MEX** *file, or a sparse matrix, or a convolution operation. The function* `Cdiffs` *represents* $\boldsymbol{C}$ *by stacking up objects generated by* `Cdiff1` *via (2.3.10). See* §2.14.

### 2.3.3 Stacked matrix implementation

The three-step approach described above in §2.3.1 has the appeal of being modular and appearing simple, but it has the serious drawback of requiring considerable extra memory for storing the intermediate results. Consider a 3D problem where $\boldsymbol{x}$ represents a $N^3$ image. For 1st-order finite differences with the 26 nearest neighbors to each voxel, the vector of differences $\boldsymbol{d}$ has length about $LN^3$, where $L = 13$. For an X-ray CT problem with $N = 512$, this is 6.5 GB for 4-byte floating point numbers, which would be inconveniently large.

To reduce memory overhead, we can use the stacked form (2.3.10) that is available for penalties of the form (2.3.5). In particular, for such penalty functions we can rewrite the regularizer (2.3.1) and its gradient (2.3.3) as follows:

$$\mathsf{R}(\boldsymbol{x}) = \sum_{l=1}^{L} \sum_{k} \psi_{k,l}([\boldsymbol{C}_l \boldsymbol{x}]_k), \qquad \nabla \mathsf{R}(\boldsymbol{x}) = \sum_{l=1}^{L} \boldsymbol{C}_l' \, \boldsymbol{D}_l(\boldsymbol{x}) \, \boldsymbol{C}_l \boldsymbol{x} \tag{2.3.11}$$ e,reg,irt,Rx,grad,Cl

where $\boldsymbol{D}_l(\boldsymbol{x}) = \mathsf{diag}\{\omega_{k,l}([\boldsymbol{C}_l \boldsymbol{x}]_k)\}$ and $k$ ranges from 1 to $n_\mathrm{p}$, where $n_\mathrm{p} = MN$ in 2D.

This expanded form suggests the following procedure for computing $\nabla \mathsf{R}(\boldsymbol{x})$.

```
g := 0
for   l = 1, ..., L
      d := C_l x
      d_k := ψ̇_{k,l}(d_k),  k = 1, ...
      b := C_l' d
      g := g + b
end
```
(2.3.12) e,reg,irt,Cl,alg

In this version the intermediate vectors $\boldsymbol{d}$ and $\boldsymbol{b}$ are the same size as $\boldsymbol{x}$, so the extra intermediate storage is only $2N^3$ instead of $LN^3$.

### 2.3.4 Reduced memory for regularization coefficients and 3D regularization

For the general type of weighting for the potential functions given in (2.3.8), the procedure in (2.3.12) would require storing all of the regularization coefficients $\{r_l[m,n] : l = 1, \dots, L\}$ (or computing these on-the-fly), which may be prohibitively large or expensive in some 3D problems. A useful family of space-variant regularizers[1] that uses less memory is

$$r_l[m,n] = \kappa[m,n] \, \kappa[m-m_l, n-n_l], \tag{2.3.13}$$ e,reg,irt,rmn_1,kap

where the 2D array $\kappa[m,n]$ describes pixel-dependent regularization factors. This form requires storing only $\{\kappa[m,n]\}$ and $\{\beta_l\}$. This family was used in [31] for example. (See Chapter 22.) However, this family is not general enough to accommodate some more complicated regularizers with direction-dependent factors [27–29].

For 3D problems, the generalization of the form (2.3.13) is $r_l[m,n,k] = \kappa[m,n,k] \, \kappa[m-m_l, n-n_l, k-k_l]$. Combining with (2.3.8) and (2.3.5) yields the following 3D regularizer

$$\mathsf{R}(f) = \sum_{l=1}^{L} \beta_l \sum_{m,n,k} \kappa[m,n,k] \, \kappa[m-m_l, n-n_l, k-k_l] \, \psi(f[m,n,k] - f[m-m_l, n-n_l, k-k_l]), \tag{2.3.14}$$ e,reg,irt,3D

where the limits on the sums over $m, n, k$ follow (2.3.6). In practice, to approach isotropic spatial resolution, it is often necessary to choose the values of $\beta_l$ that correspond to a scanner's axial direction (typically $z$, *i.e.*, $k$) *differently* from the values of $\beta_l$ that correspond to the transaxial plane (typically $x, y$, *i.e.*, $m, n$).

[1] See [30] for smoothing splines with varying regularization parameters.

### 2.3.5 2nd-order differences

In some applications one can improve image quality by using 2nd-order differences, generalizing (2.3.6) with (2.3.8) as follows:

$$\mathsf{R}_l(f) = \sum_{n=|n_l|}^{N-1-|n_l|} \sum_{m=|m_l|}^{M-1+|m_l|} r_l[m,n]\, \psi(2\, f[m,n] - f[m-m_l, n-n_l] - f[m+m_l, n+n_l])\,. \tag{2.3.15}$$

e,reg,irt,Rlf,2nd,nm

More concisely, letting $\vec{n} = (m,n)$ and $\vec{n}_l = (m_l, n_l)$ we write

$$\mathsf{R}_l(f) = \sum_{m=|m_l|}^{M-1-|m_l|} \sum_{n=|n_l|}^{N-1-|n_l|} r_l[\vec{n}]\, \psi(2\, f[\vec{n}] - f[\vec{n}-\vec{n}_l] - f[\vec{n}+\vec{n}_l])\,. \tag{2.3.16}$$

e,reg,irt,Rlf,2nd

MIRT *The form (2.3.13) is for 1st-order differences. For 2nd-order differences we use*

$$r_l[m,n] = \kappa[m,n] \sqrt{\kappa[m-m_l, n-n_l]\, \kappa[m+m_l, n+n_l]}. \tag{2.3.17}$$

e,reg,irt,rmn_l,kap,2

### 2.3.6 Non-matrix implementations

A drawback of the procedure (2.3.12) is that it accesses sequentially the memory used for storing $\boldsymbol{x}$ a total of $L$ times. Thus, for large 3D problems the execution time for this procedure can be constrained by the memory bandwidth.

To overcome this limitation, one can abandon the general matrix form (2.3.11) and focus instead on the specific form given in (2.3.5). For simplicity, consider the case where the potential functions have the common form (2.3.8). In this case, for voxels away from the image borders, the partial derivatives of $\mathsf{R}(f)$ have the form

$$\frac{\partial}{\partial f[m,n]} \mathsf{R}(f) = \sum_{l=1}^{L} \beta_l \Big( r_l[m,n]\, \dot{\psi}(f[m,n] - f[m-m_l, n-n_l]) + r_l[m+m_l, n+n_l]\, \dot{\psi}(f[m,n] - f[m+m_l, n+n_l]) \Big)\,. \tag{2.3.18}$$

e,reg,irt,d,fnm,local

(Slightly different formulas are needed for pixels that lie on the image borders.) One can loop over all the pixels and evaluate this sum using relatively local memory accesses and with only one pass over the image memory.

For 2nd-order finite differences (2.3.16) the partial derivatives have the form

$$\frac{\partial}{\partial f[\vec{n}]} \mathsf{R}(f) = \sum_{l=1}^{L} \beta_l \Big( 2\, r_l[\vec{n}]\, \dot{\psi}(2\, f[\vec{n}] - f[\vec{n}-\vec{n}_l] - f[\vec{n}+\vec{n}_l]) - r_l[\vec{n}+\vec{n}_l]\, \dot{\psi}(2\, f[\vec{n}+\vec{n}_l] - f[\vec{n}+2\vec{n}_l] - f[\vec{n}]) - r_l[\vec{n}-\vec{n}_l]\, \dot{\psi}(2\, f[\vec{n}-\vec{n}_l] - f[\vec{n}-2\vec{n}_l] - f[\vec{n}]) \Big)\,. \tag{2.3.19}$$

e,reg,irt,d,fnm,local,2nd

In particular, if $r_l[\vec{n}] = 1$ and $\dot{\psi}(z) = z$ then (2.3.19) requires at least $11L$ operations per pixel.

MIRT *The function* `Reg1.m` *creates an object that can evaluate the roughness penalty (2.3.5) and its gradient (2.3.18). It is designed to accommodate regularizers of the general form (2.3.8) and (2.3.13). The method* `R.cgrad` *computes the gradient of* $\mathsf{R}(\boldsymbol{x})$ *using (2.3.18). The* `'offsets'` *option defines what set of* $(m_l, n_l)$ *pairs are to be used. For the case (2.3.7),* `'offsets'` *would be* `[1 M M+1 M-1]`*. In general in 2D, neighbor* $(m_l, n_l)$ *corresponds to* `offset` $m_l + n_l M$ *because of* ***lexicographic ordering*** *of a* $M \times N$ *image as a vector.*

### 2.3.7 Support mask considerations

The penalty function formula (2.3.6) assumes that all pixels in the image are to be reconstructed. In practical medical imaging, often we need only to reconstruct a subset of the image, because the scanner field of view (*e.g.*, as defined by the patient portal) is often circular rather than square. Let $\chi[m,n]$ denote the binary function that is nonzero for pixels that are to be reconstructed and is zero otherwise. We refer to $\chi[m,n]$ as the reconstruction **mask**. The length $n_\mathrm{p}$ of the parameter vector $\boldsymbol{x}$ that denotes all the unknown pixel values is

$$n_\mathrm{p} = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \chi[m,n] \le MN.$$

Fig. 2.3.2 illustrates a rectangular image in which only a subset of the pixels are to be estimated. Each $\vec{n}_j$ denotes the coordinates $[m,n]$ of the pixels to be estimated. If we define $\mathcal{S} = \{\vec{n}_1, \ldots, \vec{n}_{n_\mathrm{p}}\}$ then $\chi[m,n] = \mathbb{I}_{\{[m,n] \in \mathcal{S}\}}$.

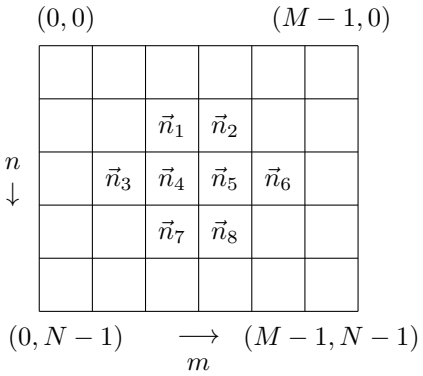



Figure 2.3.2: A $M \times N = 6 \times 5$ lattice with approximately circular FOV. Only the pixels with indices are estimated. In this example, $n_{\mathrm{p}} = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \chi[m,n] = 8$. fig,reg,mask

When designing a roughness penalty function, one choice that arises is whether to penalize the differences between pixel values that lie on the edge of the reconstruction mask but still within the mask and their neighbors that lie outside of that mask. Usually we do not want to penalize such differences, *i.e.*, we want $r_l[m,n] = 0$ if $\chi[m,n] = 1$ and $\chi[m - m_l, n - n_l] = 0$. We refer to this as a **tight** boundary condition. (Otherwise we call it a **leaky** boundary condition.) A simple way to ensure this property is to use regularization coefficients of the form (2.3.13) and to choose $\kappa[m,n]$ such that $\chi[m,n] = 0 \implies \kappa[m,n] = 0$.

MIRT *The* `'tight'` *and* `'leak'` *choices of the* `'edge_type'` *option of* `Rweights.m` *control this behavior.*

### 2.3.8 Coordinate-wise implementation

For algorithms that update one pixel at a time, such as **iterative coordinate descent** (**ICD**), instead of computing all of $\nabla \mathsf{R}(\boldsymbol{x})$ simultaneously, we only need to compute a single element of that gradient vector, *i.e.*,

$$\frac{\partial}{\partial x_j} \mathsf{R}(\boldsymbol{x}) = \sum_{k=1}^{K} c_{kj} \, \dot{\psi}_k(\boldsymbol{c}_k' \boldsymbol{x}) = \sum_{k \in \mathcal{K}_j} c_{kj} \, \dot{\psi}_k(\boldsymbol{c}_k' \boldsymbol{x}),$$

where $\mathcal{K}_j \triangleq \{k = 1, \ldots, K : c_{kj} \neq 0\}$. Note that $\boldsymbol{c}_k' \boldsymbol{x} = \sum_{j \in \mathcal{J}_k} c_{kj} x_j$ where $\mathcal{J}_k \triangleq \{j = 1, \ldots, n_{\mathrm{p}} : c_{kj} \neq 0\}$. In practice this usually is implemented like (2.3.18).

MIRT *This capability exists in the compiled ASPIRE software [32], but not in the Michigan Image Reconstruction Toolbox because coordinate-wise methods are poorly suited to interpreted languages like* MATLAB.

## 2.4 Regularization in variational formulations (s,reg,var)

s,reg,var

The regularizing roughness penalty functions introduced in §1.10 and §2.3 were formulated in terms of discrete-space images $f[m,n]$. And in practice numerical implementations of regularization always involve discretization. Nevertheless, for insight it can be useful to consider regularization functionals defined in terms of continuous-space images $f(x,y)$. These are called **variational** formulations, and they are multidimensional generalizations of the nonparametric spline methods of §2.2.

### 2.4.1 Thin membrane regularization

The 1st-order roughness penalty function (1.10.1) or (2.3.5), with quadratic potential functions, is a discrete approximation to the following roughness penalty function for a continuous-space image $f(x,y)$:

$$\mathsf{R}_{\mathrm{TM}}(f) = \iint \frac{1}{2} \left| \frac{\partial}{\partial x} f(x,y) \right|^2 + \frac{1}{2} \left| \frac{\partial}{\partial y} f(x,y) \right|^2 \mathrm{d}x \, \mathrm{d}y = \int \frac{1}{2} \left\| \nabla f \right\|^2, \tag{2.4.1}$$

e,reg,var,Rf,quad

where $\nabla f(x,y) \triangleq \left( \frac{\partial}{\partial x} f(x,y), \frac{\partial}{\partial y} f(x,y) \right)$. The functional $\mathsf{R}_{\mathrm{TM}}(f)$ is related to the bending energy of a **thin membrane** [33–35].

### 2.4.2 Rotation invariance

s,reg,var,rot,inv

One can show that the penalty function (2.4.1) is invariant to spatial rotations of $f$, *i.e.*, if we define

$$f_\theta(x,y) = f(x \cos\theta + y \sin\theta, -x \sin\theta + y \cos\theta),$$

then $\mathsf{R}_{\text{TM}}(f_\theta) = \mathsf{R}_{\text{TM}}(f)$. This invariance seems to be a desirable property for most imaging problems. However, implementation requires discretization, *e.g.*, on a 2D Cartesian grid, which usually loses rotation invariance.

### 2.4.3 Thin plate regularization

Grimson [36] considered rotationally invariant penalty functions involving second derivatives (see Problem 2.5) and presented arguments favoring the following choice in the context of surface reconstruction:

$$\mathsf{R}_{\text{TP}}(f) = \iint \left| \frac{\partial^2}{\partial x^2} f \right|^2 + 2 \left| \frac{\partial^2}{\partial x \, \partial y} f \right|^2 + \left| \frac{\partial^2}{\partial y^2} f \right|^2 \mathrm{d}x \, \mathrm{d}y \,. \tag{2.4.2}$$

e,reg,var,tps

This is the energy associated with **thin-plate splines** [37, 38], a popular deformation model for nonrigid image registration. This penalty function is zero if $f(x,y)$ is an affine function. In surface reconstruction and image registration problems often it is natural for affine functions to be unpenalized. In contrast, in many image reconstruction problems, uniform images may be more likely than affine images, so the rotationally invariant 1st-order penalty (2.4.1) is used more frequently than (2.4.2).

### 2.4.4 Edge preserving variational regularization

Suppose we replace the squaring operations in (2.4.1) with *nonquadratic* potential functions:

$$\mathsf{R}_1(f) = \iint \psi\left( \frac{\partial}{\partial x} f(x,y) \right) + \psi\left( \frac{\partial}{\partial y} f(x,y) \right) \mathrm{d}x \, \mathrm{d}y \,.$$

Although this roughness penalty function will help preserve some edges, in general it is *not* rotation invariant. One example of such an approach is a form of **total variation** (**TV**) regularization called **anisotropic TV** or **bilateral TV** [39], where $\psi(z) = |z|$. To ensure rotation invariance, one can instead use the following form, *e.g.*, [40]:

$$\mathsf{R}_2(f) = \iint \psi\left( \sqrt{\left| \frac{\partial}{\partial x} f(x,y) \right|^2 + \left| \frac{\partial}{\partial y} f(x,y) \right|^2} \right) \mathrm{d}x \, \mathrm{d}y = \iint \psi(\|\nabla f\|) \, \mathrm{d}x \, \mathrm{d}y \,. \tag{2.4.3}$$

e,reg,var,Rf2

However, the corresponding discrete representation is not of the general form given in (2.3.1). To attempt rotation invariance in the discrete case, we can generalize (2.3.1) to the form

$$\mathsf{R}(\boldsymbol{x}) = \sum_{k=1}^{K} \psi_k\left( \sqrt{|[\boldsymbol{C}_{\text{X}} \boldsymbol{x}]_k|^2 + |[\boldsymbol{C}_{\text{Y}} \boldsymbol{x}]_k|^2} \right), \tag{2.4.4}$$

e,reg,R,rot,inv

where $\boldsymbol{C}_{\text{X}}$ and $\boldsymbol{C}_{\text{Y}}$ denote, for example, the top and bottom halves of $\boldsymbol{C}$ in (1.14.2) or (2.3.10).

If we choose a **hyperbola** potential function:

$$\psi(z) = \delta^2 \left( \sqrt{1 + |z/\delta|^2} - 1 \right), \tag{2.4.5}$$

e,reg,hyperbola

then (2.4.3) and (2.4.4) become the **Beltrami regularizer** used in [41].

For notational simplicity, we focus primarily on the form (2.3.1) throughout this *book*. All of the algorithms that are suitable for regularized estimation using (2.3.1) can be generalized fairly easily to accommodate (2.4.4). Such generalizations are left as exercises for the reader, *e.g.*, Problem 12.6.

s,reg,var,tv

### 2.4.5 Total variation (TV) methods

Because the quadratic roughness penalty (2.4.1) blurs edges, a popular alternative is to replace it with the **total variation** (**TV**) regularizer [42–51].

For an arbitrary real-valued function $f$ defined on an interval $[a,b]$, its **total variation** is defined by the general formula:

$$\|f\|_{\text{TV}} \triangleq \sup_P \sum_{i=0}^{|P|-1} |f(t_{i+1}) - f(t_i)| \,, \tag{2.4.6}$$

e,reg,var,tv,1

where the supremum is taken over all partitions $P$ of the interval $[a,b]$. Strictly speaking this is a **semi-norm** because $\|f\|_{\text{TV}} = 0$ for any constant function $f$. For a 1D continuously differentiable function, the **total variation** simplifies to $\int \left| \dot{f}(t) \right| \mathrm{d}t$. For a $n$-dimensional differentiable function the total variation is given by the "**TV norm**:"

$$\mathsf{R}_{\text{TV}}(f) \triangleq \|f\|_{\text{TV}} \triangleq \int \|\nabla f(\vec{\mathrm{x}})\| \, \mathrm{d}\vec{\mathrm{x}} \,. \tag{2.4.7}$$

e,reg,var,tv

A more general definition of $\|f\|_{\mathrm{TV}}$ requires only that $f$ be absolutely integrable, *i.e.*, does not require continuity or differentiability [wiki]. However, such technicalities have limited practical interest in image reconstruction because images must be discretized for computing. The TV regularizer is not everywhere differentiable in $f$, so in practice often it is replaced by an approximation:

$$\mathsf{R}_{\mathrm{TV}}(f) \approx \int \psi(\|\nabla f(\vec{\mathrm{x}})\|) \, \mathrm{d}\vec{\mathrm{x}}, \tag{2.4.8}$$

e,reg,var,tv,corner

where $\psi(z) = \sqrt{|z|^2 + \varepsilon^2}$, for some small positive value of $\varepsilon$. This "**corner rounding**" approximation is simply the **hyperbola** potential function (2.4.5), one of many possibilities described in Table 2.1. Thus, total variation methods are simply a special case of edge-preserving regularization with a convex potential function.

Usually one uses the 2-norm in (2.4.7), which is called **isotropic TV** because it is invariant to rotations. Using the 1-norm in (2.4.7) leads to **anisotropic TV** or **bilateral TV**. To unify the anisotropic TV and isotropic TV formulations, one can use the equality [52]

$$\begin{aligned}
\sqrt{a^2 + b^2} &= \frac{1}{2} \int_0^{\pi/2} (|a\cos\varphi + b\sin\varphi| + |b\cos\varphi - a\sin\varphi|) \, \mathrm{d}\varphi \\
&= \frac{\int_0^{\pi/2} (|a\cos\varphi + b\sin\varphi| + |b\cos\varphi - a\sin\varphi|) \, \mathrm{d}\varphi}{\int_0^{\pi/2} (\cos\varphi + \sin\varphi) \, \mathrm{d}\varphi} \\
&\approx \frac{\sum_{\phi \in \left\{0, \frac{\pi}{2K}, \ldots, \frac{\pi(K-1)}{2K}\right\}} (|a\cos\varphi + b\sin\varphi| + |b\cos\varphi - a\sin\varphi|)}{\sum_{\phi \in \left\{0, \frac{\pi}{2K}, \ldots, \frac{\pi(K-1)}{2K}\right\}} (\cos\varphi + \sin\varphi)}.
\end{aligned}$$

This leads to the following approximation of the TV semi-norm:

$$\|f\|_{\mathrm{TV}} \approx \frac{\sum_{\phi \in \left\{0, \frac{\pi}{2K}, \ldots, \frac{\pi(K-1)}{2K}\right\}} \int \left( \left| \frac{\partial}{\partial x} f \cos\varphi + \frac{\partial}{\partial y} f \sin\varphi \right| + \left| \frac{\partial}{\partial y} f \cos\varphi - \frac{\partial}{\partial x} \sin\varphi \right| \right) \mathrm{d}x \, \mathrm{d}y}{\sum_{\phi \in \left\{0, \frac{\pi}{2K}, \ldots, \frac{\pi(K-1)}{2K}\right\}} (\cos\varphi + \sin\varphi)}.$$

For $K = 1$ this simplifies to the usual anisotropic TV, whereas for $K = 2$ this simplifies to

$$\|f\|_{\mathrm{TV}} \approx \frac{\int \left| \frac{\partial}{\partial x} f \right| + \left| \frac{\partial}{\partial y} f \right| \mathrm{d}x \, \mathrm{d}y + \frac{\sqrt{2}}{2} \int \left| \frac{\partial}{\partial x} f + \frac{\partial}{\partial y} f \right| + \left| \frac{\partial}{\partial x} f - \frac{\partial}{\partial y} f \right| \mathrm{d}x \, \mathrm{d}y}{1 + \sqrt{2}},$$

from which one can design a discrete-space TV approximation of the form (2.3.1).

In 2D, another way of writing the TV functional is in terms of its directional derivatives [53]:

$$\mathsf{R}_{\mathrm{TV}}(f) = \sqrt{2} \iint \left( \frac{1}{2\pi} \int_0^{2\pi} |D_\phi \, f(x,y)|^2 \, \mathrm{d}\phi \right)^{1/2} \mathrm{d}x \, \mathrm{d}y,$$

where $D_\phi \, f(x,y) \triangleq \cos\phi \frac{\partial}{\partial x} f(x,y) + \sin\phi \frac{\partial}{\partial y} f(x,y)$ . This expression invites generalizations such as using higher-order derivatives, called **higher-degree TV** (**HDTV**) [53].

Another generalization is **total generalized variation** (**TGV**) [54, 55], which encourages the image to be piecewise smooth rather than piecewise constant, thereby reducing the **stair-step artifacts** that often plaque images based on conventional TV. A method based on second derivatives [56] has similar motivations.

TV regularizers encourage piecewise constant functions. Fig. 2.4.1 illustrates this property, where one sees that

$$\int \left| \dot{f}_3 \right|^2 > \int \left| \dot{f}_2 \right|^2 > \int \left| \dot{f}_1 \right|^2$$

but

$$\int \left| \dot{f}_3 \right| = \int \left| \dot{f}_2 \right| = \int \left| \dot{f}_1 \right| = 1.$$

Given the observation model $g = \boldsymbol{\mathcal{A}} f + \varepsilon$, a typical TV approach would be the regularized approach:

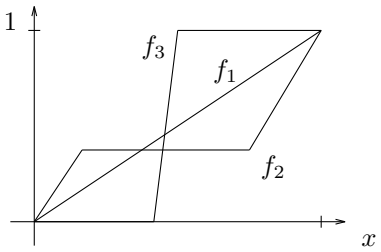
$$\arg\min_f \|\boldsymbol{\mathcal{A}} f - g\|_2^2 + \beta \, \|f\|_{\mathrm{TV}}$$

or the constrained approach:

$$\arg\min_f \|f\|_{\mathrm{TV}} \text{ sub. to } \|\boldsymbol{\mathcal{A}} f - g\|_2^2 \le n_{\mathrm{d}} \sigma^2.$$

These are somewhat challenging optimization problems. An alternative approach to TV minimization is to use augmented cost function methods (see §12.7) such as [57]:

$$\arg\min_{f,u} \|\boldsymbol{\mathcal{A}} f - g\|_2^2 + \mu \, \|f - u\|_2^2 + \beta \, \|u\|_{\mathrm{TV}},$$




fig,reg,tv

Figure 2.4.1: Three functions having different derivatives but having the same TV norm.

where the parameter $\mu$ must be chosen, or [58]:

$$\arg\min_{f,v} \|\mathcal{A}f - g\|_2^2 + \mu \|Df - v\|_2^2 + \beta \|v\|_1 .$$

In these alternatives, minimizing over $u$ or $v$, for a given $f$, does not involve $\mathcal{A}$, simplifying those updates. Such ideas date at least back to [59].

More recently, **graph cut** methods [60] and **augmented Lagrangian** methods [61, 62] have been examined for such minimization problems.

## 2.5 Regularization parameter selection (s,reg,hyper)

s,reg,hyper

One challenge in using regularized methods for image reconstruction is selecting the **regularization parameter** $\beta$, also known as the **hyperparameter** in the Bayesian terminology [63]. There are many criteria that have been proposed for selecting $\beta$, and several papers survey such methods [wiki] [64–71]. Chapter 22 describes methods for choosing regularization parameters based on **spatial resolution analysis**. Here we focus on more traditional methods that attempt to minimize the **estimation error**.

### 2.5.1 Oracle selection

Let $\hat{\boldsymbol{x}}_\beta$ denote the estimate $\hat{\boldsymbol{x}}$ as a function of the regularization parameter $\beta$. From an error point of view, we would like to choose $\beta$ so that $\hat{\boldsymbol{x}}_\beta$ is close to $\boldsymbol{x}_{\text{true}}$, *e.g.*, by minimizing the squared **estimation error**:

$$\boxed{\beta_{\text{O}} \triangleq \arg\min_\beta \|\hat{\boldsymbol{x}}_\beta - \boldsymbol{x}_{\text{true}}\|^2 .} \tag{2.5.1}$$

e,reg,hyper,se

Of course other norms could also be appropriate.

Because $\hat{\boldsymbol{x}}_\beta$ is a random vector (a function of the data $\boldsymbol{y}$), using the above criterion would lead to a somewhat different $\beta$ value for every noise realization. An alternative is to use the **mean squared error** (**MSE**):

$$\beta_{\text{MSE}} \triangleq \arg\min_\beta \text{MSE}_\beta , \qquad \text{MSE}_\beta \triangleq \mathsf{E}\left[\|\hat{\boldsymbol{x}}_\beta - \boldsymbol{x}_{\text{true}}\|^2\right] . \tag{2.5.2}$$

e,reg,hyper,mse

This is also known as the **risk criterion** for selecting $\beta$ [64]. Defining the estimator ensemble mean as

$$\bar{\boldsymbol{x}}_\beta \triangleq \mathsf{E}[\hat{\boldsymbol{x}}_\beta], \tag{2.5.3}$$

e,reg,hyper,xbr

we can write the **MSE** of any such estimator as follows:

$$\begin{aligned}
\text{MSE}_\beta = \mathsf{E}\left[\|\hat{\boldsymbol{x}}_\beta - \boldsymbol{x}_{\text{true}}\|^2\right] &= \mathsf{E}\left[\|(\hat{\boldsymbol{x}}_\beta - \bar{\boldsymbol{x}}_\beta) + (\bar{\boldsymbol{x}}_\beta - \boldsymbol{x})\|^2\right] && (2.5.4)\\
&= \mathsf{E}\left[\|\hat{\boldsymbol{x}}_\beta - \bar{\boldsymbol{x}}_\beta\|^2\right] + \|\bar{\boldsymbol{x}}_\beta - \boldsymbol{x}_{\text{true}}\|^2 && (2.5.5)\\
&= \mathsf{trace}\{\mathsf{Cov}\{\hat{\boldsymbol{x}}_\beta\}\} + \|\bar{\boldsymbol{x}}_\beta - \boldsymbol{x}_{\text{true}}\|^2 . && (2.5.6)
\end{aligned}$$

e,reg,hyper,mse,split

Thus the MSE depends on the sum of the variances and the sum of the squared biases of the estimates. If one decides to choose $\beta$ to minimize MSE, then one must somehow "balance" the variance and the bias contributions to MSE.

Neither of the above criteria ($\beta_{\text{O}}$ or $\beta_{\text{MSE}}$) can be used directly for real data because they depend on the true but unknown image $\boldsymbol{x}_{\text{true}}$. Hence they are sometimes called **oracle** or **clairvoyant** selection methods. But they can be explored in simulations (where $\boldsymbol{x}_{\text{true}}$ is known) to establish a baseline performance level. The other selection methods described hereafter typically try to approximate $\beta_{\text{MSE}}$ without using $\boldsymbol{x}_{\text{true}}$.

x,reg,hyper,pwls

**Example 2.5.1** *To explore the characteristics of the MSE approach (2.5.2), consider the linear measurement model* $\boldsymbol{y} = \bar{\boldsymbol{y}}(\boldsymbol{x}) + \boldsymbol{\varepsilon}$ *with* $\bar{\boldsymbol{y}}(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x}$ *and* $\boldsymbol{\varepsilon}$ *is zero mean with covariance* $\boldsymbol{W}^{-1}$*. For quadratic regularization* $\mathsf{R}(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}'\mathbf{R}\boldsymbol{x}$ *the* ***PWLS*** *estimator is* [2]

$$\hat{\boldsymbol{x}}_\beta = \arg\min_{\boldsymbol{x}} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_{\boldsymbol{W}^{1/2}}^2 + \beta\, \mathsf{R}(\boldsymbol{x}) = [\mathbf{F} + \beta\mathbf{R}]^{-1} \boldsymbol{A}'\boldsymbol{W}\boldsymbol{y},$$

*where* $\mathbf{F} = \boldsymbol{A}'\boldsymbol{W}\boldsymbol{A}$ *denotes the Fisher information matrix for this problem. (See §29.7.) In this case the mean is*

$$\bar{\boldsymbol{x}}_\beta = \mathsf{E}[\hat{\boldsymbol{x}}_\beta] = [\mathbf{F} + \beta\mathbf{R}]^{-1} \mathbf{F}\boldsymbol{x}_{\mathrm{true}},$$

*and the covariance is*

$$\mathsf{Cov}\{\hat{\boldsymbol{x}}_\beta\} = [\mathbf{F} + \beta\mathbf{R}]^{-1} \mathbf{F} [\mathbf{F} + \beta\mathbf{R}]^{-1}.$$

*Thus the* ***MSE*** *of this estimator is*

$$\mathrm{MSE}_\beta = \mathsf{trace}\left\{ [\mathbf{F} + \beta\mathbf{R}]^{-1} \mathbf{F} [\mathbf{F} + \beta\mathbf{R}]^{-1} \right\} + \left\| [\mathbf{F} + \beta\mathbf{R}]^{-1} \mathbf{F}\boldsymbol{x}_{\mathrm{true}} - \boldsymbol{x}_{\mathrm{true}} \right\|^2 \tag{2.5.7}$$

e,reg,hyper,pwls,mse

$$= \mathsf{trace}\left\{ [\mathbf{F} + \beta\mathbf{R}]^{-1} \mathbf{F} [\mathbf{F} + \beta\mathbf{R}]^{-1} \right\} + \beta^2 \left\| [\mathbf{F} + \beta\mathbf{R}]^{-1} \mathbf{R}\boldsymbol{x}_{\mathrm{true}} \right\|^2. \tag{2.5.8}$$

*In particular, if* $\beta = 0$*, then* $\mathrm{MSE}_0 = \mathsf{trace}\{\mathbf{F}^{-1}\}$ *as expected from §29.8.*

*Suppose* $\mathbf{F}$ *and* $\mathbf{R}$ *are both circulant*[3]*, with eigenvalues* $\mathsf{F}_k$ *and* $\mathsf{R}_k$ *respectively, and let* $X_k$ *denote the DFT of* $\boldsymbol{x}_{\mathrm{true}}$*. Then one can show (see Problem 2.6):*

e,reg,hyper,oracle,MSE

$$\mathrm{MSE}_\beta = \sum_k \frac{\mathsf{F}_k + \beta^2 \mathsf{R}_k^2 |X_k|^2 / n_{\mathrm{p}}}{(\mathsf{F}_k + \beta\mathsf{R}_k)^2}. \tag{2.5.9}$$

*In general there is no closed-form expression for* $\beta_{\mathrm{MSE}}$*, but one can find it numerically by minimizing* $\mathrm{MSE}_\beta$*.*

x,reg,hyper,pwls,I

**Example 2.5.2** *The simplest case is where* $\mathbf{R} = \boldsymbol{I}$ *and the columns of* $\sigma\boldsymbol{A}\boldsymbol{W}^{1/2}$ *are orthonormal,* i.e., $\mathbf{F} = \sigma^{-2}\boldsymbol{I}$*. Then (2.5.8) simplifies to*

e,reg,hyper,pwls,I,mse

$$\mathrm{MSE}_\beta = \frac{n_{\mathrm{p}}\sigma^{-2} + \beta^2 \|\boldsymbol{x}\|^2}{(\sigma^{-2} + \beta)^2}. \tag{2.5.10}$$

*Minimizing over* $\beta$ *per (2.5.2) yields* $\beta_{\mathrm{MSE}} = \dfrac{n_{\mathrm{p}}}{\|\boldsymbol{x}\|^2}$ *and* $\mathrm{MSE}_{\beta_{\mathrm{MSE}}} = \dfrac{n_{\mathrm{p}}\sigma^2}{1 + 1/\mathsf{SNR}}$*, where* $\mathsf{SNR} \triangleq \dfrac{\|\boldsymbol{x}\|^2}{n_{\mathrm{p}}\sigma^2}$*, and the estimator is* $\hat{\boldsymbol{x}}_{\beta_{\mathrm{MSE}}} = \frac{\sigma^2}{1+1/\mathsf{SNR}} \boldsymbol{A}'\boldsymbol{W}\boldsymbol{y}$*. This estimator is somewhat reminiscent of the* ***James-Stein shrinkage estimator*** *[72], which, for the case* $\boldsymbol{A} = \boldsymbol{W} = \boldsymbol{I}$ *and* $n_{\mathrm{p}} \geq 3$*, has the form:* $\hat{\boldsymbol{x}} = \left(1 - \frac{n_{\mathrm{p}} - 2}{\|\boldsymbol{y}\|^2}\right) \boldsymbol{y}$*.*

s,reg,hyper,rss

### 2.5.2 Residual sum of squares (s,reg,hyper,rss)

The estimation error (2.5.1) and its expectation (2.5.2) are defined in the domain of $\boldsymbol{x}$. Two other quantities of interest are the **predictive error**

$$\bar{\boldsymbol{y}}(\hat{\boldsymbol{x}}_\beta) - \bar{\boldsymbol{y}}(\boldsymbol{x}_{\mathrm{true}}),$$

defined in the data domain, and its expected (weighted) squared norm, called the **predictive risk** [73, p 97] [wiki]:

e,reg,hyper,rss,pr

$$\mathsf{PR}_\beta \triangleq \mathsf{E}\left[ \|\bar{\boldsymbol{y}}(\hat{\boldsymbol{x}}_\beta) - \bar{\boldsymbol{y}}(\boldsymbol{x}_{\mathrm{true}})\|_{\boldsymbol{W}^{1/2}}^2 \right]. \tag{2.5.11}$$

These quantities also depend on $\boldsymbol{x}_{\mathrm{true}}$ so cannot be used directly in practice for selecting $\beta$.

A quantity that *is* available in practice is the (weighted) **residual sum of squares** (**RSS**), defined in data space as:

e,reg,hyper,rss,def

$$\boxed{\mathsf{RSS}(\boldsymbol{x}) \triangleq \|\boldsymbol{y} - \bar{\boldsymbol{y}}(\boldsymbol{x})\|_{\boldsymbol{W}^{1/2}}^2.} \tag{2.5.12}$$

Several methods for regularization parameter selection are based on this quantity. (See [74] for an example showing how some such methods can be unstable.)

If the noise $\boldsymbol{y} - \bar{\boldsymbol{y}}(\boldsymbol{x}_{\mathrm{true}})$ has the gaussian distribution $\mathsf{N}(\boldsymbol{0}, \boldsymbol{W}^{-1})$, then $\mathsf{RSS}(\boldsymbol{x}_{\mathrm{true}})$ has a $\chi^2$ distribution with $n_{\mathrm{d}}$ degrees of freedom, so methods based on (2.5.12) are known as a $\chi^2$ **choice** for $\beta$ [65].

[2] For a positive definite matrix $\boldsymbol{H}$, the **weighted norm** $\|\cdot\|_{\boldsymbol{H}^{1/2}}$ is defined in terms of the **weighted inner product** $\langle \cdot, \cdot \rangle_{\boldsymbol{H}}$ as follows: $\|\boldsymbol{x}\|_{\boldsymbol{H}^{1/2}}^2 \triangleq \langle \boldsymbol{x}, \boldsymbol{x} \rangle_{\boldsymbol{H}} = \boldsymbol{x}'\boldsymbol{H}\boldsymbol{x} = \left\|\boldsymbol{H}^{1/2}\boldsymbol{x}\right\|^2$, where $\langle \boldsymbol{u}, \boldsymbol{v} \rangle_{\boldsymbol{H}} \triangleq \boldsymbol{v}'\,\boldsymbol{H}\,\boldsymbol{u}$.

[3] It suffices for $\mathbf{F}$ and $\mathbf{R}$ to have the same orthonormal eigenvectors, $\mathbf{F} = \boldsymbol{V}\,\mathsf{diag}\{\mathsf{F}_k\}\,\boldsymbol{V}'$ and $\mathbf{R} = \boldsymbol{V}\,\mathsf{diag}\{\mathsf{R}_k\}\,\boldsymbol{V}'$, with corresponding eigenvalues $\{\mathsf{F}_k\}$ and $\{\mathsf{R}_k\}$, in which case $\boldsymbol{X} = \sqrt{n_{\mathrm{p}}}\boldsymbol{V}'\boldsymbol{x}_{\mathrm{true}}$ in (2.5.9). This same generality applies hereafter to other "circulant" cases.

x,reg,hyper,rss,A

**Example 2.5.3** *To explore the characteristics of* RSS *and* $\mathsf{PR}_\beta$*, consider the linear model* $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{\varepsilon}$*, where* $\boldsymbol{\varepsilon}$ *has mean zero and covariance* $\boldsymbol{W}^{-1}$ *(and is not necessarily gaussian). Let* $\boldsymbol{b}_\beta \triangleq \mathsf{E}[\hat{\boldsymbol{x}}_\beta] - \boldsymbol{x}_{\mathrm{true}}$ *denote the estimator bias, and define the zero-mean "estimator noise" random vector* $\boldsymbol{z}_\beta \triangleq \hat{\boldsymbol{x}}_\beta - \mathsf{E}[\hat{\boldsymbol{x}}_\beta]$*. One can show (Problem 2.7) that*

e,reg,hyper,rss,A

$$\mathsf{RSS}(\hat{\boldsymbol{x}}_\beta) = (\boldsymbol{\varepsilon} - \boldsymbol{A}\boldsymbol{z}_\beta)'\boldsymbol{W}(\boldsymbol{\varepsilon} - \boldsymbol{A}\boldsymbol{z}_\beta) - 2(\boldsymbol{\varepsilon} - \boldsymbol{A}\boldsymbol{z}_\beta)'\boldsymbol{W}\boldsymbol{A}\boldsymbol{b}_\beta + \boldsymbol{b}_\beta'\mathsf{F}\boldsymbol{b}_\beta. \tag{2.5.13}$$

*Because* $\boldsymbol{z}_\beta$ *and* $\boldsymbol{\varepsilon}$ *are zero mean, it follows that*

$$\mathsf{E}[\mathsf{RSS}(\hat{\boldsymbol{x}}_\beta)] = \mathsf{E}[(\boldsymbol{\varepsilon} - \boldsymbol{A}\boldsymbol{z}_\beta)'\boldsymbol{W}(\boldsymbol{\varepsilon} - \boldsymbol{A}\boldsymbol{z}_\beta)] + \boldsymbol{b}_\beta'\mathsf{F}\boldsymbol{b}_\beta.$$

*The first term is due to variability and the second term is due to bias of the estimator* $\hat{\boldsymbol{x}}_\beta$*.*

*Similarly, one can show that the predictive risk for this linear model is*

e,reg,hyper,rss,A,pr

$$\mathsf{PR}_\beta = \mathsf{E}\left[\boldsymbol{z}_\beta'\mathsf{F}\boldsymbol{z}_\beta\right] + \boldsymbol{b}_\beta'\mathsf{F}\boldsymbol{b}_\beta. \tag{2.5.14}$$

x,reg,hyper,rss,lin

**Example 2.5.4** *Specialize the previous example by considering* linear *estimators* $\hat{\boldsymbol{x}}_\beta = \boldsymbol{L}_\beta\boldsymbol{y}$*, for which* $\boldsymbol{z}_\beta = \boldsymbol{L}_\beta\boldsymbol{\varepsilon}$ *and* $\boldsymbol{b}_\beta = \left(\boldsymbol{L}_\beta\boldsymbol{A} - \boldsymbol{I}_{n_\mathrm{p}}\right)\boldsymbol{x}_{\mathrm{true}}$*. One can show (Problem 2.7) that*

e,reg,hyper,rss,lin,r

$$\mathsf{RSS}(\hat{\boldsymbol{x}}_\beta) = \left\|\left(\boldsymbol{I} - \mathsf{M}(\beta)\right)\boldsymbol{W}^{1/2}\boldsymbol{y}\right\|^2, \tag{2.5.15}$$

e,reg,hyper,rss,lin,e

$$\mathsf{E}[\mathsf{RSS}(\hat{\boldsymbol{x}}_\beta)] = \mathsf{trace}\left\{\left(\boldsymbol{I}_{n_\mathrm{d}} - \mathsf{M}\right)'\left(\boldsymbol{I}_{n_\mathrm{d}} - \mathsf{M}\right)\right\} + \boldsymbol{x}_{\mathrm{true}}'\left(\boldsymbol{L}_\beta\boldsymbol{A} - \boldsymbol{I}_{n_\mathrm{p}}\right)'\mathsf{F}\left(\boldsymbol{L}_\beta\boldsymbol{A} - \boldsymbol{I}_{n_\mathrm{p}}\right)\boldsymbol{x}_{\mathrm{true}}, \tag{2.5.16}$$

*where the* $n_\mathrm{d} \times n_\mathrm{d}$ ***influence matrix*** *or* ***hat matrix*** *of the linear estimator is denoted*

e,reg,hyper,rss,iMr

$$\mathsf{M}(\beta) \triangleq \boldsymbol{W}^{1/2}\boldsymbol{A}\boldsymbol{L}_\beta\boldsymbol{W}^{-1/2}. \tag{2.5.17}$$

*Similarly, one can show that the predictive risk for this linear estimator is*

e,reg,hyper,rss,lin,pr

$$\mathsf{PR}_\beta = \mathsf{trace}\left\{\mathsf{M}'(\beta)\,\mathsf{M}(\beta)\right\} + \boldsymbol{b}_\beta'\mathsf{F}\boldsymbol{b}_\beta. \tag{2.5.18}$$

x,reg,hyper,rss,back

**Example 2.5.5** *To further specialize, consider linear estimators of the form* $\boldsymbol{L}_\beta = \boldsymbol{B}_\beta\boldsymbol{A}'\boldsymbol{W}$*, for some* $n_\mathrm{p} \times n_\mathrm{p}$ *matrix* $\boldsymbol{B}_\beta$*, for which* $\boldsymbol{L}_\beta\boldsymbol{A} = \boldsymbol{B}_\beta\mathsf{F}$*. Defining* $\boldsymbol{d} \triangleq \boldsymbol{A}'\boldsymbol{W}\boldsymbol{y}$*, one can show (Problem 2.7) that*

e,reg,hyper,rss,back,r

$$\mathsf{RSS}(\hat{\boldsymbol{x}}_\beta) = \|\boldsymbol{y}\|^2_{\boldsymbol{W}^{1/2}} - 2\boldsymbol{d}'\boldsymbol{B}_\beta\boldsymbol{d} + \boldsymbol{d}'\boldsymbol{B}_\beta'\mathsf{F}\boldsymbol{B}_\beta\boldsymbol{d} \tag{2.5.19}$$

e,reg,hyper,rss,back,e

$$\begin{aligned}\mathsf{E}[\mathsf{RSS}(\hat{\boldsymbol{x}}_\beta)] &= n_\mathrm{d} - n_\mathrm{p} + \mathsf{trace}\left\{\left(\boldsymbol{I}_{n_\mathrm{p}} - \mathsf{F}^{1/2}\boldsymbol{B}_\beta'\mathsf{F}^{1/2}\right)\left(\boldsymbol{I}_{n_\mathrm{p}} - \mathsf{F}^{1/2}\boldsymbol{B}_\beta\mathsf{F}^{1/2}\right)\right\} \\ &\quad + \boldsymbol{x}_{\mathrm{true}}'\left(\boldsymbol{B}_\beta\mathsf{F} - \boldsymbol{I}\right)'\mathsf{F}\left(\boldsymbol{B}_\beta\mathsf{F} - \boldsymbol{I}\right)\boldsymbol{x}_{\mathrm{true}}. \end{aligned} \tag{2.5.20}$$

*In particular, if* $\boldsymbol{B}_\beta = \left[\mathsf{F} + \beta\boldsymbol{R}\right]^{-1}$ *where* $\mathsf{F}$ *and* $\boldsymbol{R}$ *are both circulant, then (Problem 2.7)*

e,reg,hyper,rss,back,qpwls

$$\mathsf{RSS}(\hat{\boldsymbol{x}}_\beta) = \|\boldsymbol{y}\|^2_{\boldsymbol{W}^{1/2}} - \frac{1}{n_\mathrm{p}}\sum_k \frac{|D_k|^2\left(\mathsf{F}_k + 2\beta\mathsf{R}_k\right)}{\left(\mathsf{F}_k + \beta\mathsf{R}_k\right)^2} \tag{2.5.21}$$

*where* $D_k$ *denotes the* $n_\mathrm{p}$*-point DFT of* $\boldsymbol{d}$*.*

*As a special case of (2.5.20), if* $\mathsf{F}$ *is invertible and* $\boldsymbol{B}_\beta = \mathsf{F}^{-1}$*, then* $\mathsf{E}[\mathsf{RSS}(\hat{\boldsymbol{x}}_\beta)] = n_\mathrm{d} - n_\mathrm{p}$ *which is standard for LS fitting of* $n_\mathrm{p}$ *model parameters to* $n_\mathrm{d}$ *data points.*

s,reg,hyper,rss,dp

### 2.5.2.1 Discrepancy principle

For the measurement model $\boldsymbol{y} = \bar{\boldsymbol{y}}(\boldsymbol{x}) + \boldsymbol{\varepsilon}$ where the noise $\boldsymbol{\varepsilon}$ is zero mean with covariance $\boldsymbol{W}^{-1}$ (and not necessarily gaussian), then the **residual sum of squares** (**RSS**) evaluated at the *true* image $\boldsymbol{x}_{\mathrm{true}}$ satisfies:

$$\mathsf{E}[\mathsf{RSS}(\boldsymbol{x}_{\mathrm{true}})] = n_\mathrm{d}.$$

This equality suggests the following **discrepancy principle** [1, 75] for selecting $\beta$:

e,reg,hyper,dp,min

$$\boxed{\beta_{\mathrm{DP}} = \arg\min_\beta \left|\mathsf{RSS}(\hat{\boldsymbol{x}}_\beta) - n_\mathrm{d}\right|.} \tag{2.5.22}$$

Although this method is appealingly simple, it is known to produce $\beta$ values that over smooth [64]. Typically $\|\boldsymbol{y} - \bar{\boldsymbol{y}}(\hat{\boldsymbol{x}}_{\beta_{\mathrm{MSE}}})\|^2_{\boldsymbol{W}^{1/2}} < n_\mathrm{d}$, so usually $\beta_{\mathrm{DP}} > \beta_{\mathrm{MSE}}$, causing over smoothing [64]. Furthermore, $\beta_{\mathrm{DP}}$ requires knowledge of the data (co)variance $\boldsymbol{W}^{-1}$ which is not always available.

x,reg,hyper,dp,qpwls

**Example 2.5.6** *To explore this approach, consider the QPWLS estimator* $\hat{\boldsymbol{x}}_\beta = [\mathsf{F} + \beta \mathsf{R}]^{-1} \boldsymbol{A}' \boldsymbol{W} \boldsymbol{y}$ *of Example 2.5.1. This is the case of Example 2.5.5 where* $\boldsymbol{B}_\beta = [\mathsf{F} + \beta \mathsf{R}]^{-1}$ *and (2.5.20) applies directly. When* $\mathsf{F}$ *and* $\boldsymbol{R}$ *are both circulant:*

$$\begin{aligned}
\mathsf{E}[\mathsf{RSS}(\hat{\boldsymbol{x}}_\beta)] &= n_\mathrm{d} - n_\mathrm{p} + \sum_k \left(1 - \frac{\mathsf{F}_k}{\mathsf{F}_k + \beta \mathsf{R}_k}\right)^2 + \frac{|X_k|^2}{n_\mathrm{p}} \mathsf{F}_k \left(\frac{\mathsf{F}_k}{\mathsf{F}_k + \beta \mathsf{R}_k} - 1\right)^2 \\
&= n_\mathrm{d} - n_\mathrm{p} + \beta^2 \sum_k \mathsf{R}_k^2 \frac{1 + \mathsf{F}_k |X_k|^2 / n_\mathrm{p}}{(\mathsf{F}_k + \beta \mathsf{R}_k)^2} \\
&= n_\mathrm{d} - \mathsf{rank}\{\mathsf{F}\} + \beta^2 \sum_{k \,:\, \mathsf{F}_k \neq 0} \mathsf{R}_k^2 \frac{1 + \mathsf{F}_k |X_k|^2 / n_\mathrm{p}}{(\mathsf{F}_k + \beta \mathsf{R}_k)^2},
\end{aligned} \tag{2.5.23}$$

e,reg,hyper,rss,circ

*where the* ***effective model order*** *(for* $\beta = 0$*) is* $\mathsf{rank}\{\mathsf{F}\} = n_\mathrm{p} - |\{k \,:\, \mathsf{F}_k = 0\}|$.

*As* $\beta \to 0$ *the summation approaches* $n_\mathrm{d} - \mathsf{rank}\{\mathsf{F}\}$ *and as* $\beta \to \infty$ *it approaches*

$$n_\mathrm{d} - |\{k \,:\, \mathsf{R}_k = 0\}| + \sum_{\{k \,:\, \mathsf{R}_k \neq 0\}} \left(\mathsf{F}_k |X_k|^2 / n_\mathrm{p}\right). \tag{2.5.24}$$

e,reg,hyper,rss,reg,inf

*Usually these two extremes straddle* $n_\mathrm{d}$ *so there will be an intermediate value of* $\beta$ *that satisfies (2.5.22).*

x,reg,hyper,rss,dp,i

**Example 2.5.7** *In the orthogonal case where* $\mathsf{F} = \sigma^{-2} \boldsymbol{I}$ *and* $\mathsf{R} = \boldsymbol{I}$*, one can show that* (cf. *[64, eqn. (2.6)]):*

$$\mathsf{E}[\mathsf{RSS}(\hat{\boldsymbol{x}}_\beta)] = n_\mathrm{d} - n_\mathrm{p} + n_\mathrm{p} (1 + \mathsf{SNR}) \left(\frac{\sigma^2 \beta}{1 + \sigma^2 \beta}\right)^2. \tag{2.5.25}$$

e,reg,hyper,rss,dp,i

*Equating to* $n_\mathrm{d}$ *and solving yields*

$$\beta_\mathrm{DP}^* = \frac{\sigma^{-2}}{\sqrt{1 + \mathsf{SNR}} - 1}. \tag{2.5.26}$$

e,reg,hyper,dp,min,I

*One can show that* $\beta_\mathrm{DP}^* > \beta_\mathrm{MSE}$ *in this special case. Despite this drawback of the discrepancy principle, it continues to resurface in the imaging literature,* e.g., *[76]. See also Problem 2.8.*

For data with Poisson noise, related methods based on a **discrepancy principle** have been investigated [77–80].

### 2.5.2.2 Residual effective degrees of freedom (REDF) method

Using $n_\mathrm{d}$ in (2.5.22) unwisely ignores the fact that typically $\mathsf{RSS}(\hat{\boldsymbol{x}}_\beta) < \mathsf{RSS}(\boldsymbol{x}_\mathrm{true})$ because $\hat{\boldsymbol{x}}_\beta$ will fit both the signal *and* the noise in the data. An alternative that accounts for this fitting is the **residual effective degrees of freedom** (**REDF**) method [64, 81]:

$$\boxed{\beta_\mathrm{REDF} \triangleq \arg\min_{\beta > 0} \left| \mathsf{RSS}(\hat{\boldsymbol{x}}_\beta) - \mathsf{REDF}(\beta) \right|.} \tag{2.5.27}$$

e,reg,hyper,edf,min

There are various definitions of REDF [wiki]; for a linear model with a linear estimator, a natural choice based on (2.5.16) is

$$\mathsf{REDF}(\beta) \triangleq \mathsf{trace}\left\{(\boldsymbol{I}_{n_\mathrm{d}} - \mathsf{M}(\beta))' (\boldsymbol{I}_{n_\mathrm{d}} - \mathsf{M}(\beta))\right\} = n_\mathrm{d} - 2\,\mathsf{trace}\{\mathsf{M}(\beta)\} + \mathsf{trace}\left\{\mathsf{M}'(\beta)\,\mathsf{M}(\beta)\right\}. \tag{2.5.28}$$

e,reg,hyper,edf

Another popular choice is

$$\mathsf{REDF}(\beta) \triangleq n_\mathrm{d} - \mathsf{trace}\{\mathsf{M}(\beta)\} = \mathsf{trace}\{\boldsymbol{I}_{n_\mathrm{d}} - \mathsf{M}(\beta)\}. \tag{2.5.29}$$

e,reg,hyper,edf,simple

These definitions match when $\mathsf{M}$ is **idempotent**. (See [82] for complications in non-convex models.) For a well-conditioned problem, $\mathsf{REDF}(0) = n_\mathrm{d} - n_\mathrm{p}$, which is the usual (residual) **degrees of freedom** in a regression problem with $n_\mathrm{d}$ measurements and $n_\mathrm{p}$ unknowns.

x,reg,hyper,redf,qpwls

**Example 2.5.8** *For the QPWLS estimator, the* ***influence matrix*** *is*

$$\mathsf{M}(\beta) = \boldsymbol{W}^{1/2} \boldsymbol{A} [\mathsf{F} + \beta \mathsf{R}]^{-1} \boldsymbol{A}' \boldsymbol{W}^{1/2}. \tag{2.5.30}$$

e,reg,hyper,influence

*In particular,* $\mathsf{trace}\{\mathsf{M}(\beta)\} = \mathsf{trace}\left\{[\mathsf{F} + \beta \mathsf{R}]^{-1} \mathsf{F}\right\}$.

*To analyze this approach, again it is simpler to consider the expected* RSS*:*

$$\beta_\mathrm{REDF}^* \triangleq \arg\min_{\beta > 0} \left| \mathsf{E}[\mathsf{RSS}(\hat{\boldsymbol{x}}_\beta)] - \mathsf{REDF}(\beta) \right|.$$

*If* $\mathsf{F}$ *and* $\boldsymbol{R}$ *are both circulant, then with the definition (2.5.29):*

$$\mathsf{REDF}(\beta) = n_\mathrm{d} - \sum_k \frac{\mathsf{F}_k}{\mathsf{F}_k + \beta \mathsf{R}_k} = n_\mathrm{d} - \mathsf{rank}\{\mathsf{F}\} + \sum_{k\,:\,\mathsf{F}_k \neq 0} \frac{\beta \mathsf{R}_k}{\mathsf{F}_k + \beta \mathsf{R}_k}. \tag{2.5.31}$$

e,reg,hyper,rss,redf,circ

*Using (2.5.23):*

$$\begin{aligned}
\mathsf{E}[\mathsf{RSS}(\hat{\boldsymbol{x}}_\beta)] - \mathsf{REDF}(\beta) &= \sum_{k\,:\,\mathsf{F}_k \neq 0} \beta^2 \mathsf{R}_k^2 \frac{1 + \mathsf{F}_k \left|X_k\right|^2 / n_\mathrm{p}}{\left(\mathsf{F}_k + \beta \mathsf{R}_k\right)^2} - \sum_{k\,:\,\mathsf{F}_k \neq 0} \frac{\beta \mathsf{R}_k \left(\mathsf{F}_k + \beta \mathsf{R}_k\right)}{\left(\mathsf{F}_k + \beta \mathsf{R}_k\right)^2} \\
&= \sum_k \frac{\beta \mathsf{R}_k \mathsf{F}_k \left(\beta \mathsf{R}_k \left|X_k\right|^2 / n_\mathrm{p} - 1\right)}{\left(\mathsf{F}_k + \beta \mathsf{R}_k\right)^2}.
\end{aligned}$$

*In the orthogonal case where* $\mathsf{F} = \sigma^{-2} \boldsymbol{I}$ *and* $\mathsf{R} = \boldsymbol{I}$*,*

$$\mathsf{E}[\mathsf{RSS}(\hat{\boldsymbol{x}}_\beta)] - \mathsf{REDF}(\beta) = \frac{\beta \sigma^{-2}}{\left(\sigma^{-2} + \beta\right)^2} \sum_k \left(\beta \left|X_k\right|^2 / n_\mathrm{p} - 1\right) = \frac{\beta \sigma^{-2}}{\left(\sigma^{-2} + \beta\right)^2} \left(\beta \left\|\boldsymbol{x}\right\|^2 - n_\mathrm{p}\right).$$

*Equating to zero (ignoring the trivial solution* $\beta = 0$*) yields [64, eqn. (2.9)]:* $\beta_{\mathrm{REDF}}^* = n_\mathrm{p} / \left\|\boldsymbol{x}\right\|^2 = \beta_{\mathrm{MSE}}$*. See [64, 65] and Problem 2.9 for more approaches and related analyses.*

x,reg,hyper,rss,redf

**Example 2.5.9** *Fig. 2.5.1 shows an example of* $n_\mathrm{d} = 100$ *noisy samples of one cycle of a sinusoid denoised by fitting polynomials of various degrees (*$n_\mathrm{p} = 1+$ *degree) and by using a quadratic roughness penalty* $\hat{\boldsymbol{x}}_\beta = \left[\boldsymbol{I} + \beta \boldsymbol{C}' \boldsymbol{C}\right]^{-1} \boldsymbol{y}$ *with periodic boundary conditions. As polynomial degree increases, naturally* RSS *decreases. Similarly, as regularization parameter* $\beta$ *decreases,* RSS *decreases. The* $\beta_{\mathrm{REDF}}$ *based on (2.5.27) and the corresponding value for the polynomial fit are marked with stars. The RMSE plot shows that in this case the REDF criterion picked nearly the option* $\beta$ *for the regularized method, but a slightly higher degree polynomial than would have been best here.*

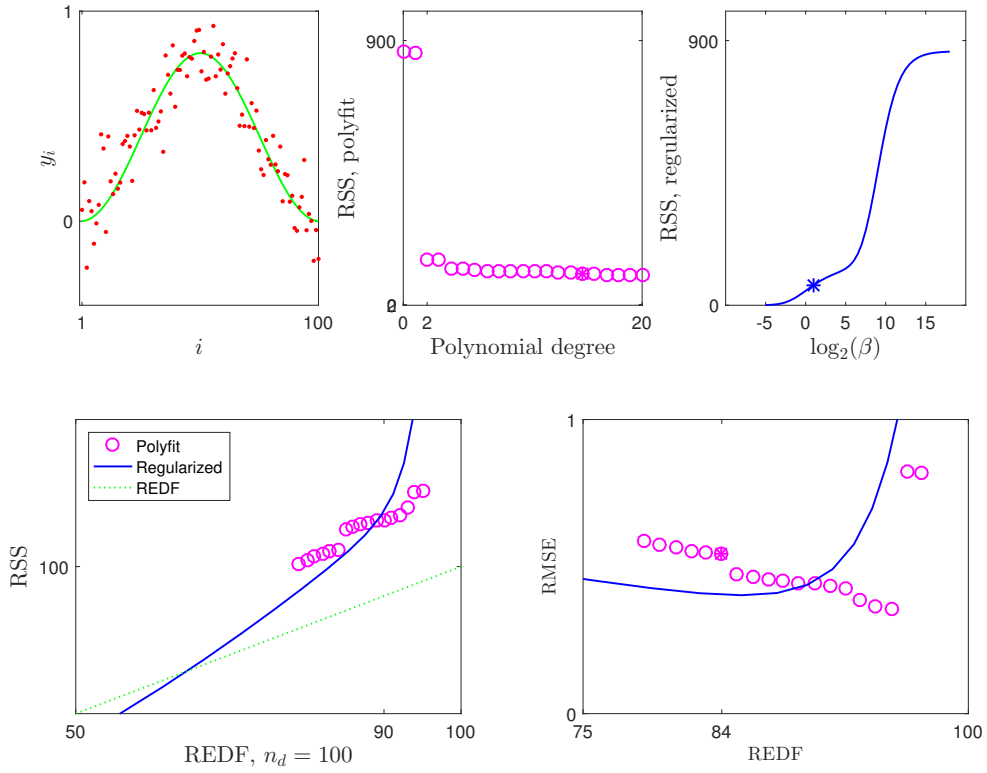



Figure 2.5.1: Using REDF for selecting polynomial order and regularization parameter for a simple denoising problem.

fig_reg_redf

### 2.5.2.3 Unbiased predictive risk estimator (UPRE)

Yet another variation is the **unbiased predictive risk estimator** (**UPRE**) [73, p. 98] that minimizes

$$\mathsf{UPRE}_\beta \triangleq \mathsf{RSS}(\hat{\boldsymbol{x}}_\beta) + 2\,\mathsf{trace}\{\mathsf{M}(\beta)\} - n_\mathrm{d}. \tag{2.5.32}$$

e,reg,hyper,upre

For the linear measurement model and linear estimator considered in Example 2.5.4, one can verify that $\mathsf{UPRE}_\beta$ is an unbiased estimate of the **predictive risk**, *i.e.*, $\mathsf{E}[\mathsf{UPRE}_\beta] = \mathsf{PR}_\beta$, by comparing (2.5.16) and (2.5.18).

s,reg,hyper,cv

### 2.5.3 Cross validation method (s,reg,hyper,cv)

In **cross validation** methods, we set aside part of the data, perform model fitting on the rest, and then see how well the fitted model predicts the data that we set aside. The idea is that if $\beta$ is too small or too large, then the predictions of the data values that were set aside will be worse than if $\beta$ is chosen appropriately. The simplest form is called **leave-one-out** cross validation, and is our focus here [83, 84].

Let $\hat{\boldsymbol{x}}_\beta^{(-i)}$ denote the estimate that is formed using all the data *except* $y_i$. For the model in Example 2.5.1:

$$\hat{\boldsymbol{x}}_\beta^{(-i)} = \arg\min_{\boldsymbol{x}} \sum_{k \neq i} w_k \left| y_k - [\boldsymbol{A}\boldsymbol{x}]_k \right|^2 + \beta \boldsymbol{x}' \mathbf{R} \boldsymbol{x} \tag{2.5.33}$$

e,reg,hyper,cv,xhri

$$= \left[ \boldsymbol{A}' \boldsymbol{W}_{(-i)} \boldsymbol{A} + \beta \mathbf{R} \right]^{-1} \boldsymbol{A}' \boldsymbol{W}_{(-i)} \boldsymbol{y}, \tag{2.5.34}$$

where $\boldsymbol{W}_{(-i)} = \boldsymbol{W} - w_i \boldsymbol{e}_i \boldsymbol{e}_i' = \boldsymbol{W} \left( \boldsymbol{I} - \boldsymbol{e}_i \boldsymbol{e}_i' \right)$ is like $\boldsymbol{W}$ but with a 0 in its $i$th diagonal element. To choose $\beta$, we compare the "left out" data value $y_i$ with its predicted value $\bar{y}_i\left(\hat{\boldsymbol{x}}_\beta^{(-i)}\right)$ as follows:

e,reg,hyper,cv

$$\boxed{\beta_{\mathrm{CV}} = \arg\min_\beta \Phi_{\mathrm{CV}}(\beta),} \qquad \boxed{\Phi_{\mathrm{CV}}(\beta) \triangleq \sum_{i=1}^{n_\mathrm{d}} w_i \left| y_i - \bar{y}_i\left(\hat{\boldsymbol{x}}_\beta^{(-i)}\right) \right|^2.} \tag{2.5.35}$$

Apparently this would be a computationally intensive procedure because it appears to require that one perform $n_\mathrm{d}$ separate estimations for each value of $\beta$. However, one can show (see Problem 2.14) for linear problems that[4]

e,reg,hyper,cv,ybi

$$\bar{y}_i\left(\hat{\boldsymbol{x}}_\beta^{(-i)}\right) = \boldsymbol{a}_i' \hat{\boldsymbol{x}}_\beta^{(-i)} = \frac{1}{1 - \mathsf{M}_{ii}(\beta)} \left( \boldsymbol{a}_i' \hat{\boldsymbol{x}}_\beta - \mathsf{M}_{ii}(\beta)\, y_i \right), \tag{2.5.36}$$

where $\mathsf{M}_{ii}(\beta)$ is the $i$th diagonal element of the influence matrix in (2.5.30). Thus, the summation in (2.5.35) simplifies to the following form [10, p. 51] [85]:

e,reg,hyper,cv,simple

$$\Phi_{\mathrm{CV}}(\beta) = \sum_{i=1}^{n_\mathrm{d}} w_i \frac{\left| y_i - \bar{y}_i(\hat{\boldsymbol{x}}_\beta) \right|^2}{\left(1 - \mathsf{M}_{ii}(\beta)\right)^2}. \tag{2.5.37}$$

Although this expression appears simpler than (2.5.35), it remains impractical because the influence matrix $\mathbf{M}(\beta)$ is too large for imaging problems. See Problem 2.10.

Cross validation methods have been reported to have undesirable variability, though variance reduction methods have been proposed [86].

A variation on CV is called **estimation stability with cross validation** (**ESCV**) [87]. This method examines the "estimation stability" of the estimates obtained by each leave-one-out estimator:

$$\Phi_{\mathrm{ES}}(\beta) = \frac{\sum_{i=1}^{n_\mathrm{d}} \left\| \hat{\boldsymbol{x}}_\beta^{(-i)} - \hat{\boldsymbol{x}}_\beta \right\|^2}{\left\| \sum_{i=1}^{n_\mathrm{d}} \hat{\boldsymbol{x}}_\beta^{(-i)} \right\|^2}.$$

Instead of simply minimizing $\Phi_{\mathrm{ES}}$ over $\beta$, which could lead to over-smoothing, the procedure is to choose $\beta_{\mathrm{ES}}$ as a local minimizer of $\Phi_{\mathrm{ES}}$ that is *smaller* than $\beta_{\mathrm{CV}}$. This choice can compensate for the tendency of $\beta_{\mathrm{CV}}$ to be too large.

s,reg,hyper,gcv

#### 2.5.3.1 Generalized cross validation (GCV) (s,reg,hyper,gcv)

The ordinary cross validation method is not invariant to orthonormal transformations (rotations) of the data, *i.e.*, $\boldsymbol{y} \mapsto \boldsymbol{Q}\boldsymbol{y}$ and $\boldsymbol{A} \mapsto \boldsymbol{Q}\boldsymbol{A}$, for some orthonormal matrix $\boldsymbol{Q}$, even if $\boldsymbol{W} = \boldsymbol{I}$. This lack of invariance motivated the development of the **generalized cross validation** (**GCV**) method [85, 88]:

e,reg,hyper,gcv

$$\boxed{\beta_{\mathrm{GCV}} \triangleq \arg\min_\beta \Phi_{\mathrm{GCV}}(\beta),} \qquad \boxed{\Phi_{\mathrm{GCV}}(\beta) \triangleq \sum_{i=1}^{n_\mathrm{d}} w_i \frac{\left| y_i - \bar{y}_i(\hat{\boldsymbol{x}}_\beta) \right|^2}{\left(1 - \bar{\mathsf{M}}(\beta)\right)^2},} \tag{2.5.38}$$

where $\bar{\mathsf{M}}(\beta) \triangleq \frac{1}{n_\mathrm{d}} \sum_{i=1}^{n_\mathrm{d}} \mathsf{M}_{ii}(\beta) = \frac{1}{n_\mathrm{d}} \mathsf{trace}\{\mathbf{M}(\beta)\}$ is the average value of the diagonal elements of the influence matrix. It is useful to write $\Phi_{\mathrm{GCV}}$ using the definition of REDF in (2.5.29) as follows:

e,reg,hyper,gcv,redf

$$\boxed{\Phi_{\mathrm{GCV}}(\beta) = n_\mathrm{d}^2 \frac{\mathsf{RSS}(\hat{\boldsymbol{x}}_\beta)}{\mathsf{REDF}^2(\beta)}.} \tag{2.5.39}$$

[4] One can show that $\mathsf{M}_{ii}(\beta) < 1$ for $\beta > 0$ for (2.5.34), so the ratio is well defined. (See Problem 2.14.) More generally, $0 < \mathsf{M}_{ii}(\beta) < 1$ for useful estimators.

Both RSS and REDF decrease as $\beta \to 0$.

GCV has various optimality properties [89] [10, p. 55] for linear problems. See §2.5.3.3 for nonlinear extensions. Unfortunately, GCV is prohibitively expensive computationally to evaluate exactly for imaging problems. However, see §2.5.3.2 for approximations to $\Phi_{\mathrm{GCV}}$ based on stochastic methods that are feasible for imaging problems. GCV has been used to optimize not only the regularization parameter $\beta$, but also other parameters of the regularizer [90] and of the blur [91].

x,reg,hyper,gcv,circ

**Example 2.5.10** *Continuing Example 2.5.1, if $\boldsymbol{A}$ and $\mathsf{R}$ are both circulant, with eigenvalues $B_k$ and $\mathsf{R}_k$ respectively, and if $\boldsymbol{W} = \sigma^{-2}\boldsymbol{I}$, then the diagonal elements of the influence matrix simplify to the same value:*

$$\mathsf{M}_{ii}(\beta) = \frac{1}{n_{\mathrm{p}}} \sum_k \frac{|B_k|^2 / \sigma^2}{|B_k|^2 / \sigma^2 + \beta \mathsf{R}_k}.$$

*In this special case, $\Phi_{\mathrm{CV}} = \Phi_{\mathrm{GCV}}$. See [66, eqn. (19)] for further details.*

*Slightly more generally, if $\mathsf{F}$ and $\mathsf{R}$ are both circulant, with eigenvalues $\mathsf{F}_k$ and $\mathsf{R}_k$ respectively, then using (26.1.7):*

$$\bar{\mathsf{M}}(\beta) = \frac{1}{n_{\mathrm{d}}} \mathsf{trace}\{\mathbf{M}(\beta)\} = \frac{1}{n_{\mathrm{d}}} \mathsf{trace}\left\{\mathbf{F}\left[\mathbf{F} + \beta \mathbf{R}\right]^{-1}\right\} = \frac{1}{n_{\mathrm{d}}} \sum_k \frac{\mathsf{F}_k}{\mathsf{F}_k + \beta \mathsf{R}_k}. \tag{2.5.40}$$

e,reg,hyper,gcv,m,circ

*Note that $\bar{\mathsf{M}}(\beta) \to \mathsf{rank}\{\mathbf{F}\}/n_{\mathrm{d}}$ as $\beta \to 0$. One could use (2.5.40) to evaluate $\Phi_{\mathrm{GCV}}$ in large (linear) problems that are locally shift invariant. Interestingly, because the ratio inside the above summation is the frequency response of the estimator, the value of $\bar{\mathsf{M}}(\beta)$ in this case is proportional to the central value of the PSF (1.9.2).*

*Combining (2.5.40) and (2.5.21), the GCV criterion in the circulant case is*

$$\Phi_{\mathrm{GCV}}(\beta) = \frac{\|\boldsymbol{y}\|_{\boldsymbol{W}^{1/2}}^2 - \frac{1}{n_{\mathrm{p}}} \sum_k \frac{|D_k|^2 (\mathsf{F}_k + 2\beta \mathsf{R}_k)}{(\mathsf{F}_k + \beta \mathsf{R}_k)^2}}{\left(1 - \frac{1}{n_{\mathrm{d}}} \sum_k \frac{\mathsf{F}_k}{\mathsf{F}_k + \beta \mathsf{R}_k}\right)^2} \tag{2.5.41}$$

e,reg,hyper,gcv,phi,circ

*where $D_k$ is the $n_{\mathrm{p}}$-point DFT of $\boldsymbol{d} = \boldsymbol{A}'\boldsymbol{W}\boldsymbol{y}$. One can minimize this over $\beta$ numerically. See Problem 2.11.*

x,res,hyper,cv,orth

**Example 2.5.11** *Continuing Example 2.5.2, if $\mathbf{F} = \sigma^{-2}\boldsymbol{I}$ and $\mathbf{R} = \boldsymbol{I}$ then $\bar{\mathsf{M}}(\beta) = \frac{1}{n_{\mathrm{d}}} \mathsf{trace}\{\mathbf{M}(\beta)\} = \frac{n_{\mathrm{p}}}{n_{\mathrm{d}}} \frac{1}{1+\sigma^2\beta}$ so using (2.5.25)*

$$\mathsf{E}[\Phi_{\mathrm{GCV}}(\beta)] = \frac{n_{\mathrm{d}} - n_{\mathrm{p}} + n_{\mathrm{p}} (1 + \mathsf{SNR}) \left(\frac{\sigma^2\beta}{1+\sigma^2\beta}\right)^2}{\left(1 - \frac{n_{\mathrm{p}}}{n_{\mathrm{d}}} \frac{1}{1+\sigma^2\beta}\right)^2} = n_{\mathrm{d}} \frac{(1-f)(1+\gamma)^2 + f(1+\mathsf{SNR})\gamma^2}{(1+\gamma-f)^2}$$

*where $\gamma \triangleq \sigma^2\beta$ and $f = n_{\mathrm{p}}/n_{\mathrm{d}}$. One can show the minimizer is $\beta_{\mathrm{GCV}}^* = \sigma^{-2}/\mathsf{SNR} = n_{\mathrm{p}}/\|\boldsymbol{x}\|^2 = \beta_{\mathrm{MSE}}$, so at least for this highly idealized case, minimizing the (expectation) of $\Phi_{\mathrm{GCV}}$ provides the MSE-optimal value of $\beta$, unlike the discrepancy principle choice in (2.5.26).*

s,reg,hyper,trace

### 2.5.3.2 Monte Carlo methods for matrix trace (s,reg,hyper,trace)

Several of the preceding expressions depend on the trace of a (large) square matrix, namely the influence matrix in (2.5.32), (2.5.29) and (2.5.38). For circulant problems one can compute such traces easily in the frequency domain, *e.g.*, (2.5.40). For non-circulant imaging problems, exact trace computation can be prohibitively expensive. However, the following *stochastic* (**Monte Carlo**) approach to estimating the trace of a matrix $\boldsymbol{M}$ is simple and effective [92–97].

Let $\boldsymbol{w}$ be an IID random vector in $\mathbb{R}^{n_{\mathrm{d}}}$ with $\mathsf{E}[\boldsymbol{w}] = \mathbf{0}$ and $\mathsf{Cov}\{\boldsymbol{w}\} = \boldsymbol{I}_{n_{\mathrm{d}}}$. Then by (29.5.1):

$$\mathsf{E}[\boldsymbol{w}'\boldsymbol{M}\boldsymbol{w}] = \mathsf{E}[\mathsf{trace}\{\boldsymbol{w}'\boldsymbol{M}\boldsymbol{w}\}] = \mathsf{trace}\{\boldsymbol{M}\,\mathsf{E}[\boldsymbol{w}\boldsymbol{w}']\} = \mathsf{trace}\{\boldsymbol{M}\}. \tag{2.5.42}$$

e,reg,hyper,trace

Thus $\hat{t} \triangleq \boldsymbol{w}'\boldsymbol{M}\boldsymbol{w}$ is an unbiased estimate of $t = \mathsf{trace}\{\boldsymbol{M}\}$. To reduce the variance of $\hat{t}$, one could average several realizations. However, in imaging problems usually $n_{\mathrm{d}}$ is large enough that $\hat{t}$ has small variance. Using an IID Bernoulli $\pm 1$ distribution for $\boldsymbol{w}$ is preferable [92, 98, 99].

s,reg,hyper,ngcv

### 2.5.3.3 GCV for nonlinear estimators (s,reg,hyper,ngcv)

Various methods have been proposed for extending GCV for nonlinear estimators [93, 99–103]. Here explore one heuristic method based on (2.5.42).

Consider a linear estimator $\hat{\boldsymbol{x}}_\beta(\boldsymbol{y}) = \boldsymbol{L}_\beta \boldsymbol{y}$, linear model $\bar{\boldsymbol{y}}(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x}$, and white noise $\mathsf{Cov}\{\boldsymbol{y}\} = \sigma^2 \boldsymbol{I}$, with corresponding influence matrix $\mathsf{M}(\beta) = \boldsymbol{A}\boldsymbol{L}_\beta$. If $\mathsf{E}[\boldsymbol{w}] = \boldsymbol{0}$ and $\mathsf{Cov}\{\boldsymbol{w}\} = \boldsymbol{I}$, then using (2.5.42), an unbiased estimate of the trace of $\mathsf{M}(\beta)$ is $\boldsymbol{w}' \mathsf{M}(\beta) \boldsymbol{w}$. We exploit linearity to rewrite this unbiased trace estimate as follows:

$$\boldsymbol{w}' \mathsf{M}(\beta) \boldsymbol{w} = \boldsymbol{w}' \boldsymbol{A} \boldsymbol{L}_\beta \boldsymbol{w} = \boldsymbol{w}' \bar{\boldsymbol{y}}(\hat{\boldsymbol{x}}_\beta(\boldsymbol{w})) = \boldsymbol{w}' \frac{\bar{\boldsymbol{y}}(\hat{\boldsymbol{x}}_\beta(\boldsymbol{y} + \varepsilon \boldsymbol{w})) - \bar{\boldsymbol{y}}(\hat{\boldsymbol{x}}_\beta(\boldsymbol{y}))}{\varepsilon} \triangleq \hat{\mathsf{M}}_\beta(\boldsymbol{w}, \varepsilon) . \tag{2.5.43}$$

e,reg,hyper,gcv,tr,w

For linear estimators, $\mathsf{E}\left[\hat{\mathsf{M}}_\beta(\boldsymbol{w}, \varepsilon)\right] = \mathsf{E}[\boldsymbol{w}' \mathsf{M}(\beta) \boldsymbol{w}] = \mathsf{trace}\{\mathsf{M}(\beta)\}$ for any $\varepsilon \neq 0$.

For nonlinear estimators, we can form a heuristic version of GCV by replacing (2.5.38) with

$$\boxed{\beta_{\text{NGCV}} \triangleq \arg\min_\beta \Phi_{\text{NGCV}}(\beta),} \quad \boxed{\Phi_{\text{NGCV}}(\beta) \triangleq \frac{\mathsf{RSS}(\hat{\boldsymbol{x}}_\beta)}{\left(1 - \frac{1}{n_{\text{d}}} \hat{\mathsf{M}}_\beta(\boldsymbol{w}, \varepsilon)\right)^2}.} \tag{2.5.44}$$

e,reg,hyper,ngcv

This approach requires applying the estimator twice for each candidate $\beta$: once for data $\boldsymbol{y}$ and once for the perturbed data $\boldsymbol{y} + \varepsilon \boldsymbol{w}$. Choosing $\varepsilon$ such that $\|\varepsilon \boldsymbol{w}\| \ll \|\boldsymbol{y}\|$ seems desirable so that the estimator behaves approximately linearly. (However, if $\varepsilon$ is too small, there can be numerical precision issues in evaluating (2.5.43) numerically.) An even simpler approach is to use $\hat{\mathsf{M}}_\beta \triangleq \boldsymbol{w}' \boldsymbol{A} \hat{\boldsymbol{x}}_\beta(\boldsymbol{w})$, which is unbiased in the linear case and may work acceptably even for some nonlinear problems [103].

An alternative way of deriving $\hat{\mathsf{M}}_\beta(\boldsymbol{w}, \varepsilon)$ in (2.5.43) is as follows. When both $\bar{\boldsymbol{y}}(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x}$ and $\hat{\boldsymbol{x}}_\beta(\boldsymbol{y}) = \boldsymbol{L}_\beta \boldsymbol{y}$ are linear, then the influence matrix represents the differential change in the predicted measurements $\bar{\boldsymbol{y}}$ as a function of the measured values $\boldsymbol{y}$:

$$\mathsf{M}(\beta) = \boldsymbol{A}\boldsymbol{L}_\beta = \nabla_{\boldsymbol{y}} \bar{\boldsymbol{y}}(\hat{\boldsymbol{x}}_\beta(\boldsymbol{y})) .$$

To generalize the notion of influence matrix to for nonlinear models and/or estimators, we can *define* the **influence matrix** as this gradient:

$$\mathsf{M}(\beta) \triangleq \nabla_{\boldsymbol{y}} \bar{\boldsymbol{y}}(\hat{\boldsymbol{x}}_\beta(\boldsymbol{y})) . \tag{2.5.45}$$

e,reg,hyper,ngcv,M

Defining $\boldsymbol{\mu}_\beta(\boldsymbol{y}) \triangleq \bar{\boldsymbol{y}}(\hat{\boldsymbol{x}}_\beta(\boldsymbol{y}))$ as a mapping from $\mathbb{R}^{n_{\text{d}}}$ into $\mathbb{R}^{n_{\text{d}}}$, then

$$\mathsf{trace}\{\mathsf{M}(\beta)\} = \mathsf{trace}\{\nabla_{\boldsymbol{y}} \boldsymbol{\mu}_\beta(\boldsymbol{y})\} = \sum_{i=1}^{n_{\text{d}}} \frac{\partial}{\partial y_i} [\boldsymbol{\mu}_\beta(\boldsymbol{y})]_i,$$

where the last sum is called the **divergence** of $\boldsymbol{\mu}_\beta(\boldsymbol{y})$ [97].

Using a first-order Taylor expansion for a small $\varepsilon$:

$$\boldsymbol{\mu}_\beta(\boldsymbol{y} + \varepsilon \boldsymbol{w}) \approx \boldsymbol{\mu}_\beta(\boldsymbol{y}) + \nabla \boldsymbol{\mu}_\beta(\boldsymbol{y})(\varepsilon \boldsymbol{w})$$

so

$$\boldsymbol{w}' \frac{\boldsymbol{\mu}_\beta(\boldsymbol{y} + \varepsilon \boldsymbol{w}) - \boldsymbol{\mu}_\beta(\boldsymbol{y})}{\varepsilon} \approx \boldsymbol{w}' \nabla \boldsymbol{\mu}_\beta(\boldsymbol{y}) \boldsymbol{w} \approx \mathsf{trace}\{\mathsf{M}(\beta)\} .$$

In summary, a reasonable approximation for the trace of the influence matrix for use in (2.5.44) is

$$\mathsf{trace}\{\mathsf{M}(\beta)\} \approx \boldsymbol{w}' \frac{\bar{\boldsymbol{y}}(\hat{\boldsymbol{x}}_\beta(\boldsymbol{y} + \varepsilon \boldsymbol{w})) - \bar{\boldsymbol{y}}(\hat{\boldsymbol{x}}_\beta(\boldsymbol{y}))}{\varepsilon}.$$

This approximation should become unbiased as $\varepsilon \to 0$ for nonlinear estimators that are continuously differentiable.

MIRT *See* `ir_deblur_gcv1.m`.

s,reg,hyper,ml

### 2.5.4 Maximum likelihood and Bayesian methods (s,reg,hyper,ml)

A regularization method with penalty function $\beta \mathsf{R}_\delta(\boldsymbol{x})$ can be interpreted as a Bayesian method with prior distribution $\mathsf{p}(\boldsymbol{x}; \beta, \delta) = c_{\beta,\delta} \, \mathrm{e}^{-\beta \mathsf{R}_\delta(\boldsymbol{x})}$, where $c_{\beta,\delta}$ is a constant known as the **partition function** in the **Markov random field** literature that ensures the density function integrates to unity. Given a noiseless training image $\boldsymbol{x}$, in principle one could estimate the parameters $\beta$ and $\delta$ by maximum likelihood:

$$\hat{\beta}, \hat{\delta} = \arg\max_{\beta,\delta} \log \mathsf{p}(\boldsymbol{x}; \beta, \delta),$$

*e.g.*, [104]. Alternatively, if one supposes a prior for the regularization parameters, then one can estimate them from a training image by a Bayesian MAP approach:

$$\hat{\beta}, \hat{\delta} = \arg\max_{\beta,\delta} \log \mathsf{p}(\beta, \delta \,|\, \boldsymbol{x}) .$$

Many such Bayesian methods have been investigated [67, 68, 105, 106]. In practice, these methods are difficult to realize because of the complexity of the partition function. Approximations that disregard the dependence of the rows of $\boldsymbol{C}\boldsymbol{x}$ have been investigated, *e.g.*, [107].

s,reg,hyper,lcurve

### 2.5.5 L-curve method (s,reg,hyper,lcurve)

Regularized methods involve minimizing a cost function consisting of a data fit term and a regularization term:

$$\hat{\boldsymbol{x}}_\beta = \arg\min_{\boldsymbol{x}} \mathsf{L}(\boldsymbol{x}) + \beta\, \mathsf{R}(\boldsymbol{x}) \,.$$

The values of $\mathsf{L}(\hat{\boldsymbol{x}}_\beta)$ and $\mathsf{R}(\hat{\boldsymbol{x}}_\beta)$ change as one varies $\beta$. If one graphs $(\mathsf{L}(\hat{\boldsymbol{x}}_\beta), \mathsf{R}(\hat{\boldsymbol{x}}_\beta))$ as $\beta$ is varied, the curve has an "L" shape [108]. It has been argued that reasonable values for $\beta$ lie somewhere near the "corner" in this **L-curve** [109–113]. However, there also have been critiques of this method [114]. It requires substantial computation in general to trace out the L-curve, because one must find $\hat{\boldsymbol{x}}_\beta$ for several values of $\beta$. The location of the "corner" of the L-curve does not have a canonical definition. And the properties of $\hat{\boldsymbol{x}}_\beta$ in terms of spatial resolution, noise, or MSE are unknown when $\beta$ is chosen using the L-curve method. So we do not consider this approach further here.

s,reg,hyper,sure

### 2.5.6 SURE methods (s,reg,hyper,sure)

The MSE in (2.5.2) depends on the true parameter $\boldsymbol{x}_{\text{true}}$ so it cannot be used in practice for choosing $\beta$. However, one can *estimate* the MSE (also known as the **risk**) as a function of $\beta$ and then minimize the estimate. The best known such method is **Stein's unbiased risk estimate** (**SURE**) [74, 97, 115–124].

#### 2.5.6.1 Weighted MSE

In this section we consider a **weighted mean-squared error** (**WMSE**) that generalizes (2.5.6) as follows:

e,reg,hyper,sure,wmse

$$\mathsf{WMSE}_\beta \triangleq \mathsf{E}[(\hat{\boldsymbol{x}}_\beta - \bar{\boldsymbol{x}}_\beta)' \boldsymbol{J}_1 (\hat{\boldsymbol{x}}_\beta - \bar{\boldsymbol{x}}_\beta)] + (\bar{\boldsymbol{x}}_\beta - \boldsymbol{x}_{\text{true}})' \boldsymbol{J}_2 (\bar{\boldsymbol{x}}_\beta - \boldsymbol{x}_{\text{true}}) \,, \tag{2.5.46}$$

where the estimator mean $\bar{\boldsymbol{x}}_\beta$ was defined in (2.5.3). The first term quantifies the variability (noise) of the estimator $\hat{\boldsymbol{x}}_\beta$, and the second term quantifies the systematic error (bias) of the estimator. If $\boldsymbol{J}_1 = \boldsymbol{J}_2 = \boldsymbol{I}$ then this WMSE simplifies to the standard definition of MSE in (2.5.2).

Expanding (2.5.46) and simplifying yields

e,reg,hyper,sure,wmse,all4

$$\mathsf{WMSE}_\beta = \mathsf{E}\big[\hat{\boldsymbol{x}}_\beta' \boldsymbol{J}_1 \hat{\boldsymbol{x}}_\beta\big] + \bar{\boldsymbol{x}}_\beta' (\boldsymbol{J}_2 - \boldsymbol{J}_1) \bar{\boldsymbol{x}}_\beta - 2 \operatorname{real}\big\{\bar{\boldsymbol{x}}_\beta' \boldsymbol{J}_2 \boldsymbol{x}_{\text{true}}\big\} + c_2, \tag{2.5.47}$$

where $c_2 = \boldsymbol{x}_{\text{true}}' \boldsymbol{J}_2 \boldsymbol{x}_{\text{true}}$ is a constant independent of $\beta$ that can be ignored. The middle two terms are the primary challenge for choosing $\beta$.

#### 2.5.6.2 Linear model and estimator

Consider the case of a linear estimator $\hat{\boldsymbol{x}}_\beta = \boldsymbol{L}_\beta \boldsymbol{y}$ for which $\bar{\boldsymbol{x}}_\beta \triangleq \mathsf{E}[\hat{\boldsymbol{x}}_\beta] = \boldsymbol{L}_\beta \boldsymbol{A} \boldsymbol{x}_{\text{true}}$, assuming $\mathsf{E}[\boldsymbol{y}] = \boldsymbol{A}\boldsymbol{x}_{\text{true}}$. In this case the middle two terms of $\mathsf{WMSE}_\beta$ in (2.5.47) become

e,reg,hyper,sure,wmse,23

$$\boldsymbol{x}_{\text{true}}' \boldsymbol{A}' \boldsymbol{M}_1 \boldsymbol{A} \boldsymbol{x}_{\text{true}} - 2 \operatorname{real}\big\{\boldsymbol{x}_{\text{true}}' \boldsymbol{A}' \boldsymbol{L}_\beta' \boldsymbol{J}_2 \boldsymbol{x}_{\text{true}}\big\}, \tag{2.5.48}$$

where $\boldsymbol{M}_1 \triangleq \boldsymbol{L}_\beta' (\boldsymbol{J}_2 - \boldsymbol{J}_1) \boldsymbol{L}_\beta$. To proceed, we assume $\boldsymbol{y} \sim \mathsf{N}\big(\boldsymbol{A}\boldsymbol{x}_{\text{true}}, \boldsymbol{W}^{-1}\big)$ and use (29.5.1) for a general $n_{\text{d}} \times n_{\text{d}}$ matrix $\boldsymbol{M}$:

$$\mathsf{E}[\boldsymbol{y}' \boldsymbol{M} \boldsymbol{y}] = \bar{\boldsymbol{y}}' \boldsymbol{M} \bar{\boldsymbol{y}} + \mathsf{trace}\{\mathsf{Cov}\{\boldsymbol{y}\} \boldsymbol{M}\} = \boldsymbol{x}_{\text{true}}' \boldsymbol{A}' \boldsymbol{M} \boldsymbol{A}' \boldsymbol{x}_{\text{true}} + \mathsf{trace}\big\{\boldsymbol{W}^{-1} \boldsymbol{M}\big\} \,.$$

Thus $\boldsymbol{y}' \boldsymbol{M}_1 \boldsymbol{y} - \mathsf{trace}\big\{\boldsymbol{W}^{-1} \boldsymbol{M}_1\big\} = \hat{\boldsymbol{x}}_\beta' (\boldsymbol{J}_2 - \boldsymbol{J}_1) \hat{\boldsymbol{x}}_\beta - \mathsf{trace}\big\{\boldsymbol{W}^{-1} \boldsymbol{M}_1\big\}$ is an unbiased estimate of the first term in (2.5.48).

The second term in (2.5.48) is more challenging. We describe two approaches for estimating it next.

##### 2.5.6.2.1 Case where an unbiased estimator exists

One approach is to assume there exists an *unbiased* estimator $\hat{\boldsymbol{x}}_0$ of $\boldsymbol{x}_{\text{true}}$, with some covariance $\boldsymbol{K}_0$. Then using (29.5.1) again, $\hat{\boldsymbol{x}}_0' \boldsymbol{M}_2 \hat{\boldsymbol{x}}_0 - \mathsf{trace}\{\boldsymbol{K}_0 \boldsymbol{M}_2\}$ is an unbiased estimator of $\boldsymbol{x}_{\text{true}}' \boldsymbol{M}_2 \boldsymbol{x}_{\text{true}}$ where $\boldsymbol{M}_2 \triangleq \boldsymbol{A}' \boldsymbol{L}_\beta' \boldsymbol{J}_2$ is a $n_{\text{p}} \times n_{\text{p}}$ matrix. Collecting terms leads to the following unbiased estimate of the WMSE:

e,reg,hyper,sure,ure1

$$\Phi_{\text{URE},1}(\beta) = \hat{\boldsymbol{x}}_\beta' \boldsymbol{J}_2 \hat{\boldsymbol{x}}_\beta - \mathsf{trace}\big\{\boldsymbol{W}^{-1} \boldsymbol{M}_1\big\} - 2 (\hat{\boldsymbol{x}}_0' \boldsymbol{M}_2 \hat{\boldsymbol{x}}_0 - \mathsf{trace}\{\boldsymbol{K}_0 \boldsymbol{M}_2\}) + c_2. \tag{2.5.49}$$

To further simplify, suppose $\boldsymbol{W} = \sigma^{-2} \boldsymbol{I}$, $\boldsymbol{L}_\beta = \big[\boldsymbol{\mathsf{F}} + \beta \boldsymbol{\mathsf{R}}\big]^{-1} \boldsymbol{A}' \boldsymbol{W} = \big[\boldsymbol{A}'\boldsymbol{A} + \sigma^2 \beta \boldsymbol{\mathsf{R}}\big]^{-1} \boldsymbol{A}'$, $\boldsymbol{J}_1 = \boldsymbol{I}$, and $\boldsymbol{J}_2 = \alpha \boldsymbol{I}$, for which $\boldsymbol{M}_1 = (\alpha - 1) \boldsymbol{A} \big[\boldsymbol{A}'\boldsymbol{A} + \sigma^2 \beta \boldsymbol{\mathsf{R}}\big]^{-2} \boldsymbol{A}'$ and $\boldsymbol{M}_2 = \alpha \boldsymbol{A}' \boldsymbol{A} \big[\boldsymbol{A}'\boldsymbol{A} + \sigma^2 \beta \boldsymbol{\mathsf{R}}\big]^{-1}$. Furthermore, $\hat{\boldsymbol{x}}_0 = [\boldsymbol{A}'\boldsymbol{A}]^{-1} \boldsymbol{A}' \boldsymbol{y}$ and $\boldsymbol{K}_0 = \sigma^2 [\boldsymbol{A}'\boldsymbol{A}]^{-1}$. Then

$$\Phi_{\text{URE},1}(\beta) = \alpha \boldsymbol{y}' \boldsymbol{A} \big[\boldsymbol{A}'\boldsymbol{A} + \sigma^2 \beta \boldsymbol{\mathsf{R}}\big]^{-2} \boldsymbol{A}' \boldsymbol{y} - \sigma^2 (\alpha - 1) \operatorname{trace}\Big\{ \big[\boldsymbol{A}'\boldsymbol{A} + \sigma^2 \beta \boldsymbol{\mathsf{R}}\big]^{-2} \boldsymbol{A}'\boldsymbol{A} \Big\}$$

e,reg,hyper,sure,urel,qpwls

$$- 2\alpha \boldsymbol{y}' \boldsymbol{A} \left[\boldsymbol{A}'\boldsymbol{A} + \sigma^2 \beta \mathbf{R}\right]^{-1} \left[\boldsymbol{A}'\boldsymbol{A}\right]^{-1} \boldsymbol{A}'\boldsymbol{y} + 2\sigma^2 \alpha \,\mathsf{trace}\left\{\left[\boldsymbol{A}'\boldsymbol{A} + \sigma^2 \beta \mathbf{R}\right]^{-1}\right\} + c_2. \tag{2.5.50}$$

Assuming $\boldsymbol{A}$ and $\mathbf{R}$ are both circulant, with corresponding eigenvalues $B_k$ and $\mathsf{R}_k$, then

e,reg,hyper,sure,urel,circ

$$\Phi_{\mathrm{URE},1}(\beta) = \frac{1}{N} \sum_k \left( \frac{|B_k Y[k]|^2}{\left(|B_k|^2 + \sigma^2 \beta \mathsf{R}_k\right)^2} + (\alpha - 1) \frac{|B_k|^2 \left(|Y[k]|^2 - N\sigma^2\right)}{\left(|B_k|^2 + \sigma^2 \beta \mathsf{R}_k\right)^2} - 2\alpha \frac{|Y[k]|^2 - N\sigma^2}{|B_k|^2 + \sigma^2 \beta \mathsf{R}_k} \right) + c_2, \tag{2.5.51}$$

where $Y[k]$ denotes the DFT of $\boldsymbol{y}$. The case $\alpha = 1$ simplifies to the usual MSE, for which (2.5.51) reduces to [66, eqn. (21)-(22)].

x,fig_reg_hyper_sure1

**Example 2.5.12** *See Fig. 2.5.2 for an illustration of choosing* $\beta$ *by minimizing* $\Phi_{\mathrm{URE},1}(\beta)$ *in (2.5.51), using a quadratic roughness penalty with periodic boundary conditions.*

*MIRT* See `fig_reg_hyper_sure1.m`.

#### 2.5.6.2.2 Case where certain matrices commute (e.g., for denoising)

Unbiased estimators $\hat{\boldsymbol{x}}_0$ do not exist when $\boldsymbol{A}$ has a non-trivial null space. An alternative approach is to make the following restrictive assumption:

$$\boldsymbol{J}_2 \boldsymbol{L}_\beta = \boldsymbol{A}' \boldsymbol{M}_3,$$

for some $n_\mathrm{d} \times n_\mathrm{d}$ matrix $\boldsymbol{M}_3$, so that $\boldsymbol{A}'\boldsymbol{L}_\beta'\boldsymbol{J}_2 = \boldsymbol{A}'\boldsymbol{M}_3\boldsymbol{A}$. (This holds for certain circulant problems, even when $\boldsymbol{A}$ is singular, and for denoising problems where $\boldsymbol{A} = \boldsymbol{I}$, but perhaps not much more generally.) Then $-2\left(\boldsymbol{y}'\boldsymbol{M}_3\boldsymbol{y} - \mathsf{trace}\{\boldsymbol{W}^{-1}\boldsymbol{M}_3\}\right)$ is an unbiased estimate of the second term in (2.5.48).

Collecting terms leads to the following unbiased estimate of the WMSE:

e,reg,hyper,sure,ure2

$$\Phi_{\mathrm{URE},2}(\beta) = \hat{\boldsymbol{x}}_\beta' \boldsymbol{J}_2 \hat{\boldsymbol{x}}_\beta - \mathsf{trace}\{\boldsymbol{W}^{-1}\boldsymbol{M}_1\} - 2\left(\boldsymbol{y}'\boldsymbol{M}_3\boldsymbol{y} - \mathsf{trace}\{\boldsymbol{W}^{-1}\boldsymbol{M}_3\}\right) + c_2. \tag{2.5.52}$$

Further simplifying, suppose $\boldsymbol{W} = \sigma^{-2}\boldsymbol{I}$, $\boldsymbol{L}_\beta = \left[\mathbf{F} + \beta \boldsymbol{R}\right]^{-1} \boldsymbol{A}'\boldsymbol{W} = \left[\boldsymbol{A}'\boldsymbol{A} + \sigma^2\beta\mathbf{R}\right]^{-1}\boldsymbol{A}'$, $\boldsymbol{J}_1 = \boldsymbol{I}$, and $\boldsymbol{J}_2 = \alpha\boldsymbol{I}$, for which $\boldsymbol{M}_1 = (\alpha - 1)\boldsymbol{A}\left[\boldsymbol{A}'\boldsymbol{A} + \sigma^2\beta\mathbf{R}\right]^{-2}\boldsymbol{A}'$.

Assuming $\boldsymbol{M}_3$ and $\boldsymbol{A}'$ commute (*e.g.*, they are both circulant, or when $\boldsymbol{A} = \boldsymbol{I}$) then we also have $\boldsymbol{M}_3 = \alpha\left[\boldsymbol{A}'\boldsymbol{A} + \sigma^2\beta\mathbf{R}\right]^{-1}$. This leads an expression similar (but not identical!) to (2.5.50):

$$\begin{aligned}\Phi_{\mathrm{URE},2}(\beta) = {} & \alpha \boldsymbol{y}'\boldsymbol{A}\left[\boldsymbol{A}'\boldsymbol{A} + \sigma^2\beta\mathbf{R}\right]^{-2}\boldsymbol{A}'\boldsymbol{y} - \sigma^2(\alpha-1)\,\mathsf{trace}\left\{\left[\boldsymbol{A}'\boldsymbol{A} + \sigma^2\beta\mathbf{R}\right]^{-2}\boldsymbol{A}'\boldsymbol{A}\right\} \\ & - 2\alpha\boldsymbol{y}'\left[\boldsymbol{A}'\boldsymbol{A} + \sigma^2\beta\mathbf{R}\right]^{-1}\boldsymbol{y} + 2\sigma^2\alpha\,\mathsf{trace}\left\{\left[\boldsymbol{A}'\boldsymbol{A} + \sigma^2\beta\mathbf{R}\right]^{-1}\right\} + c_2.\end{aligned}$$

Interestingly, assuming $\boldsymbol{A}$ and $\mathbf{R}$ are both circulant leads to an expression for $\Phi_{\mathrm{URE},2}(\beta)$ that is *identical* to (2.5.51). In other words, for circulant problems, $\Phi_{\mathrm{URE},1}$ in (2.5.51) is valid even when no unbiased estimator $\hat{\boldsymbol{x}}_0$ exists.

For the denoising case where $\boldsymbol{A} = \boldsymbol{I}$, and for the usual case where $\alpha = 1$, this simplifies to

$$\Phi_{\mathrm{URE},2}(\beta) = \boldsymbol{y}'\left[\boldsymbol{I} + \sigma^2\beta\mathbf{R}\right]^{-2}\boldsymbol{y} - 2\boldsymbol{y}'\left[\boldsymbol{I} + \sigma^2\beta\mathbf{R}\right]^{-1}\boldsymbol{y} + 2\sigma^2\,\mathsf{trace}\left\{\left[\boldsymbol{I} + \sigma^2\beta\mathbf{R}\right]^{-1}\right\} + c_2,$$

which one can show is equivalent to [97, eqn. (6)].

Generalizing this WMSE approach to non-circulant problems, even for a linear model and linear estimators, is an *open problem.*

x,reg,hyper,sure,mc

**Example 2.5.13** *This example applies the Monte Carlo trace estimate of §2.5.3.2 to a denoising problem with* $\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{\varepsilon}$ *where* $\boldsymbol{\varepsilon} \sim \mathsf{N}(\mathbf{0}, \sigma^2\boldsymbol{I})$ *and a linear estimator* $\hat{\boldsymbol{x}}_\beta = \boldsymbol{L}_\beta\boldsymbol{y}$. *One can verify that* $\Phi_{\mathrm{SURE}}(\beta) \triangleq \hat{\boldsymbol{x}}_\beta'\hat{\boldsymbol{x}}_\beta - 2(\boldsymbol{y}'\boldsymbol{L}_\beta\boldsymbol{y} - \sigma^2\,\mathsf{trace}\{\boldsymbol{L}_\beta\}) + c_2$ *is an unbiased estimate of* $\mathrm{MSE}(\beta) = \mathsf{E}\left[\|\hat{\boldsymbol{x}}_\beta - \boldsymbol{x}_\mathrm{true}\|^2\right]$. *Furthermore the following is also an unbiased estimator when* $\boldsymbol{w}$ *has an IID Bernoulli* $\pm 1$ *distribution:*

$$\Phi_{\mathrm{SURE}}(\beta) \triangleq \hat{\boldsymbol{x}}_\beta'\hat{\boldsymbol{x}}_\beta - 2(\boldsymbol{y}'\hat{\boldsymbol{x}}_\beta - \sigma^2\boldsymbol{w}'\boldsymbol{L}_\beta\boldsymbol{w}) + c_2.$$

### 2.5.6.3 Nonlinear estimators

The MSE of an estimator $\hat{\boldsymbol{x}}_\beta$ can be expanded:

reg,hyper,sure,mse

$$\mathrm{MSE}_\beta = \mathsf{E}\left[\|\hat{\boldsymbol{x}}_\beta - \boldsymbol{x}_\mathrm{true}\|^2\right] = \mathsf{E}\left[\|\hat{\boldsymbol{x}}_\beta\|^2\right] - 2\,\mathsf{E}\left[\hat{\boldsymbol{x}}_\beta'\,\boldsymbol{x}_\mathrm{true}\right] + \|\boldsymbol{x}_\mathrm{true}\|^2. \tag{2.5.53}$$

The middle term is the challenging one because it depends both on the estimator $\hat{\boldsymbol{x}}_\beta$ and the unknown parameter $\boldsymbol{x}_\mathrm{true}$.

To proceed we need the following property of the gaussian distribution [115] [125] [118, p. 396]. This result generalizes to exponential families [119, eqn. (15)]. See also see (29.4.2). Generalizing the methods below to a WMSE of the form (2.5.46) is an *open problem.*

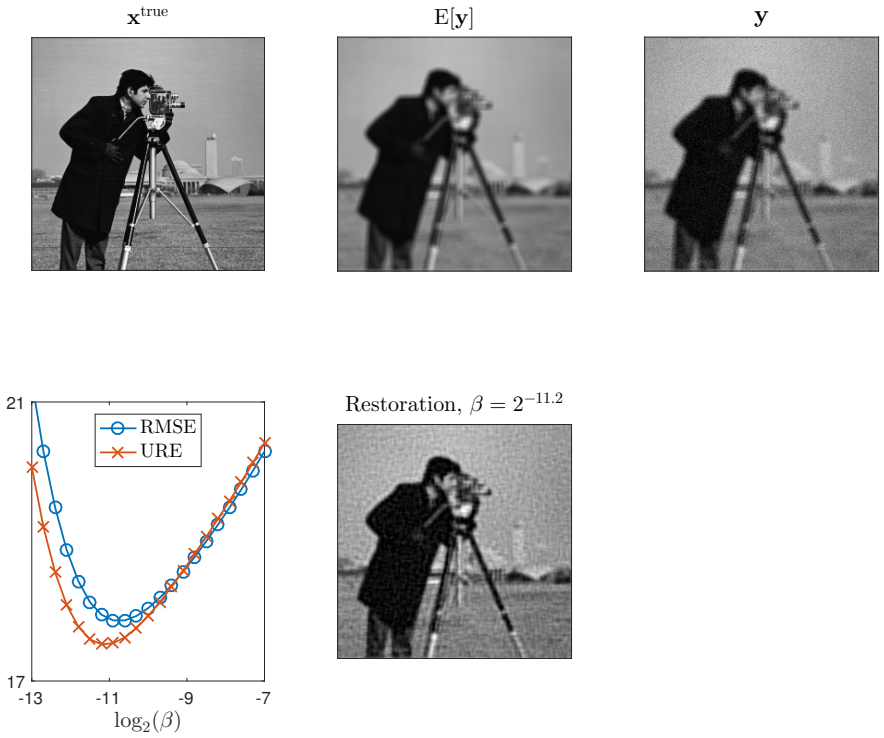



Figure 2.5.2: Example of using unbiased risk estimator (2.5.51) to choose regularization parameter β for an image restoration problem. fig_reg_hyper_sure1

l,prob,gauss,sure1

**Lemma 2.5.14** *If* $\boldsymbol{z} \sim \mathsf{N}(\boldsymbol{\mu}, \boldsymbol{K}) \in \mathbb{R}^{n_\mathrm{p}}$ *and each* $h_j(\boldsymbol{z})$ *is a differentiable function of* $\boldsymbol{z}$ *for which* $\mathsf{E}[|h_j(\boldsymbol{z})|]$ *is bounded for* $j = 1, \dots, n_\mathrm{p}$*, then*

$$\mathsf{E}[\boldsymbol{h}'(\boldsymbol{z})\boldsymbol{\mu}] = \mathsf{E}[\boldsymbol{h}'(\boldsymbol{z})\boldsymbol{z}] - \mathsf{trace}\{\mathsf{E}[\boldsymbol{K}\nabla\boldsymbol{h}(\boldsymbol{z})]\}\,. \tag{2.5.54}$$

e,prob,gauss,sure1

Proof:
Differentiating (29.4.1):

$$\nabla_{\boldsymbol{z}}\,\mathsf{p}(\boldsymbol{z}) = -\,\mathsf{p}(\boldsymbol{z})\,\boldsymbol{K}^{-1}(\boldsymbol{z}-\boldsymbol{\mu}) \implies \boldsymbol{K}\nabla_{\boldsymbol{z}}\,\mathsf{p}(\boldsymbol{z}) = \mathsf{p}(\boldsymbol{z})\,\boldsymbol{\mu} - \mathsf{p}(\boldsymbol{z})\,\boldsymbol{z}.$$

Multiplying by $\boldsymbol{h}'$ and taking the expectation:

$$\mathsf{E}[\boldsymbol{h}'(\boldsymbol{z})\boldsymbol{\mu}] = \mathsf{E}[\boldsymbol{h}'(\boldsymbol{z})\boldsymbol{z}] + \int \boldsymbol{h}'(\boldsymbol{z})\boldsymbol{K}\nabla\,\mathsf{p}(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z}\,.$$

Letting $\boldsymbol{v}(\boldsymbol{z}) = \boldsymbol{K}\boldsymbol{h}(\boldsymbol{z})$:

$$\begin{aligned}\int \boldsymbol{h}'(\boldsymbol{z})\boldsymbol{K}\nabla\,\mathsf{p}(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z} &= \int \boldsymbol{v}'(\boldsymbol{z})\nabla\,\mathsf{p}(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z} = \sum_j \int v_j(\boldsymbol{z})\frac{\partial}{\partial z_j}\,\mathsf{p}(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z} \\ &= -\sum_j \int \left(\frac{\partial}{\partial z_j} v_j(\boldsymbol{z})\right)\mathsf{p}(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z} = -\,\mathsf{trace}\{\mathsf{E}[\nabla\,\boldsymbol{v}(\boldsymbol{z})]\} = -\,\mathsf{trace}\{\mathsf{E}[\boldsymbol{K}\nabla\boldsymbol{h}(\boldsymbol{z})]\},\end{aligned}$$

using integration by parts and the assumptions on $\boldsymbol{h}$. □

The utility of (2.5.54) is that the right-hand side terms do not depend on $\boldsymbol{\mu}$, unlike the left-hand side.

Now suppose that $\hat{\boldsymbol{x}}_0$ is an **unbiased** estimator for $\boldsymbol{x}_{\text{true}}$ with covariance $\boldsymbol{K}$, and suppose that $\hat{\boldsymbol{x}}_\beta = \boldsymbol{L}_\beta\,\hat{\boldsymbol{x}}_0$ is a linear function of $\hat{\boldsymbol{x}}_0$. Then applying (2.5.54) with $\boldsymbol{\mu} \mapsto \boldsymbol{x}_{\text{true}}$ and $\boldsymbol{z} \mapsto \hat{\boldsymbol{x}}_0$ yields:

$$\mathsf{E}\big[\hat{\boldsymbol{x}}_\beta'\,\boldsymbol{x}_{\text{true}}\big] = \mathsf{E}\big[\hat{\boldsymbol{x}}_\beta'\,\hat{\boldsymbol{x}}_0\big] - \mathsf{trace}\big\{\boldsymbol{K}\boldsymbol{L}_\beta'\big\}\,.$$

Therefore the following criterion is a **unbiased estimate** of the MSE:

$$\begin{aligned}\Phi_{\text{SURE}}(\beta) &\triangleq \|\hat{\boldsymbol{x}}_\beta\|^2 - 2\,\hat{\boldsymbol{x}}_\beta'\,\hat{\boldsymbol{x}}_0 + 2\,\mathsf{trace}\big\{\boldsymbol{K}\boldsymbol{L}_\beta'\big\} + \|\boldsymbol{x}_{\text{true}}\|^2 \\ &= \|\hat{\boldsymbol{x}}_\beta - \hat{\boldsymbol{x}}_0\|^2 + 2\,\mathsf{trace}\big\{\boldsymbol{K}\boldsymbol{L}_\beta'\big\} + \left(\|\boldsymbol{x}_{\text{true}}\|^2 - \|\hat{\boldsymbol{x}}_0\|^2\right),\end{aligned} \tag{2.5.55}$$

e,reg,hyper,kost,sure

*i.e.*, we select β as follows

$$\boxed{\beta_{\text{SURE}} \triangleq \arg\min_\beta \Phi_{\text{SURE}}(\beta),} \quad \text{where} \quad \boxed{\mathsf{E}[\Phi_{\text{SURE}}(\beta)] = \text{MSE}_\beta.}$$

(The final term is a constant independent of $\beta$ so it does not affect parameter selection.)

For example, when $\hat{\boldsymbol{x}}_0 = \mathsf{F}^{-1}\boldsymbol{A}'\boldsymbol{W}\boldsymbol{y}$ and $\boldsymbol{W} = [\mathsf{Cov}\{\boldsymbol{y}\}]^{-1}$ so $\mathsf{Cov}\{\hat{\boldsymbol{x}}_0\} = \mathsf{F}^{-1}$ and $\boldsymbol{L}_\beta = [\mathsf{F} + \beta\mathsf{R}]^{-1}\mathsf{F}$ we have (*cf.* [118, eqn. (5.51)]):

$$\Phi_{\mathrm{SURE}}(\beta) \equiv \left\| [\mathsf{F} + \beta\mathsf{R}]^{-1} \beta\mathsf{R}\mathsf{F}^{-1}\boldsymbol{A}'\boldsymbol{W}\boldsymbol{y} \right\|^2 + 2\,\mathsf{trace}\left\{ [\mathsf{F} + \beta\mathsf{R}]^{-1} \right\} + c_0, \tag{2.5.56}$$

e,reg,hyper,kost,sure,pwls

where $c_0 \triangleq \|\boldsymbol{x}_{\mathrm{true}}\|^2 - \|\hat{\boldsymbol{x}}_0\|^2$ is independent of $\beta$.

The criterion $\Phi_{\mathrm{SURE}}$ in (2.5.56) requires that $\mathsf{F}$ be non-singular, which limits its applicability in image reconstruction problems. See [103, 119] consideration of cases where $\mathsf{F}$ is singular, using a modified MSE of the form $\mathsf{E}\left[ \|\mathcal{P}_{\boldsymbol{A}}(\hat{\boldsymbol{x}}_\beta - \boldsymbol{x}_{\mathrm{true}})\|^2 \right]$, where $\mathcal{P}_{\boldsymbol{A}}$ denotes the orthogonal **projection** onto the range space of $\boldsymbol{A}$. The approach in [103, 119] requires computing the **pseudo-inverse** solution which is impractical except for special cases like circulant problems. A general practical solution for the singular case remains an *open problem*.

Monte Carlo methods are another approach to computing unbiased estimates of $\mathrm{MSE}_\beta$ [97, 121, 123].

### 2.5.7 Other regularization parameter selection methods (s,reg,hyper,other)

s,reg,hyper,other

A variety of other selection methods have been proposed, including **predictive sum of squares** (**PRESS**) [126, 127]. Despite many publications on this topic, it seems that none of the methods are used widely in medical imaging practice. All of the methods described in this section attempt to approximate the "optimal" value $\beta_{\mathrm{o}}$ in (2.5.1). In practice, squared error may be a suboptimal metric for imaging, which may limit the practical impact of such methods.

A drawback of most methods for selecting $\beta$ is that one must compute $\hat{\boldsymbol{x}}_\beta$ for many values of $\beta$. One can reduce computation by pruning poor choices of $\beta$ while iterating [128]. Frommer and Maass [129] describe a more efficient method for applying CG to the case of Tikhonov–Phillips Regularization (where $\mathsf{R} = \boldsymbol{I}$) for multiple $\beta$ values. Another option is to find a scheme that chooses $\beta$ adaptively during an iterative algorithm using a **feedback** mechanism [77, 103].

To avoid computing $\hat{\boldsymbol{x}}_\beta$ for many values of $\beta$, another alternative is to use $\beta = 0$ and initialize some iterative algorithm with a uniform image $\boldsymbol{x}^{(0)}$ and then stop the iterations before $\boldsymbol{x}^{(n)}$ becomes "too noisy." A drawback of this approach is that the final image depends on the choice of iterative algorithm, not just on the cost function $\Psi$. Numerous publications have explored **stopping rules** for such methods [96, 130–136].

This section considered methods for choosing a single regularization parameter $\beta$. There are also data-driven methods for selecting space-variant regularization parameters adaptively, *e.g.*, [137].

## 2.6 Limiting behavior (s,reg,limit)

s,reg,limit

This section analyzes the properties of a QPWLS estimator as the regularization parameter increases. (See Problem 2.3 for extensions to penalize-likelihood estimation with nonquadratic regularization, and Problem 2.3 for extensions to temporal regularization for dynamic scans.)

As seen in Example 2.5.1, for quadratic regularization the PWLS estimator has the form

$$\hat{\boldsymbol{x}}_\beta = [\mathsf{F} + \beta\mathsf{R}]^{-1} \boldsymbol{A}'\boldsymbol{W}\boldsymbol{y},$$

where $\mathsf{F} = \boldsymbol{A}'\boldsymbol{W}\boldsymbol{A}$ is a (Hermitian) symmetric positive-semidefinite matrix, as is $\mathsf{R} = \boldsymbol{C}'\boldsymbol{C}$. We further assume that $\mathsf{F}$ and $\mathsf{R}$ have **disjoint null spaces**, so that $\mathsf{F} + \beta\mathsf{R}$ is positive definite for any $\beta > 0$.

Because $\mathsf{R}$ is symmetric positive-semidefinite, it has an orthonormal eigen-decomposition of the form

$$\mathsf{R} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}' = \begin{bmatrix} \boldsymbol{U}_1 & \boldsymbol{U}_0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{U}_1 & \boldsymbol{U}_0 \end{bmatrix}',$$

where $\boldsymbol{U}$ is unitary and $\boldsymbol{\Sigma}_1$ is positive definite. The columns of the matrix $\boldsymbol{U}_0$ span the null space of $\mathsf{R}$. For a typical penalty function based on 1st-order differences, the null space of $\mathsf{R}$ is uniform images, *i.e.*,

$$\boldsymbol{U}_0 = \frac{1}{\sqrt{n_{\mathrm{p}}}}\boldsymbol{1}, \tag{2.6.1}$$

e,reg,limit,null1

where $\boldsymbol{1}$ denotes the vector of ones of length $n_{\mathrm{p}}$.

Because $\mathsf{R}$ and $\mathsf{F}$ have disjoint null spaces, one can verify that

$$\boldsymbol{B} \triangleq \boldsymbol{U}_0'\mathsf{F}\boldsymbol{U}_0$$

is positive definite. To proceed we express $\mathsf{F}$ in terms of the basis $\boldsymbol{U}$ as follows:

$$\boldsymbol{U}'\mathsf{F}\boldsymbol{U} = \begin{bmatrix} \boldsymbol{N} & \boldsymbol{M}' \\ \boldsymbol{M} & \boldsymbol{B} \end{bmatrix}.$$

Note that even though $\boldsymbol{\Sigma}$ is diagonal, $\boldsymbol{B}$ and $\boldsymbol{N}$ are not diagonal in general. The PWLS estimator involves the term

$$\begin{aligned}[\mathsf{F} + \beta \mathsf{R}]^{-1} &= \boldsymbol{U} \left[ \begin{bmatrix} \boldsymbol{N} & \boldsymbol{M}' \\ \boldsymbol{M} & \boldsymbol{B} \end{bmatrix} + \beta \begin{bmatrix} \boldsymbol{\Sigma}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \right]^{-1} \boldsymbol{U}' = \boldsymbol{U} \begin{bmatrix} \boldsymbol{N} + \beta \boldsymbol{\Sigma}_1 & \boldsymbol{M}' \\ \boldsymbol{M} & \boldsymbol{B} \end{bmatrix}^{-1} \boldsymbol{U}' \\ &= \boldsymbol{U} \begin{bmatrix} \left[\boldsymbol{N} + \beta \boldsymbol{\Sigma}_1 - \boldsymbol{M}' \boldsymbol{B}^{-1} \boldsymbol{M}\right]^{-1} & -\left[\boldsymbol{N} + \beta \boldsymbol{\Sigma}_1\right]^{-1} \boldsymbol{M}' \boldsymbol{\Delta}^{-1} \\ -\boldsymbol{\Delta}^{-1} \boldsymbol{M} \left[\boldsymbol{N} + \beta \boldsymbol{\Sigma}_1\right]^{-1} & \boldsymbol{B}^{-1} \end{bmatrix} \boldsymbol{U}', \end{aligned}$$

using (26.1.11), where the Schur complement is $\boldsymbol{\Delta} \triangleq \boldsymbol{B} - \boldsymbol{M} \left[\boldsymbol{N} + \beta \boldsymbol{\Sigma}_1\right]^{-1} \boldsymbol{M}'$. Because $\boldsymbol{\Sigma}_1$ is positive definite, $\left[\boldsymbol{N} + \beta \boldsymbol{\Sigma}_1\right]^{-1} \to \boldsymbol{0}$ and $\boldsymbol{\Delta} \to \boldsymbol{B}$ as $\beta \to \infty$. Thus,

$$\lim_{\beta \to \infty} \hat{\boldsymbol{x}}_\beta = \boldsymbol{U} \begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{B}^{-1} \end{bmatrix} \boldsymbol{U}' \boldsymbol{A}' \boldsymbol{W} \boldsymbol{y} = \boldsymbol{U}_0 \left[\boldsymbol{U}_0' \mathsf{F} \boldsymbol{U}_0\right]^{-1} \boldsymbol{U}_0' \boldsymbol{A}' \boldsymbol{W} \boldsymbol{y}. \tag{2.6.2}$$

e,reg,limit,xh

In particular, in the usual case (2.6.1),

$$\lim_{\beta \to \infty} \hat{\boldsymbol{x}}_\beta = \boldsymbol{1} (\boldsymbol{1}' \boldsymbol{A}' \boldsymbol{W} \boldsymbol{A} \boldsymbol{1})^{-1} \boldsymbol{1}' \boldsymbol{A}' \boldsymbol{W} \boldsymbol{y}.$$

As expected, this limit is the same estimator that is found by assuming the image is uniform, *i.e.*, $\boldsymbol{x} = \boldsymbol{1}\alpha$ and then estimating the coefficient by WLS:

$$\hat{\boldsymbol{x}} = \boldsymbol{1} \hat{\alpha}, \qquad \hat{\alpha} = \arg\min_\alpha \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{1}\alpha\|_{\boldsymbol{W}^{1/2}}^2 .$$

## 2.7 Potential functions (s,reg,pot)

s,reg,pot

The analysis in §1.10.3 showed that the potential weighting function $\omega_\psi$ determines the properties of the restored image $\hat{\boldsymbol{x}}$. Table 2.1 and Table 2.2 summarize many of the options and Fig. 2.7.1 shows many of the weighting functions $\omega_\psi$.

| Name | $\psi(z)$ | $\omega_\psi(z)$ | Comments |
|---|---|---|---|
| quadratic (gaussian pdf) | $\frac{\lvert z\rvert^2}{2}$ | $1$ | simplest not edge preserving |
| Huber | $\begin{cases} \lvert z\rvert^2/2, & \lvert z\rvert \le \delta \\ \delta \lvert z\rvert - \delta^2/2, & \lvert z\rvert > \delta \end{cases}$ | $\begin{cases} 1, & \lvert z\rvert \le \delta \\ \delta / \lvert z\rvert, & \lvert z\rvert > \delta \end{cases}$ | not strictly convex not twice different. |
| hyperbola [138–140] | $\delta^2 \left[\sqrt{1 + \lvert z/\delta\rvert^2} - 1\right]$ | $1/\sqrt{1 + \lvert z/\delta\rvert^2}$ | approximate methods for **total variation** |
| log cosh [141, 142] | $\delta^2 \log\cosh(\lvert z/\delta\rvert)$ | $\frac{\tanh(\lvert z/\delta\rvert)}{\lvert z/\delta\rvert}$ | |
| Lange1 [143] | $\frac{\lvert z\rvert^2/2}{1 + \lvert z/\delta\rvert}$ | $\frac{1 + \lvert z/\delta\rvert/2}{(1 + \lvert z/\delta\rvert)^2}$ | |
| Lange3 [143] Fair [144–146] | $\delta^2 [\lvert z/\delta\rvert - \log(1 + \lvert z/\delta\rvert)]$ | $\frac{1}{1 + \lvert z/\delta\rvert}$ | |
| Li [147] | $\delta^2 \left[\left\lvert\frac{z}{\delta}\right\rvert \arctan\left(\left\lvert\frac{z}{\delta}\right\rvert\right) - \frac{1}{2}\log(1 + \left\lvert\frac{z}{\delta}\right\rvert^2)\right]$ | $\frac{\arctan(z/\delta)}{z/\delta}$ | $\approx$ hyperbola requires arctan |
| Absolute value (TV) (Laplacian pdf) | $\lvert z\rvert$ | $\frac{1}{\lvert z\rvert}$ | not differentiable $\omega_\psi$ unbounded |
| Generalized gaussian [148, 149] | $\lvert z\rvert^p, \quad 1 < p \le 2$ | $p \lvert z\rvert^{p-2}$ | $\omega_\psi$ unbounded for $p < 2$, not twice differentiable |
| Absolute entropy [150] | $\delta^2 (1 + \lvert z/\delta\rvert) \log(1 + \lvert z/\delta\rvert)$ | $\frac{1 + \log(1 + \lvert z/\delta\rvert)}{\lvert z/\delta\rvert}$ | $\omega_\psi$ unbounded |

Table 2.1: Table of (symmetric) *convex* potential functions. The parameter $\delta$ is positive throughout. All of the cases with bounded surrogate curvatures are normalized to $\omega_\psi(0) = 1$.

tab,potent

| Name | $\psi(z)$ | $\omega_\psi(z)$ | Comments |
|---|---|---|---|
| arctan [151] | $\frac{\delta^2}{2} \arctan\left(\left\|z/\delta\right\|^2\right)$ | $\frac{1}{1+\left\|z/\delta\right\|^4}$ | not convex |
| Beaton/Tukey biweight [152] | $\frac{\delta^2}{6}\left[1-\max\left(1-\left\|\frac{z}{\delta}\right\|^2, 0\right)^3\right]$ | $z \max\left(1-\left\|\frac{z}{\delta}\right\|^2, 0\right)^2$ | not convex |
| Cauchy ($t$ pdf) [145, 153–155] | $\frac{\delta^2}{2} \log\left(1+\left\|z/\delta\right\|^2\right)$ | $\frac{1}{1+\left\|z/\delta\right\|^2}$ | not convex aka Lorentzian [156] |
| mixture-of-exp's [157, 158] | $\log(1+\left\|z/\delta\right\|)$ | $\frac{1}{\left\|z/\delta\right\|\left(1+\left\|z/\delta\right\|\right)}$ | not differentiable at 0 $\omega_\psi$ unbounded |
| Geman & McClure [159] | $\frac{\delta^2}{2}\frac{\left\|z/\delta\right\|^2}{1+\left\|z/\delta\right\|^2}$ | $1/\left(1+\left\|z/\delta\right\|^2\right)^2$ | not convex |
| Geman & Reynolds [160] | $\frac{\left\|z\right\|}{1+\left\|z\right\|}$ | $\frac{1}{\left\|z\right\|}\frac{1}{(1+\left\|z\right\|)^2}$ | not convex not differentiable at 0 $\omega_\psi$ unbounded |
| Potts [wiki] [161] | $\mathbb{I}_{\{\left\|z\right\|>\delta\}}$ | undefined | not convex not differentiable at $\pm\delta$ |
| CEL0 [162] | $1-\left(\left\|z/\delta\right\|-1\right)^2 \mathbb{I}_{\{\left\|z\right\|\le\delta\}}$ | undefined | not convex not differentiable at $\pm\delta$ |
| Welsh [163] | $\delta^2\left(1-\mathrm{e}^{-\left\|z/\delta\right\|^2/2}\right)$ | $\mathrm{e}^{-\left\|t/\delta\right\|^2/2}$ | not convex |

Table 2.2: Table of (symmetric) *non-convex* potential functions. The parameter $\delta$ is positive throughout. All of the cases with bounded surrogate curvatures are normalized to $\omega_\psi(0) = 1$. tab,potent,non-convex

Most of the choices in Table 2.1 and Table 2.2 have a selectable "shape" parameter, $\delta$, that controls the **edge-preserving** characteristics[5]; see §1.10.2 and §1.10.3. Potential functions with more shape parameters have also been proposed, *e.g.*, [164].

A variety of desiderata for $\omega_\psi$ have been proposed, *e.g.*, [147, 160], including the following properties: continuity, symmetry, and positivity. It is logical to require that $\omega_\psi$ be nonincreasing for $z > 0$, and for edge preservation: $\lim_{z\to\infty} z\,\omega_\psi(z) \in (0, \infty)$. Some authors argue that $\psi$ should be **convex**, *i.e.*, $\dot{\psi}(z) = \frac{\mathrm{d}}{\mathrm{d}z}(z\,\omega_\psi(z)) \ge 0$, whereas others have argued that $\psi$ should be *concave* on $(0, \infty)$, and should have a finite asymptote: $\lim_{z\to\infty} \psi(z) < \infty$, *e.g.*, [160].

From a computational perspective, one might add to this list that we would like to be able to evaluate $\omega_\psi$ quickly, avoiding transcendental function evaluations if possible. The importance of such considerations depends on computing resources; often the computation demands of the log-likelihood term far outweigh those of the roughness penalty.

Several "generalized families" of potential functions have been proposed in the literature. Some of these are summarized and generalized next. To my knowledge, there is no theory that establishes optimality any of these families; the "best" choice is application dependent.

### 2.7.1 Generalized Gaussian

The **generalized gaussian** family, defined by [148]

$$\psi(z) = |z|^p, \quad 1 < p \le 2, \tag{2.7.1}$$ e,reg,pot,gg

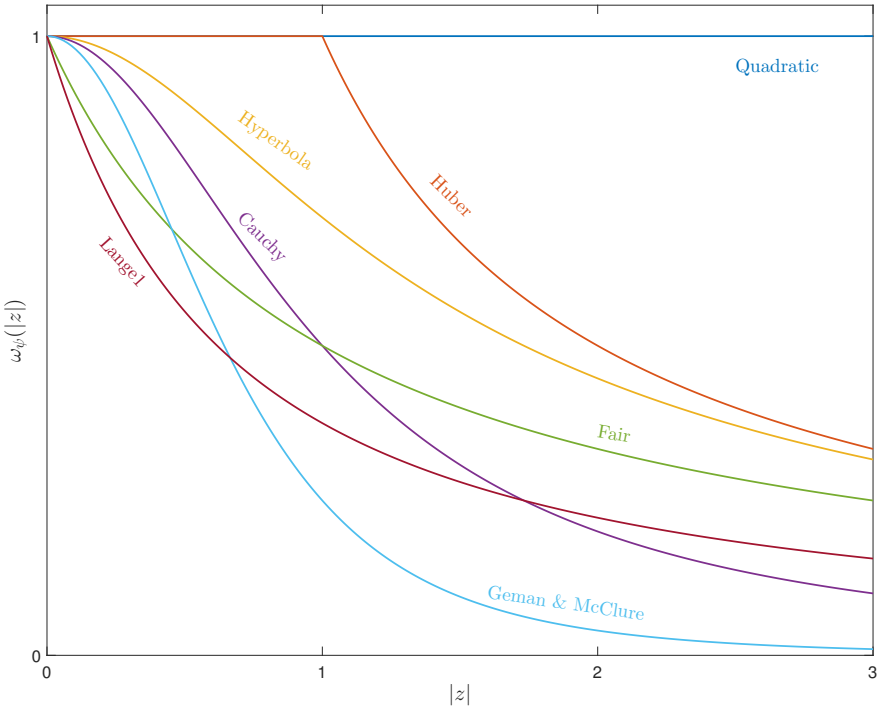
includes the quadratic function as a special case, and has a desirable **scale invariance** property [149]. Unfortunately this function is not twice differentiable at zero for $p < 2$, which complicates some optimization methods.

### 2.7.2 Generalized Huber

A further generalization of the generalized gaussian is to have a transition point $\delta$ where the potential switches (with continuous derivative) from $|z|^p$ to a different power $|z|^q$. An example is:

$$\psi(z) = \begin{cases} \frac{1}{2}|z|^p, & |z| \le \delta \\ \frac{1}{2}\frac{p}{q}\delta^{p-q}|z|^q + \frac{1}{2}\left(1-\frac{p}{q}\right)\delta^p, & |z| > \delta, \end{cases} \qquad \omega_\psi(z) = \begin{cases} \frac{p}{2}|z|^{p-2}, & |z| \le \delta \\ \frac{p}{2}\delta^{p-q}|z|^{q-2}, & |z| > \delta, \end{cases} \tag{2.7.2}$$ e,reg,pot,genhub

[5] It is claimed in [150] that the "absolute entropy" function does "not require the selection of structural parameters." That paper uses $\delta = e^{-1}$, which surely is a selection...




fig_reg_wpot1

Figure 2.7.1: Bounded potential weighting functions $\omega_\psi(z)$ from Table 2.1 and Table 2.2.

where typically $1 \leq q \ll p \leq 2$. (Stevenson et al. proposed a similar potential [165].) Unfortunately, for $p < 2$ both the original generalized gaussian and the above generalization have **unbounded curvature** at the origin, precluding the use of algorithms like (1.11.4). Chartrand considers the case $p = 2$ and $q < 1$ as a (non-convex) sparsity prior [166].

Taking $p = 2$ and $q = 1$ above, the expression simplifies to the **Huber potential** (1.10.9):

$$\psi(z) = \begin{cases} \frac{1}{2}|z|^2, & |z| \leq \delta \\ \delta|z| - \frac{1}{2}\delta^2, & |z| > \delta, \end{cases} \qquad \omega_\psi(z) = \begin{cases} 1, & |z| \leq \delta \\ 1/|z/\delta|, & |z| > \delta. \end{cases} \tag{2.7.3}$$

e,reg,pot,huber

This choice originated in robust statistics and has certain min-max optimality properties in that context [167].

One can write the Huber potential in a **dual formulation** [168]:

$$\begin{aligned} \psi(z) &= \underset{\gamma \in [-1,1]}{\arg\min}\, \delta^2 \left( \gamma |z/\delta| - \frac{1}{2}\gamma^2 \right) \\ &= \underset{\gamma \in [-1,1]}{\arg\min}\, \delta^2 \left( \frac{1}{2}|z/\delta|^2 - \frac{1}{2}(\gamma - |z/\delta|)^2 \right). \end{aligned}$$

Writing the Generalized Huber potential (2.7.2) in a dual formulation is left as an exercise.

s,reg,pot,qgg

### 2.7.3 Generalized Gaussian "$q$-generalized" (s,reg,pot,qgg)

Instead of "switching" abruptly from $|z|^p$ to $|z|^q$ as in (2.7.2), an alternative is to transition gradually between the two, *e.g.*, by using the following family of potential functions:

$$\psi(z) = \frac{\frac{1}{2}|z|^p}{\left(1 + |z/\delta|^{(p-q)r}\right)^{1/r}}, \qquad \omega_\psi(z) = |z|^{p-2} \frac{\frac{p}{2} + \frac{q}{2}|z/\delta|^{(p-q)r}}{\left(1 + |z/\delta|^{(p-q)r}\right)^{1+1/r}}, \tag{2.7.4}$$

e,reg,pot,qgg

where $r > 0$ and usually $1 \leq q \ll p \leq 2$. De Man et al. explored the case $p = 2$ and $(p-q)r = 2$ [169–171], generalizing the Geman & McClure potential [159, 172]. Thibault et al. studied the case $r = 1$ and found the choice $p = 2$ and $q \approx 1.2$ to be particularly desirable for X-ray CT [173]. The sub-family where $r = 1$ and $q = 0$ generalize the Geman & Reynolds potential functions [159, 160]. Special cases are tabulated below, where $*$ denotes arbitrary values.

| $p$ | $q$ | $r$ | name |
|---|---|---|---|
| 2 | 2 | $*$ | quadratic |
| $*$ | $p$ | $*$ | generalized gaussian |
| 2 | 1 | 1 | Lange1 [143] |
| 2 | 0 | 1 | Geman & McClure [159, 172] |
| 3/2 | 0 | 1 | in [174], according to [160] |
| 1 | 0 | 1 | Geman & Reynolds [160] |
| $*$ | 0 | 1 | Generalized Geman & Reynolds |

Letting $m = 1/r$, $n = (p-q)r$, and $x = |z/\delta|^n$, the curvature of this potential function is:

$$\ddot{\psi}(z) = |z|^{p-2} \frac{ax^2 + bx + c}{(1+x)^{m+2}},$$

where $a = (p - mn)(p - 1 - mn)/2 = q(q-1)/2$, $b = [2p(p-1) + mn(1-2p) - mn^2]/2$, $c = p(p-1)/2$. To ensure convexity (by nonnegativity of $\ddot{\psi}$) it is necessary to have $a \geq 0$, so hereafter we assume that $mn \leq p - 1$. Because $mn = p - q$, equivalently $1 \leq q$. We also need $c \geq 0$ or equivalently $1 \leq p$. To explore convexity further, recall that polynomials of the form $ax^2 + bx + c$ with $a \geq 0$ and $c \geq 0$ are nonnegative for $x \geq 0$ if $b \geq 0$ or if $b^2 \leq 4ac$. Here, one can verify that

$$2b = 2(r+1)(p-1)(q-1) + (p-1)\left[(2-p)r + (1-r)\right] + (q-1)\left[(2-q)r + (1-r)\right].$$

Thus $b \geq 0$, and hence $\psi$ is convex, if $0 < r \leq 1$ and $1 \leq p, q \leq 2$, because these conditions ensure that all the parenthesized terms are nonnegative. These conditions generalize slightly those derived in [173]. The most useful case is probably where $r = 1$ and $1 \leq q \leq p \leq 2$.
Futhermore, one can verify that

$$b^2 - 4ac = \frac{1}{4} mn^2 \left[ m\left(n^2 + (4p-2)n + 1\right) - 4p(p-1) \right],$$

so for convexity of $\psi$ it suffices to have

$$m \leq \frac{4p(p-1)}{n^2 + (4p-2)n + 1}.$$

In particular, in the typical case where $p = 2$, it suffices to have $m \leq \frac{8}{n^2+6n+1}$. Specifically, when $n = 2$ it suffices to have $m \leq 8/17$, consistent with [169].

s,reg,pot,gf1

### 2.7.4 Generalized Fair potential: 1st order (s,reg,pot,gf1)

A drawback of (2.7.4) is that evaluating $\omega_\psi$ requires computing powers (unless $p$ and $(2-q)r$ and $1 + 1/r$ are integers). A family of potential functions that avoids power operations for $\omega_\psi$ is the following **generalized Fair potential** functions:

$$\psi(z) = \frac{\delta^2}{2b^3}\left( ab^2 |z/\delta|^2 + 2b(b-a)|z/\delta| + 2(a-b)\log(1 + b|z/\delta|) \right), \qquad \omega_\psi(z) = \frac{1 + a|z/\delta|}{1 + b|z/\delta|}, \tag{2.7.5}$$

e,reg,pot,gf1

where $b \geq a \geq 0$. Special cases are tabulated below.

| $a$ | $b$ | name |
|---|---|---|
| any | $a$ | quadratic |
| 0 | 1 | Lange3 [143] / Fair [144–146] |

One can show that

$$\ddot{\psi}(z) = \frac{1 + 2a|z/\delta| + ab|z/\delta|^2}{(1 + b|z/\delta|)^2}$$

so this potential function is strictly convex when $b \geq a \geq 0$. Although the potential function itself in (2.7.5) is somewhat complicated looking, often what matters most for implementation is $\omega_\psi$, which is very simple here.

By choosing $a$ and $b$, one can make the weighting function $\omega_\psi$ in (2.7.5) approximate another potential weighting function $\tilde{\omega}_\psi$ that has a "cusp" at 0, such as the Lange1 potential shown in Fig. 2.7.1. Suppose we match such that $\omega_\psi(s_k\delta) = w_k \triangleq \tilde{\omega}_\psi(s_k\delta)$ where $0 < s_1 < s_2$. Solving for $a$ and $b$ yields

$$\begin{bmatrix} a \\ b \end{bmatrix} = \frac{1}{(w_1 - w_2)s_1 s_2} \begin{bmatrix} w_2 s_2(1 - w_1) - w_1 s_1(1 - w_2) \\ s_2(1 - w_1) - s_1(1 - w_2) \end{bmatrix}. \tag{2.7.6}$$

e,reg,pot,gf1,coef

MIRT *See the* `'gf1'` *and* `'gf1-fit'` *options of* `potential_fun.m`.

Fig. 2.7.2 compares $\omega_\psi(z)$ for the $q$-generalized gaussian potential with $r = 1$, $p = 2$ and $q = 1.2$ and a generalized Fair potential with parameters chosen using (2.7.6) so that the $\omega_\psi$ values are matched at $|z/\delta| \in \{1, 10\}$. Qualitatively they match very closely.

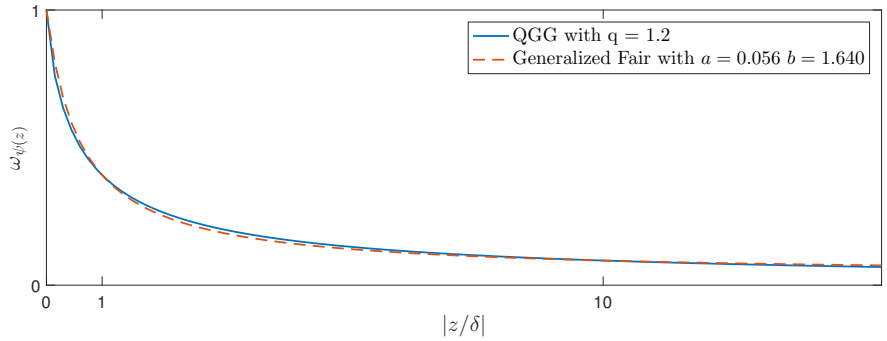



Figure 2.7.2: Comparison of $\omega_\psi(z)$ for the $q$-generalized gaussian potential with $q = 1.2$ and $r = 1$ and a generalized Fair potential with selected parameters. fig_reg_wpot_fair

s,reg,pot,gf2

### 2.7.5 Generalized Fair potential: 2nd order (s,reg,pot,gf2)

The weighting function for (2.7.5) has only two degrees of freedom: $\delta$ and the ratio $b/a$. Furthermore, $\omega_\psi$ does not decrease all the way to zero as $|z/\delta| \to \infty$, unless $a = 0$. To overcome these limitations, consider the following family:

$$\psi(z) = \frac{\delta^2}{b^2 + ac^2} \left[ (b + ac) \left| z/\delta \right| - \log(1 + b \left| z/\delta \right|) - a \log(1 + c \left| z/\delta \right|) \right], \tag{2.7.7}$$

e,reg,pot,gf2

where $a \geq 0,\ b, c > 0$. Using the Taylor expansion $\log(1 + x) \approx x - x^2/2$, one can verify that $\psi(z) \approx |z|^2 / 2$ for $|z/\delta| \ll 1$. One can also verify that

$$\dot\psi(z) = z \frac{1 + bc \frac{b+ac}{b^2+ac^2} \left| z/\delta \right|}{1 + (b + c) \left| z/\delta \right| + bc \left| z/\delta \right|^2}, \qquad \ddot\psi(z) = \frac{1 + 2bc \frac{b+ac}{b^2+ac^2} \left| z/\delta \right| + (1 + a) \frac{(bc)^2}{b^2+ac^2} \left| z/\delta \right|^2}{\left( 1 + (b + c) \left| z/\delta \right| + bc \left| z/\delta \right|^2 \right)^2},$$

so $\psi$ is strictly convex. By design, the weighting function for (2.7.10) has the following rational form:

$$\omega_\psi(z) = \frac{1 + bc \frac{b+ac}{b^2+ac^2} \left| z/\delta \right|}{1 + (b + c) \left| z/\delta \right| + bc \left| z/\delta \right|^2}. \tag{2.7.8}$$

e,reg,pot,gf2,wpot

One can verify that

$$\frac{\mathrm{d}}{\mathrm{d}z} \omega_\psi(0^+) = -\frac{1}{\delta} \frac{b^3 + ac^3}{b^2 + ac^2}, \tag{2.7.9}$$

e,reg,pot,gf2,wpot0

which is also negative. When $a = 0$ and $b = 1$, this potential function degenerates to the Lange3 [143] or Fair [144–146] choice. Determining whether this potential function could match others better than (2.7.5) is an *open problem.*

s,reg,pot,p12

### 2.7.6 Convex arctan potential (s,reg,pot,p12)

A limitation of (2.7.5) is that $\frac{\mathrm{d}}{\mathrm{d}t} \omega_\psi(0^+) = (a - b)/\delta < 0$ in the usual case where $a < b$. The 2nd-order case (2.7.9) is similar. So those weighting functions always have a cusp at zero. Some of the weighting functions illustrated in Fig. 2.7.1 have zero slope at $t = 0$, such as the hyperbola. But the hyperbola weighting function requires a square root operation. For a family of potential functions that can approximate weighting functions having zero slope at $t = 0$ while also having a simple weighting function, consider the following:

$$\psi(z) = \delta^2 \frac{1 + \alpha}{2} \left[ \left| z/\delta \right| - \frac{1 + \alpha}{\sqrt{\alpha}} \arctan\left( \frac{\left| z/\delta \right| + 1}{\sqrt{\alpha}} \right) \right], \tag{2.7.10}$$

e,reg,pot,p12

where $\alpha > 0$. One can verify that

$$\dot\psi(z) = z \frac{1 + \frac{1}{2} \left| z/\delta \right|}{1 + \frac{2}{1+\alpha} \left| z/\delta \right| + \frac{1}{1+\alpha} \left| z/\delta \right|^2}, \qquad \ddot\psi(z) = \frac{1 + \left| z/\delta \right|}{\left( 1 + \frac{2}{1+\alpha} z + \frac{1}{1+\alpha} \left| z \right|^2 \right)^2},$$

so $\psi$ is strictly convex. By design, the weighting function for the potential (2.7.10) has the following rational form:

$$\omega_\psi(z) = \frac{1 + \frac{1}{2} \left| z/\delta \right|}{1 + \frac{2}{1+\alpha} \left| z/\delta \right| + \frac{1}{1+\alpha} \left| z/\delta \right|^2}. \tag{2.7.11}$$

e,reg,pot,p12,wpot

One can verify that

$$\frac{\mathrm{d}}{\mathrm{d}z} \omega_\psi(0^+) = \frac{1}{\delta} \left( \frac{1}{2} - \frac{2}{\alpha + 1} \right) = \frac{1}{\delta} \frac{1}{2} \frac{\alpha - 3}{\alpha + 1},$$

so for $\omega_\psi$ to be decreasing for $z \geq 0$ we want $\alpha \leq 3$. Choosing $\alpha = 3$ provides a flat weighting function at $z = 0$.

Fig. 2.7.3 compares weighting functions $\omega_\psi(z)$ for the hyperbola potential and for the convex arctan potential. The two agree very closely (within about 7%) but the convex arctan potential avoids the square root function.

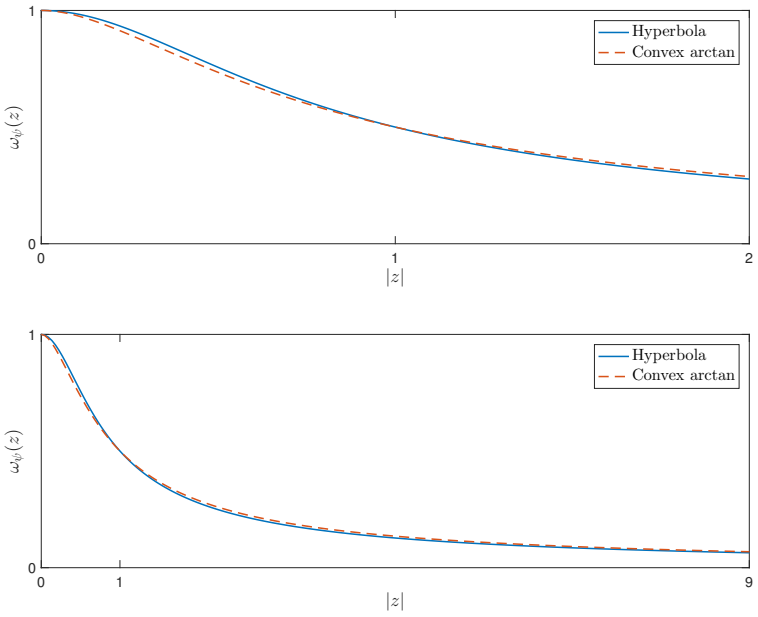



Figure 2.7.3: The weighting functions $\omega_\psi(z)$ for the hyperbola potential with $\delta = 1/\sqrt{3}$ and for the convex arctan potential of (2.7.10) with $\delta = 1/(1+\sqrt{5})$ and $\alpha = 3$. These $\delta$ values ensure that $\omega_\psi(1) = 1/2$. fig_reg_wpot_hyper

For even more flexibility, one might try to design a family of potential functions with the following general rational form for the weighting function:

$$\omega_\psi(z) = \frac{1 + a\,|z/\delta|}{1 + b\,|z/\delta| + c\,|z/\delta|^2}, \tag{2.7.12}$$

e,reg,pot,p12,wpot-gen

where $a, b, c \geq 0$. Because $\frac{\mathrm{d}}{\mathrm{d}z}\,\omega_\psi(z) = \frac{1}{\delta}\frac{a-b-2c|z/\delta|-ac|z/\delta|^2}{(1+b|z/\delta|+c|z/\delta|^2)^2}$, for $t > 0$, usually we will want to choose $a \leq b$ so that $\omega_\psi$ is a decreasing function. Special cases are tabulated below.

| $a$ | $b$ | $c$ | name |
|---|---|---|---|
| any | $a$ | 0 | quadratic |
| 1/2 | 2 | 1 | Lange1 [143] |
| 0 | 1 | 0 | Lange3 [143] / Fair [144–146] |
| 0 | 0 | 1 | Cauchy ($t$ pdf) [145, 153–155] |

One can show that

$$\ddot{\psi}(z) = \frac{1 + 2a\,|z/\delta| + (ab - c)\,|z/\delta|^2}{\left(1 + b\,|z/\delta| + c\,|z/\delta|^2\right)^2}$$

so this potential function is strictly convex if and only if $c \leq ab$. Unfortunately, it is difficult to determine the potential function $\psi$ that leads to the weighting function (2.7.12) in general. Algorithms that require a line search need to have $\psi$ available. However, algorithms that that use only $\omega_\psi$ and $\dot{\psi}(z) = z\,\omega_\psi(z)$ could use the general form (2.7.12).

### 2.7.7 Hypergeometric (generalized hyperbola) (s,reg,pot,hyper2)

s,reg,pot,hyper2

Many of the potential functions in Table 2.1 are special cases of the following very general form:

$$\psi(z) = \delta^2 \int_0^{|z/\delta|} s\,\frac{c + as^p}{(1 + bs^q)^r}\,\mathrm{d}s,$$

where $a, b, c \geq 0$. To avoid degeneracy, we require $a > 0$ and/or $c > 0$. For rational values of $p, q, r$, this integral relates to the **hypergeometric** function of Gauss [175, p. 555]. This family is designed to satisfy

$$\omega_\psi(z) = \frac{c + a\,|z/\delta|^p}{(1 + b\,|z/\delta|^q)^r}, \tag{2.7.13}$$

e,reg,pot,hyper2,wpot

and to ensure that $\omega_\psi$ is bounded for large values of $|z|$ we require that $0 \leq p \leq qr$. For this family, one can show that

$$\dot{\psi}(z) = z\frac{c + a\,|z/\delta|^p}{(1 + b\,|z/\delta|^q)^r}, \qquad \ddot{\psi}(z) = \frac{c + a(p+1)\,|z/\delta|^p + bc(1-qr)\,|z/\delta|^q + ab(p+1-qr)\,|z/\delta|^{p+q}}{(1 + b\,|z/\delta|^q)^{r+1}}.$$

Thus, this potential function is strictly convex if: $c > 0$ and $qr \leq 1$, or if: $c = 0$ and $qr \leq 1 + p$. Otherwise typically it is not. Special cases are tabulated below.

| $p$ | $q$ | $r$ | $a$ | $b$ | $c$ | name |
|---|---|---|---|---|---|---|
| 0 | * | 0 | * | * | 1 | quadratic |
| 0 | 2 | 1/2 | 0 | 1 | 1 | hyperbola |
| 0 | 2 | 1 | 0 | 1 | 1 | Cauchy |
| 0 | 2 | 2 | 0 | 1 | 1 | Geman & McClure [159, 172] |
| 1 | 1 | 2 | 1/2 | 1 | 1 | Lange1 |
| 0 | 1 | 1 | 0 | 1 | 1 | Lange3 / Fair |
| 0 | 4 | 1 | 0 | 1 | 1 | arctan |
| 1 | 1 | 1 | * | * | 1 | generalized Fair (2.7.5) |
| * | 0 | 0 | 1 | 0 | 1 | generalized gaussian |

There is an explicit expression for this potential when $q = 2$, $b = 1$, and $a = p = 0$:

$$\psi(z) = c \begin{cases} \frac{\delta^2}{2} \log\left(1 + |z/\delta|^2\right), & r = 1 \\ \frac{\delta^2}{2(1-r)} \left[\left(1 + |z/\delta|^2\right)^{1-r} - 1\right], & r \geq 0,\ r \neq 1. \end{cases}$$

There is also an explicit, but lengthy, expression when $q = p = 1$ and $c = 0$.

s,reg,pot,tab

### 2.7.8 Tabulated potential functions (s,reg,pot,tab)

Several of the potential functions described above and in Table 2.1 have weighting functions that involve somewhat expensive operations such as powers (2.7.4) (2.7.13), exponentials, or trigonometric functions. One way to avoid such operations is to use a **look-up table**. This section describes methods for representing $\psi$ using tabulated values.

One natural approach is to sample the values of $\dot{\psi}$, *i.e.*, to tabulate $d_k = \dot{\psi}(t_k)$ for $k = 0, \dots, K$, where $t_0 = 0$ and $d_0 = 0$. The design question then becomes how to interpolate $\dot{\psi}$ between these sample values. The following sections describe a few options.

For each method, we will need to use the following table indexing function:

$$k' \triangleq k'(t) = \max \{k \in \{0, 1, \dots, K\} : t_k \le |z|\}. \tag{2.7.14}$$

e,reg,pot,tab,k'

Naturally, table look-up is simplest when the sample points are spaced equally: $t_k = k\Delta$, for $k = 0, \dots, K$, because in this case the indexing function simplifies to a floor function:

$$k' \triangleq k'(t) = \min(\lfloor |z| / \Delta \rfloor, K). \tag{2.7.15}$$

e,reg,pot,tab,k',equal

s,reg,pot,tab,0

#### 2.7.8.1 Zeroth-order interpolation of $\dot{\psi}$ samples

The simplest approach is to use (mostly) **sample and hold interpolation** of the $\dot{\psi}$ samples:

$$\dot{\psi}(z) = \operatorname{sgn}(z) \left( \frac{d_1}{t_1} |z| \, \mathbb{I}_{\{|z| < t_1\}} + \sum_{k=1}^{K} d_k \mathbb{I}_{\{t_k \le |z| < t_{k+1}\}} \right). \tag{2.7.16}$$

e,reg,pot,tab,dpot,0

We set $t_{K+1} = \infty$ so that $\psi(z)$ is a line with slope $d_K$ for $t > t_K$.

The corresponding potential function is piecewise linear (except for being quadratic near 0):

$$\begin{aligned}
\psi(z) &= \int_0^{|z|} \dot{\psi}(\tau) \, \mathrm{d}\tau = \int_0^{|z|} \left( \frac{d_1}{t_1} \tau \mathbb{I}_{\{\tau < t_1\}} + \sum_{k=1}^{K} d_k \mathbb{I}_{\{t_k \le \tau < t_{k+1}\}} \right) \mathrm{d}\tau \\
&= \frac{1}{2} \frac{d_1}{t_1} \left( \min(|z|, t_1) \right)^2 + \left( \sum_{k=1}^{k'-1} d_k \left( t_{k+1} - t_k \right) \right) + d_{k'} \left( |z| - t_{k'} \right) \mathbb{I}_{\{k' > 0\}} \\
&= \begin{cases} \frac{1}{2} \frac{d_1}{t_1} |z|^2, & |z| < t_1 \\ s_{k'} + d_{k'} \left( |z| - t_{k'} \right), & \text{otherwise,} \end{cases}
\end{aligned} \tag{2.7.17}$$

e,reg,pot,tab,pot,0

where $k'$ was defined in (2.7.14) and $s_{k'} \triangleq \frac{1}{2} d_1 t_1 + \sum_{k=1}^{k'-1} d_k \left( t_{k+1} - t_k \right)$ for $k' = 1, \dots, K$. A drawback of this model is that $\psi$ is not differentiable at $\pm t_k$ for $k > 1$ where $d_{k+1} \neq d_k$. The potential function $\psi$ is convex provided the samples are all nondecreasing: $d_{k-1} \le d_k$.

The corresponding weighting function is

$$\omega_\psi(z) = \frac{\dot{\psi}(z)}{z} = \frac{d_1}{t_1} \mathbb{I}_{\{|z| < t_1\}} + \sum_{k=1}^{K} \frac{d_k}{|z|} \mathbb{I}_{\{t_k \le |z| < t_{k+1}\}}. \tag{2.7.18}$$

e,reg,pot,tab,wpot,0

We chose to use a linear segment for $|z| < t_1$ in (2.7.16) so that $\omega_\psi$ would be finite over that range. Note that $\omega_\psi\left(t_k^-\right) = d_{k-1}/t_k$ whereas $\omega_\psi\left(t_k^+\right) = d_k/t_k$ so $\omega_\psi$ is discontinuous at every $t_k$ in general for $k > 1$, which seems undesirable.

For optimization transfer algorithms based on quadratic surrogates, we need a curvature function $\breve{c}$ that is no smaller than $\omega_\psi$. Usually we simply use $\omega_\psi$ itself, but to save the effort of computing the ratio in (2.7.17) we could use the following precomputed ratios:

$$\breve{c}(z) = \frac{d_1}{t_1} \mathbb{I}_{\{|z| < t_1\}} + \sum_{k=1}^{K} \frac{d_k}{t_k} \mathbb{I}_{\{t_k \le |z| < t_{k+1}\}} = \frac{d_{k'}}{t_{k'}}. \tag{2.7.19}$$

e,reg,pot,tab,curv,0

x,reg,pot,tab0,huber

**Example 2.7.1** *If we choose $K = 1$, $t_0 = 0$, $t_1 = \delta$, $t_2 = \infty$, $d_0 = 0$, $d_1 = \delta$, then (2.7.16) corresponds to the Huber function (2.7.3).*

For sparsity-based regularizers, it is important to be able to solve the **shrinkage problem** (see Problem 1.12) also known as the **Moreau proximity operator** (see §27.9.3.6) [62]:

$$\hat{z}(c) = \arg\min_z \frac{1}{2} |z - c|^2 + \beta \, \psi(z). \tag{2.7.20}$$

e,reg,pot,tab,shrink,kost,0

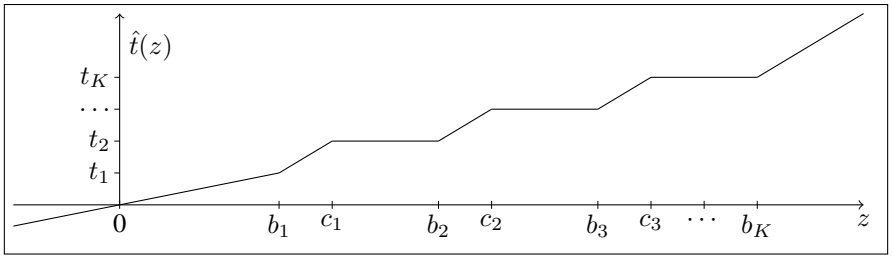



Figure 2.7.4: Shrinkage function (2.7.22) for tabulated potential function using sample-and-hold interpolation of derivative. fig,reg,pot,tab,0

We can solve this problem exactly for the tabulated model (2.7.16). Zeroing the derivative requires solving

$$c = z + \beta\, \dot{\psi}(z) \tag{2.7.21}$$ e,reg,pot,tab,fzero,0

for $z = \hat{z}(c)$, at points where $\dot{\psi}$ is differentiable. If $|z| \le t_1$ then $c = z + \beta \frac{d_1}{t_1} t$ so $\hat{z} = \frac{t_1}{b_1} z$ where $b_k \triangleq t_k + \beta d_k$ provided $|\hat{z}| \le t_1$ or equivalently $|z| \le b_1$. If $t_k < |z| < t_{k+1}$ then $c = z + \beta\, \mathrm{sgn}(z)\, d_k$ so the shrinkage rule is $\hat{z} = \mathrm{sgn}(c)\,(|c| - \beta d_k)$. Focusing on $c > 0$, this solution holds when $t_k < c - \beta d_k < t_{k+1}$ or equivalently when $b_k < c < c_k$ where $c_k \triangleq t_{k+1} + \beta d_k > b_k$. In particular, $\hat{z}(b_k^+) = b_k - \beta d_k = t_k$ and $\hat{z}(c_k^-) = c_k - \beta d_k = t_{k+1}$. Summarizing yields the following piecewise linear shrinkage function, illustrated in Fig. 2.7.4:

$$\hat{z} = \begin{cases} \frac{t_1}{b_1} c, & |c| \le b_1 \\ \mathrm{sgn}(c)\,(|c| - \beta d_k)\,, & b_k \le |c| \le c_k \\ t_{k+1}, & c_k \le |c| \le b_{k+1}. \end{cases} \tag{2.7.22}$$ e,reg,pot,tab,shrink,th,0

In the usual case when the $t_k$ and $d_k$ values are monotone nondecreasing (*e.g.*, when $\psi$ is convex) then there is a unique solution. Unfortunately the breakpoints are unequally spaced in general, so (2.7.22) appears to require many comparison operations to implement. Nevertheless, at least there is an exact solution that is fairly simple so when we minimize a regularized cost function using the tabulated potential (2.7.16), one should be able to reach identical minimizers using algorithms that do and do not use a shrinkage operation (2.7.20).

Because (2.7.22) is piecewise linear, we can implement it exactly using linear interpolation. However, this may require numerous comparison operations because the breakpoints in (2.7.22) are unequally spaced. An alternative may be to use the minimum breakpoint spacing

$$\Delta c = \min(b_1, \{c_k - b_k\}, \{b_{k+1} - c_k\}) = \min(b_1, \{t_{k+1} - t_k\}, \{\beta(d_{k+1} - d_k)\})$$

to tabulate the relationship between $c$ and a nearby $k$ to reduce the number of comparisons. Implementing this version efficiently is an *open problem.*

MIRT *See* `potential_fun.m`.

x,reg,pot,tab,0

**Example 2.7.2** *Fig. 2.7.5 compares the QGG potential function (2.7.4) with $p = 2$ and $q = 1.2$ and $\delta = 10$ to a tabulated approximation (2.7.16) with $K = 50$ and $\Delta = 0.5$. For large $|z|$, the approximation rises linearly whereas QGG rises as $|z|^q$ so an accurate match requires that $K\Delta$ be sufficiently large. The large discontinuities in the weighting function are somewhat disconcerting, although the corresponding derivative $\dot{\psi}(z) = z\, \omega_\psi(z)$ is piecewise constant as dictated by (2.7.16).*

*Fig. 2.7.6 compares the shrinkage function (2.7.20) for QGG (found numerically) and the tabulated approximation (2.7.22); these agree very well.*

### 2.7.8.2 Linear interpolation of $\dot{\psi}$ samples

Another option is to use linear interpolation for $\dot{\psi}$, which seems reasonable because $\dot{\psi}$ is piecewise linear for the quadratic and Huber potentials, leading to the following mathematical model:

$$\begin{aligned} \dot{\psi}(z) &= \mathrm{sgn}(z) \sum_{k=0}^{K} \left( d_k + \frac{|z| - t_k}{t_{k+1} - t_k} (d_{k+1} - d_k) \right) \mathbb{I}_{\{t_k \le |z| < t_{k+1}\}} \\ &= \mathrm{sgn}(z) \sum_{k=0}^{K} (d_k + (|z| - t_k)\, c_k)\, \mathbb{I}_{\{t_k \le |z| < t_{k+1}\}}, \end{aligned} \tag{2.7.23}$$ e,reg,pot,tab,dpot,1

where $c_k \triangleq \frac{d_{k+1} - d_k}{t_{k+1} - t_k}$ for $k = 0, \ldots, K-1$. Usually we set $t_{K+1} = \infty$ and set $c_K = 0$ so that $\psi(z)$ is a line with slope $d_K$ for $z > t_K$. For this design, $\dot{\psi}$ is continuous with $\dot{\psi}(t_k^-) = \dot{\psi}(t_k^+) = d_k$ for $k = 0, \ldots, K$. For this model,

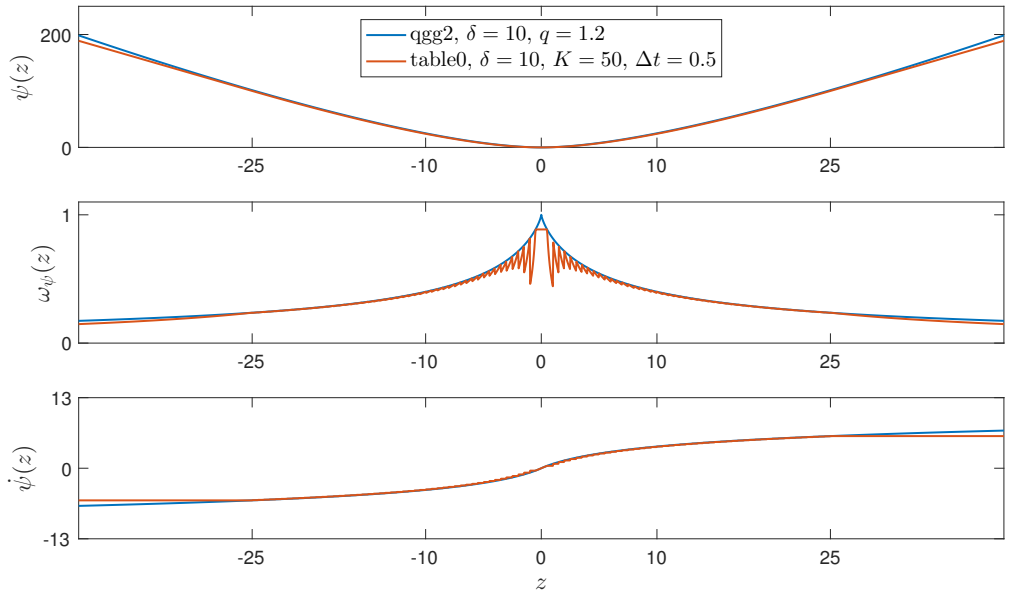



Figure 2.7.5: QGG potential function for $p = 2$ and $q = 1.2$ and tabulated approximation using sample-and-hold interpolation of $\dot{\psi}$ per (2.7.16). fig_reg_pot_table0_qgg2

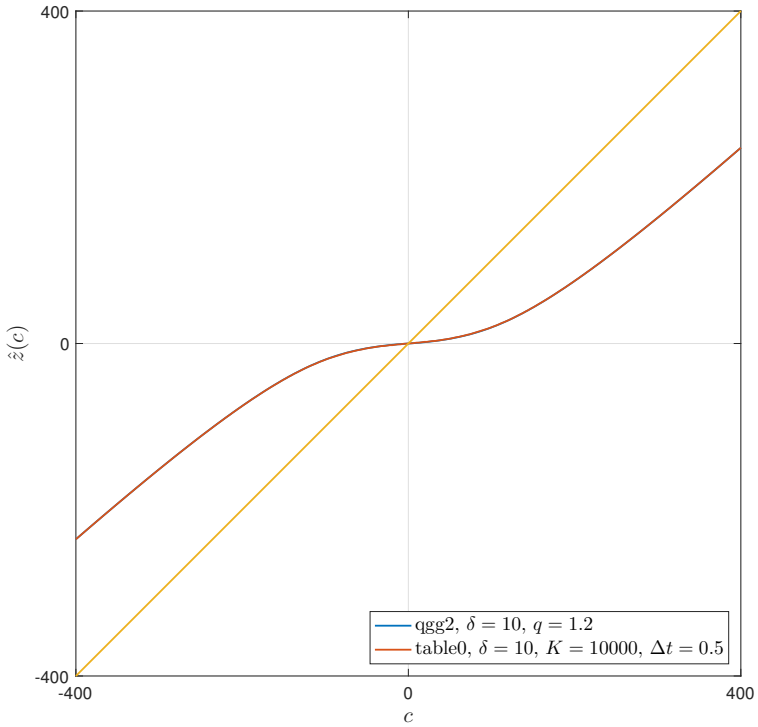
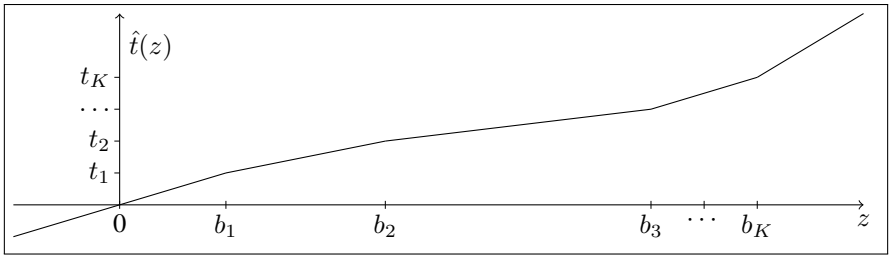



Figure 2.7.6: Shrinkage function $\hat{z}(c)$ for QGG potential function with $p = 2$ and $q = 1.2$ and for its tabulated approximation (2.7.22) using sample-and-hold interpolation of $\dot{\psi}$ per (2.7.16). fig_reg_pot_table0_qgg2_shrink




fig,reg,pot,tab,1

Figure 2.7.7: Shrinkage function (2.7.27) for tabulated potential function using linear interpolation of derivative.

$\psi$ has piecewise constant second derivative:

$$\ddot{\psi}(z) = \sum_{k=0}^{K} c_k \mathbb{I}_{\{t_k \le |z| < t_{k+1}\}},$$

so if the slopes $c_k$ of $\dot{\psi}$ are nonnegative, then $\psi$ is convex, and if each $c_k$ is positive except possibly for $c_K$ then $\psi$ is strictly convex over $(-t_K, t_K)$. This property is similar to the Huber function.

The weighting function for this model has a piecewise reciprocal form:

$$\omega_\psi(z) = \frac{\dot{\psi}(z)}{z} = \begin{cases} d_1/t_1, & |z| < t_1 \\ \frac{d_k + (|z| - t_k)c_k}{|z|}, & t_k \le |z| < t_{k+1},\ 1 \le k \le K. \end{cases} \tag{2.7.24}$$

e,reg,pot,tab,wpot,1

Unlike (2.7.18), here $\omega_\psi(t_k^-) = \omega_\psi(t_k^+) = d_k/t_k$ so here $\omega_\psi$ is continuous. Again, to save computing the ratio in (2.7.24), we can use curvatures based on the upper bound for each interval:

$$\breve{c}(z) = \max_{t_{k'} \le |z| < t_{k'+1}} \omega_\psi(z) = \omega_\psi(t_{k'}) = \frac{d_{k'}}{t_{k'}}. \tag{2.7.25}$$

e,reg,pot,tab,curv,1

The potential function has the form

$$\begin{aligned} \psi(z) &= \int_0^{|z|} \dot{\psi}(\tau)\, \mathrm{d}\tau = \sum_{k=0}^{K} \int_0^{|z|} (d_k + (\tau - t_k)c_k)\, \mathbb{I}_{\{t_k \le \tau < t_{k+1}\}}\, \mathrm{d}\tau \\ &= s_{k'} + (d_{k'} - t_{k'} c_{k'})\,(|z| - t_{k'}) + \frac{c_k}{2}\left(t^2 - t_{k'}^2\right), \end{aligned} \tag{2.7.26}$$

e,reg,pot,tab,pot,1

where for compute efficiency we tabulate the following sum for $k' = 0, 1, \ldots, K$:

$$s_{k'} \triangleq \sum_{k=0}^{k'-1} (d_k - t_k c_k)\,(t_{k+1} - t_k) + \frac{c_k}{2}\left(t_{k+1}^2 - t_k^2\right).$$

This table is needed only if we plan to evaluate $\psi(z)$, whereas many algorithms do not need it.

x,reg,pot,tab1,huber

**Example 2.7.3** *If we choose $K = 1$, $t_0 = 0$, $t_1 = \delta$, $t_2 = \infty$, $d_0 = 0$, $d_1 = \delta$, $c_0 = 1$, $c_1 = 0$, then (2.7.23) corresponds to the Huber function (2.7.3). On the other hand, if we set $c_1 = 1$ then we get the ordinary parabola $\psi(z) = |z|^2/2$. So the choice of $c_K$ affects the properties of $\psi(z)$ for large $|z|$.*

We can again solve the **shrinkage problem** (2.7.20) exactly for the tabulated model (2.7.23). If $t_k \le z < t_{k+1}$ then using (2.7.21):

$$c = z + \beta\,(d_k + (z - t_k)c_k)$$

so the shrinkage rule is again a piecewise linear function illustrated in Fig. 2.7.7:

$$\hat{z}(c) = \operatorname{sgn}(c)\,\frac{|c| - \beta d_k + \beta t_k c_k}{1 + \beta c_k} = \operatorname{sgn}(c)\left(\frac{|c| - b_k}{1 + \beta c_k} + t_k\right), \tag{2.7.27}$$

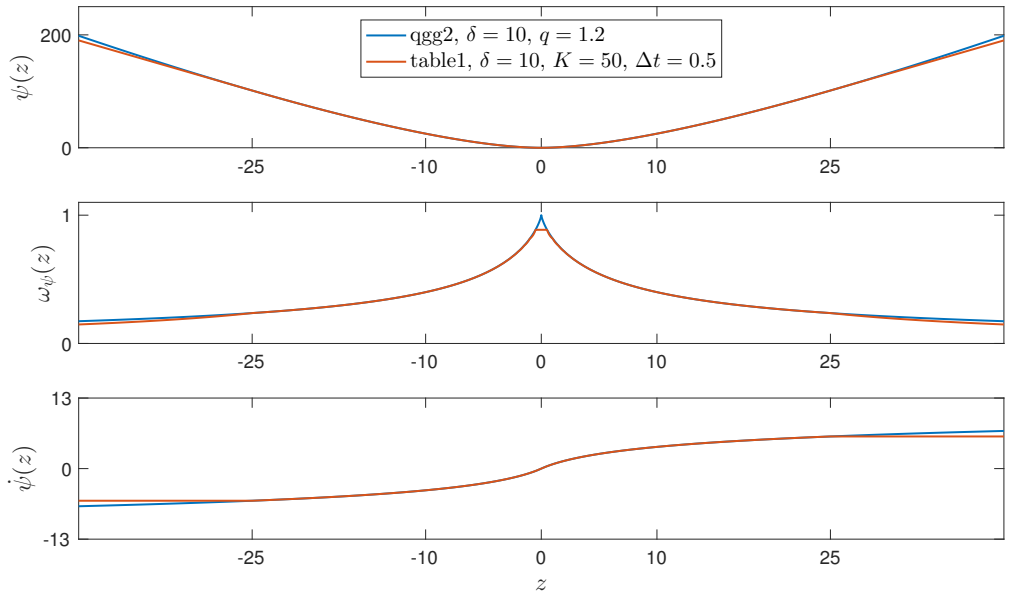
e,reg,pot,tab,shrink,th,1

where $b_k = t_k + \beta d_k$. This solution is correct when $t_k \le |\hat{z}| < t_{k+1}$ or equivalently when $b_k \le |c| < b_{k+1}$. In the usual case when the $t_k$ and $d_k$ values are monotone nondecreasing (*e.g.*, when $\psi$ is convex) then the intervals are non-overlapping and there is a unique solution.

x,reg,pot,tab

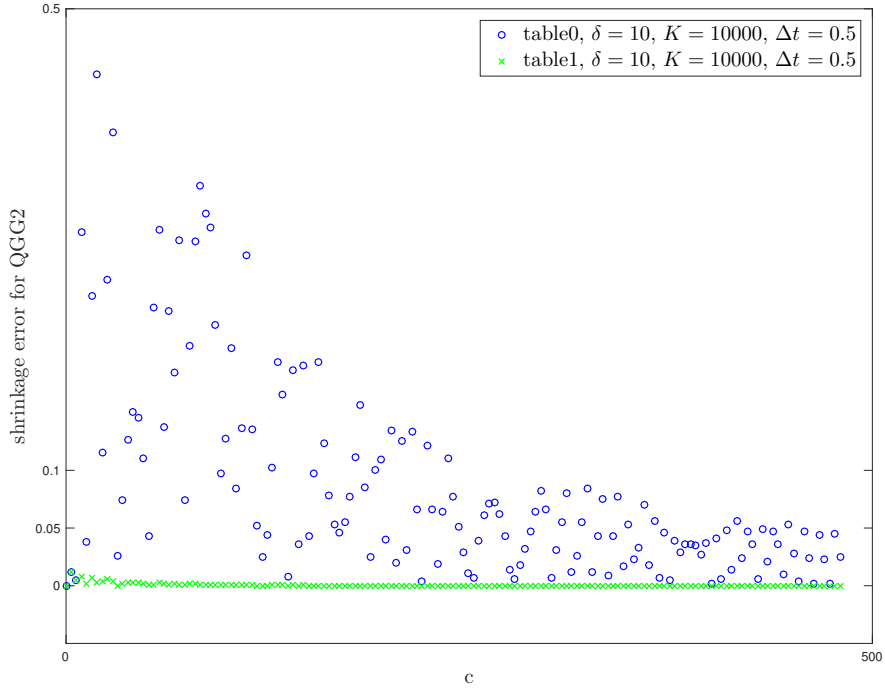
**Example 2.7.4** *Fig. 2.7.8 compares the QGG potential function (2.7.4) with $p = 2$ and $q = 1.2$ and $\delta = 10$ to the tabulated approximation (2.7.23) with $K = 50$ and $\Delta = 0.5$. For large $|z|$, the approximation rises linearly whereas QGG rises as $|z|^q$ so an accurate match requires that $K\Delta$ be sufficiently large.*

*The shrinkage function $\hat{z}(c)$ looks similar to that shown in Fig. 2.7.6 when viewed over a large range of $c$ values. Fig. 2.7.9 shows the error between the true shrinkage function for QGG2 versus the approximation from the tabulated versions (2.7.22) and (2.7.27). The linear interpolation method yields much lower errors for the same $K$ and $\Delta$.*




fig_reg_pot_table1_qgg2

Figure 2.7.8: QGG potential function for $p = 2$ and $q = 1.2$ and tabulated approximation.




fig_reg_pot_table01_qgg2_shrinker

Figure 2.7.9: Shrinkage function $\hat{t}(z)$ errors for tabulated versions versus truth for QGG2 with $q = 1.2$ and $\delta = 10$.

#### 2.7.8.3 Alternative tabulation methods

It might be tempting to define the weighting function to be piecewise constant, but that seemingly simpler approach leads to a convex potential function only in the degenerate case where $\omega_\psi$ is a constant.

To elaborate on this, because $\dot\psi(z) = z\,\omega_\psi(z)$ it follows that $\ddot\psi(z) = z\,\dot\omega_\psi(z) + \omega_\psi(z)$ so if we want $\ddot\psi(z) \geq 0$ as a sufficient condition for convexity, then we need

$$\dot\omega_\psi(z) \geq \frac{-\omega_\psi(z)}{z}.$$

In other words, we need $\omega_\psi(z)$ not to decrease too rapidly. In particular this condition prohibits a step decrease in $\omega_\psi$.

Another option would be to tabulate $\psi(t_k)$ using linear interpolation, but this approach cannot provide strictly convex potential functions.

Yet another approach would be to define $\omega_\psi$ piecewise using the simple ratios of the generalized Fair potential weighting functions (2.7.5). This approach might require fewer sample points for approximating some $\psi$ cases.

### 2.7.9 Summary

Clearly there are numerous possibilities for $\psi$. A variety of other non-convex potentials have also been studied, *e.g.*, [176–180]. The best choice can depend greatly on the image properties in a given application.

## 2.8 Multiple-channel regularization (s,reg,multi)

s,reg,multi

Most of the regularization methods described here are for a single "grayscale" image. There are a variety of imaging problems that involve multiple "channels" of images, such as dual-energy X-ray CT imaging, color photographs, polarimetric imaging [181, 182], PET/CT scanning, hyperspectral imaging, and dual-isotope SPECT imaging [183, 184].

In most of these applications, it is plausible that many of the edges between object regions will appear in more than one channel. Apply conventional edge-preserving regularizers to each channel independently would ignore the edge correspondences between channels. Some regularization methods have been proposed to account for such correlations, *e.g.*, Farsiu et al. [39] used regularization that encourages similar edge orientation in different color channels using cross products related to the angle of edge orientation. Weisenseel et al. [185] used a PDE-based approach to estimate a common boundary (edge) field for multiple images. This section summarizes some of the options for multiple-channel regularization.

### 2.8.1 Conventional channel-separable regularization

If $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M$ denote candidate vectors for the $M$ image channels, conventional regularization would be

$$\mathsf{R}(\boldsymbol{x}) = \sum_{m=1}^{M} \mathsf{R}_0(\boldsymbol{x}_m),$$

where $\mathsf{R}_0(\boldsymbol{x}_m)$ denotes a "conventional" regularizer for a single image, *e.g.*, (2.3.1). This approach ignores any correlation between images, so it provides a baseline for comparing alternate methods.

### 2.8.2 Convex multiple-channel regularization

One alternative is modify the arguments of the potential functions in (2.3.1) so that if an edge is present in one channel, the regularization is relaxed for the other channels. The following approach provides a convex regularizer and has been investigated in [182, 184]:

$$\mathsf{R}(\boldsymbol{x}) = \sum_{k=1}^{K} \psi\left( \sqrt{\sum_{m=1}^{M} \left| \frac{[\boldsymbol{C}\boldsymbol{x}_m]_k}{\delta_{mk}} \right|^2} \right), \tag{2.8.1}$$

e,reg,multi,convex

where $\psi$ is a convex edge-preserving potential function, such as the **hyperbola** (2.4.5). Although this modified regularizer is not quite of the form (2.3.1), one can develop efficient optimization methods for it, *e.g.*, Problem 12.6.

A drawback of (2.8.1) is that it can be challenging to control the **spatial resolution properties** of the different channels, particularly when the statistics of the corresponding data terms differ or when different values of the parameters $\delta_{mk}$ are needed for each channel. (See Chapter 22.)

### 2.8.3 Rank-based multiple-channel regularization

Consider patches around a single spatial location extracted from $M$ images in a multiple-channel setting. If the edges in those patches have similar locations, then the corresponding Jacobian matrix is likely to have low rank, *i.e.*, rank less than $M$. Specifically, let $\boldsymbol{C}_1, \dots, \boldsymbol{C}_L$ denote the finite difference matrices in $L$ different directions, *e.g.*, as defined in §2.14.3. Then the **total nuclear variation** (**TNV**) defined in [186, 187] is given by the following semi-norm:

$$\mathsf{R}_{\mathrm{TNV}}(\boldsymbol{x}) = \sum_{j=1}^{n_\mathrm{p}} \left\| \begin{bmatrix} [\boldsymbol{C}_1 \boldsymbol{x}_1]_j & \dots & [\boldsymbol{C}_L \boldsymbol{x}_1]_j \\ \vdots & \vdots & \vdots \\ [\boldsymbol{C}_1 \boldsymbol{x}_M]_j & \dots & [\boldsymbol{C}_L \boldsymbol{x}_M]_j \end{bmatrix} \right\|_*, \tag{2.8.2}$$

e,reg,multi,tnv

where $\|\cdot\|_*$ denotes the **nuclear norm** (sum of singular values) of a matrix. Results for dual-energy X-ray CT with this regularizer are encouraging [187].

For other **vector TV** (**VTV**) definitions, see [188–193].

### 2.8.4 Line-site based multiple-channel regularization

If we let $\boldsymbol{l}$ denote a common set of boundaries, *e.g.*, line sites (*cf.* §1.12.1) an alternative is

$$\mathsf{R}(\boldsymbol{x}, \boldsymbol{l}) = U(\boldsymbol{l}) + \sum_{m=1}^{M} \mathsf{R}_0(\boldsymbol{x}_m, \boldsymbol{l}),$$

where $\mathsf{R}_0(\boldsymbol{x}_m, \boldsymbol{l})$ was defined in (1.12.2) and $U(\boldsymbol{l})$ in (1.12.4), for example. This approach is a discretized version of the PDE approach in [185]. A drawback of this approach is that usually $\mathsf{R}(\boldsymbol{x}, \boldsymbol{l})$ is not convex as function of both arguments.

For example, consider the regularizer

$$\mathsf{R}(\boldsymbol{x}, \boldsymbol{l}) = \sum_{k=1}^{K} \left[ \left( \sum_{m=1}^{M} \frac{1}{2} |[\boldsymbol{C}\boldsymbol{x}_m]_k|^2 \right) l_k + u(l_k) \right] \tag{2.8.3}$$

e,reg,multi,line

where for $l \in (0, 1]$:

$$u(l) = \frac{(1-l)^2}{2l}. \tag{2.8.4}$$

e,reg,multi,line,ul

One can show that for this choice, minimizing over $l_k$ for a given estimate $\{\boldsymbol{x}_m^{(n)}\}$ yields

$$l_k = \left( 1 + \sum_{m=1}^{M} |[\boldsymbol{C}\boldsymbol{x}_m^{(n)}]_k|^2 \right)^{-1/2}.$$

For insight into the choice (2.8.4) and generalizations thereof, see Problem 2.19.

### 2.8.5 Sparsity-based multiple-channel regularization

In the language of compressed sensing, another option is to look for regularization that encourages "common sparsity" between the $\boldsymbol{x}_m$ images, *e.g.*, [194–203]. Consider the case of two images ($M = 2$). The traditional "ideal" sparsity regularizer would be $\|\boldsymbol{x}_1\|_0 + \|\boldsymbol{x}_2\|_0$, which fails to capture joint sparsity. Instead, we might want to use

$$\mathsf{R}(\boldsymbol{x}) = \sum_{j=1}^{n_\mathrm{p}} h(x_{1j}, x_{2j}),$$

where $f$ satisfies the following axioms (all of which generalize readily to $M > 2$):

$$\begin{aligned}
&h(0,0) = 0 \\
&h(a,b) = h(b,a) \text{ (symmetry)} \\
&h(a,b) \geq 0 \\
&h(a,0) > 0 \text{ if } a \neq 0 \\
&h(a,0) < h(a,b) \text{ if } b \neq 0 \text{ (monotonicity)} \\
&h(a,b) < h(a,0) + h(0,b) \text{ if } a, b \neq 0 \text{ (commonality)}.
\end{aligned}$$

A particularly popular example that satisfies these conditions is the convex function

$$h(a,b) = \sqrt{a^2 + b^2}$$

which is akin to (2.8.1). This choice is called the mixed $\ell_{1,2}$ norm of the matrix $[\boldsymbol{x}_1 \ \boldsymbol{x}_2]$ [202].

Another approach is to write [196, 204, 205]: $\boldsymbol{x}_1 = \boldsymbol{z}_c + \boldsymbol{z}_1$, $\boldsymbol{x}_2 = \boldsymbol{z}_c + \boldsymbol{z}_2$ and penalize $\|\boldsymbol{z}_c\|_0 + \|\boldsymbol{z}_1\|_0 + \|\boldsymbol{z}_2\|_0$.

s,reg,complex

## 2.9 Regularization of complex-valued images (s,reg,complex)

When regularizing complex images, there are several options depending on the application. All of the potential functions in §2.7 are defined for complex-valued arguments, so the general regularizer form (2.3.1) is applicable. For the usual case of finite differences, a typical term in the regularizer has the form $\psi(|x_j - x_k|)$, *i.e.*, is a function of the complex difference between neighboring pixel values. The main subtlety here is computing $\nabla\Psi$ properly; see Appendix 28.

In some applications it is beneficial to regularize the real and imaginary parts separately [206–209], *e.g.*,

e,reg,complex,r,i

$$\mathsf{R}(\boldsymbol{x}) = \beta_1 \mathsf{R}_1(\text{real}\{\boldsymbol{x}\}) + \beta_2 \mathsf{R}_1(\text{imag}\{\boldsymbol{x}\}) . \tag{2.9.1}$$

In other applications, it is beneficial to regularize the magnitude and phase separately [210–214], *e.g.*,

$$\mathsf{R}(\boldsymbol{x}) = \beta_1 \mathsf{R}_1(|\boldsymbol{x}|) + \beta_2 \mathsf{R}_1(\angle \boldsymbol{x}) .$$

Often it is reasonable to assume the magnitude and the real and imaginary parts are all piecewise smooth, for which edge-preserving regularization is appropriate. In some applications the phase is smooth, and in other cases it is sparse or piecewise smooth. In any case, when regularizing the phase of a complex image using finite differences, it may be more appropriate to penalize the differences between values raised to a complex exponential to avoid phase wrap issues [212, 214] *e.g.*, for first-order finite differences:

$$\left| \mathrm{e}^{\imath \angle x_j} - \mathrm{e}^{\imath \angle x_k} \right| .$$

As noted in [212]:

$$|a - b| = \left| |a| \, \mathrm{e}^{\imath \angle a} - |b| \, \mathrm{e}^{\imath \angle b} \right| = \left| |a| - |b| \right|^2 + 2 \, |a| \, |b| \, (1 - \cos(\angle a - \angle b)) .$$

This type of weighted $1 - \cos$ term for the phase is helpful in areas where the magnitude approaches zero, and hence the phase is not well defined [215].

s,reg,side

## 2.10 Regularization with side information (s,reg,side)

In some imaging applications, there one has available a prior image $\bar{\boldsymbol{x}}$ that is expected to be related in some way to the image $\boldsymbol{x}$ being reconstructed. There are many methods for reconstructing an image $\hat{\boldsymbol{x}}$ using both the measurements $\boldsymbol{y}$ and the **side information** present in the prior image $\bar{\boldsymbol{x}}$.

A simple option is to initialize the iterative algorithm for finding $\hat{\boldsymbol{x}}$ with the prior image $\bar{\boldsymbol{x}}$ [216, 217]. This may be reasonable when the data is highly under-sampled. In such problems, the solution typically is under-determined without regularization, and initializing with $\bar{\boldsymbol{x}}$ can steer $\hat{\boldsymbol{x}}$ towards a solution near $\bar{\boldsymbol{x}}$.

Another option is to use a regularizer that encourages the estimate to agree with the prior image, such as [216, 218]:

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \mathsf{L}(\boldsymbol{x}) + \beta \, \|\boldsymbol{x} - \bar{\boldsymbol{x}}\|^2 .$$

In multimodality systems like PET/CT scanners, the grayscale values of the PET and CT images are entirely different, but some of the edges between regions should be in similar locations. Therefore, another widely studied option is to extract the region boundaries from $\bar{\boldsymbol{x}}$, and then use a modified regularizer akin to (1.10.17) that relaxes the regularization between neighboring voxels that lies in two different regions. Early work in this area used **line site** models (*cf.* §1.12.1) [219–230]. Modified regularizers have also been investigated widely [231–241]. Some such approaches allow for mixtures [242–244]. In some cases the region boundaries are estimated jointly with the reconstruction [185, 245, 246].

Another option is to use **image segmentation** to identify regions in the prior image $\bar{\boldsymbol{x}}$, and then assume the corresponding regions in $\hat{\boldsymbol{x}}$ are homogeneous [247–250].

One way to avoid the need for finding edges or segmenting regions in the prior image $\bar{\boldsymbol{x}}$ is to use a regularizer based on a information theoretic principles such as **cross entropy** [251, 252], **mutual information** [253, 254] and **joint entropy** [255, 256].

Many post-processing methods have been proposed [257]. Sufficiently accurate boundary information can improve image detection tasks [258].

Multi-modality systems are of increasing interest in many imaging areas, so reconstruction methods for such problems will remain an active research area.

s,reg,values

## 2.11 Regularization using specific voxel values (s,reg,values)

The primary focus of this chapter is on regularizers that involve differences between neighboring voxels. There have also been methods proposed that penalize the pixel values themselves, such as

$$\mathsf{R}(\boldsymbol{x}) = \sum_{j=1}^{n_\mathrm{p}} \psi(x_j - \mu_j),$$

for some **prior image** $\boldsymbol{\mu}_{\boldsymbol{x}}$. As discussed in §1.7.3.3, often the prior image $\boldsymbol{\mu}_{\boldsymbol{x}}$ does not add useful information to the reconstructed image.

However, in some applications we know (or expect) that the pixel values $x_j$ will tend to cluster around a small number of mean values. For example, in X-ray CT imaging, we expect most voxel values to be near the typical values of air, lung, water (soft tissue), or bone. In other words, we expect a histogram of the image to have several distinct peaks. A typical statistical model for such a histogram is a gaussian mixture model:

$$\mathsf{p}(x) = \sum_{k=1}^{K} \mathsf{p}_k \frac{1}{\sqrt{2\pi}\sigma_k} \mathrm{e}^{-(x-\mu_k)^2/(2\sigma_k^2)},$$

where $\mathsf{p}_k \geq 0$ and $\sum_{k=1}^{K} \mathsf{p}_k = 1$. One could use the negative logarithm of this prior distribution as a regularizer [259–261]:

$$\mathsf{R}(\boldsymbol{x}) = -\sum_{j=1}^{n_\mathrm{p}} \log \mathsf{p}(x_j) = -\sum_{j=1}^{n_\mathrm{p}} \log \left( \sum_{k=1}^{K} \mathsf{p}_k \frac{1}{\sqrt{2\pi}\sigma_k} \mathrm{e}^{-(x_j-\mu_k)^2/(2\sigma_k^2)} \right).$$

The summation within the logarithm is slightly inconvenient for optimization. An alternative is to use a piecewise quadratic regularizer of the following form [262, 263]:

$$\mathsf{R}(\boldsymbol{x}) = \sum_{j=1}^{n_\mathrm{p}} \psi(x_j), \quad \psi(x) = \sum_{k=1}^{K} \frac{(x-\mu_k)^2}{2\sigma_k^2} \mathbb{I}_{\{a_k < x \leq b_k\}}$$

where $a_1 = -\infty$, $b_1 = (\mu_1+\mu_2)/2$, $a_k = (\mu_{k-1}+\mu_k)/2$, $b_k = (\mu_k+\mu_{k+1})/2$, $a_K = (\mu_{K-1}+\mu_K)/2$, $b_K = \infty$, for $k = 2, \ldots, K-1$. This regularizer corresponds to an approximation of the negative logarithm of a gaussian mixture. The approximation is most accurate when the mixture components are well separated. (See also Problem 2.18.) Both options for $\mathsf{R}(\boldsymbol{x})$ are highly nonconvex functions so local minimizers are a significant challenge for optimization.

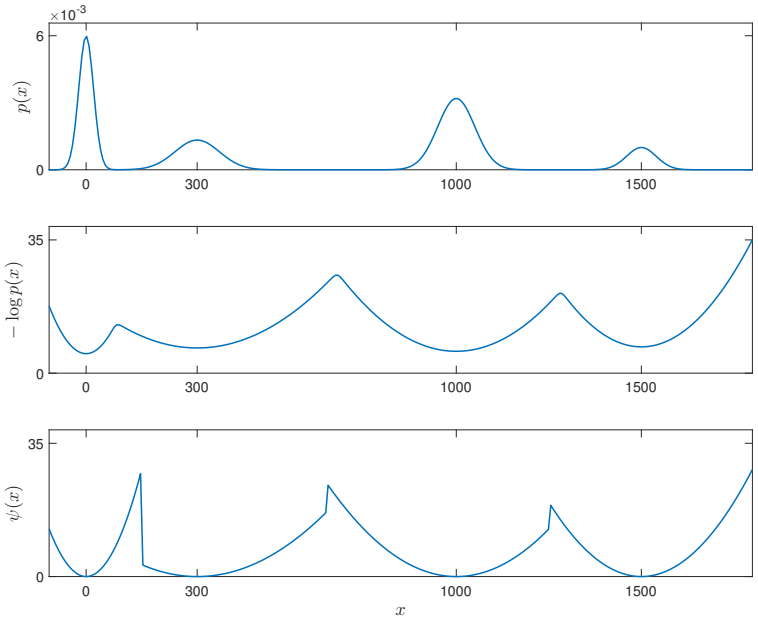



Figure 2.11.1: Top: density of gaussian mixture model. Model: its negative logarithm $-\log \mathsf{p}(x)$. Bottom: piecewise quadratic regularizer that approximates the negative logarithm. fig_reg_gauss_mix

Fig. 2.11.1 illustrates the functions described above.

## 2.12 Regularization using non-local means (s,reg,nlm)

s,reg,nlm

Buades [264] proposed an effective method for **image denoising** using **non-local means**. For the denoising model $\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{\varepsilon}$, a nonlocal means estimator has the form

$$\hat{x}_j = \hat{x}_j(\boldsymbol{y}) = [\mathrm{NLM}(\boldsymbol{y})]_j = \frac{\sum_{k \in \mathcal{N}_j} w_{k,j}(\boldsymbol{y}) y_k}{\sum_{k \in \mathcal{N}_j} w_{k,j}(\boldsymbol{y})},$$

where $\mathcal{N}_j$ is a neighborhood of the $j$th pixel and $w_{k,j}(\boldsymbol{y})$ are data-adaptive weights. (If the weights are independent of $\boldsymbol{y}$ then this simplifies to ordinary linear filtering.) The weights used in the nonlocal means method have the form

$$w_{k,j}(\boldsymbol{y}) = \mathrm{e}^{-\|\boldsymbol{R}_k \boldsymbol{y} - \boldsymbol{R}_j \boldsymbol{y}\|^2 / c} f(\|\vec{n}_j - \vec{n}_k\|),$$

where $\vec{n}_j$ denotes the spatial coordinates of the $j$th pixel, $\boldsymbol{R}_k$ is a linear operator that extracts a local patch of values around the $j$th pixel, and typically $f(\cdot)$ is a decreasing function.

Let $\mathrm{NLM}(\boldsymbol{y})$ denote the non-local means image denoising function. This function can be used as a regularizer for inverse problems as follows [265, 266]

e,reg,nlm

$$\mathsf{R}(\boldsymbol{x}) = \|\boldsymbol{x} - \mathrm{NLM}(\boldsymbol{x})\|, \tag{2.12.1}$$

for some norm. See [265] for a steepest descent minimization method. This topic is evolving rapidly [266–271].

## 2.13 Summary (s,reg,summ)

s,reg,summ

This chapter and Chapter 1 have described numerous possible methods for regularization. More methods continue to be developed; see Chapter 10 for regularizers based on **dictionary learning**. No single method is universally optimal, and the results depend on the properties of the object and the imaging system. Empirical investigation is required to evaluate various options; the *Michigan Image Reconstruction Toolbox* can facilitate such explorations.

## 2.14 Appendix: Implementing finite differences: $\boldsymbol{Cx}$ (s,reg,irt,Cx)

s,reg,irt,Cx

This section describes methods for implementing the matrix-vector multiplication operation $\boldsymbol{d} = \boldsymbol{C}\boldsymbol{x}$ and the transpose operation $\boldsymbol{z} = \boldsymbol{C}'\boldsymbol{d}$ corresponding to finite differences. These operations are useful for some implementations of regularization, as described in §1.8.1 and §1.10. See §2.3 for alternative implementations that often have advantages.

### 2.14.1 Implementing 1D finite differences (s,reg,irt,c1)

s,reg,irt,c1

We begin with the case of 1D signals, primarily for illustration. For 1D signals $\boldsymbol{x}$ of length $N$, we focus here on the following $N \times N$ 1st-order finite differencing matrix:

$$\boldsymbol{C} \triangleq \begin{bmatrix} 0 & 0 & 0 & 0 & \dots & 0 \\ -1 & 1 & 0 & 0 & \dots & 0 \\ 0 & -1 & 1 & 0 & \dots & 0 \\ & & \ddots & \ddots & & \\ 0 & \dots & 0 & -1 & 1 & 0 \\ 0 & \dots & 0 & 0 & -1 & 1 \end{bmatrix}, \implies \boldsymbol{d} = \boldsymbol{C}\boldsymbol{x} = \begin{bmatrix} 0 \\ x_2 - x_1 \\ \vdots \\ x_N - x_{N-1} \end{bmatrix}. \tag{2.14.1}$$

e,reg,C,N,N

For periodic boundary conditions, one replaces the first row with $[1\ 0\ \dots\ 0\ -1]$. Otherwise the first row of $\boldsymbol{C}$ is superfluous, but harmless because we always use potential functions for which $\psi(0) = 0$. Using a square matrix here can simplify implementation, particularly in higher dimensions.

MIRT *The function* `Cdiff1` *generates* $\boldsymbol{C}$ *objects that can perform* $\boldsymbol{d} = \boldsymbol{C}\boldsymbol{x}$ *using several different methods, as described below.*

#### 2.14.1.1 loop

s,reg,irt,Cx,ld,loop

In most compiled languages, the natural way to implement $\boldsymbol{d} = \boldsymbol{C}\boldsymbol{x}$ is to use a **loop** as follows.

```
for n=2:N
        d(n) = x(n) - x(n-1);
end
d(1) = x(1) - x(N); % for periodic boundary conditions
d(1) = 0; % otherwise
```

MIRT `Cdiff1` *with* `'mex'` *option uses such a loop, compiled in ANSI C, which is quite fast.*

MIRT `Cdiff1` *with* `'for1'` *option uses such a loop, but is quite slow because* MATLAB *is an interpreted language.*

#### 2.14.1.2 matrix

s,reg,irt,Cx,ld,matrix

One can create $\boldsymbol{C}$ directly as a matrix as follows:

```
C = diag([0 ones(1,N-1)]) + diag(-ones(N-1,1), -1);
```

However, this approach fails to exploit the sparsity of $\boldsymbol{C}$ and computing $\boldsymbol{d} = \boldsymbol{C}\boldsymbol{x}$ would use $O(N^2)$ operations. Its only practical use is didactic.

s,reg,irt,Cx,ld,sparse

#### 2.14.1.3 sparse

The 1D matrix $\boldsymbol{C}$ given in (2.14.1) is a **sparse matrix**, because most of its elements are zeros. For 1st-order differences, each row of $\boldsymbol{C}$ has at most two nonzero elements (out of $N$). A natural way to store $\boldsymbol{C}$ is as a **sparse matrix**, meaning a data structure that stores only the nonzero values and the locations of those values in a list. A concise description of $\boldsymbol{C}$ is the following enumeration of its $2(N-1)$ nonzero entries $c_{kj}$.

| row | $k$ | 2 | 3 | ... | $N$ | 2 | 3 | ... | $N$ |
|---|---|---|---|---|---|---|---|---|---|
| column | $j$ | 2 | 3 | ... | $N$ | 1 | 2 | ... | $N-1$ |
| element | $c_{kj}$ | 1 | 1 | ... | 1 | $-1$ | $-1$ | ... | $-1$ |

One can generate such a matrix using MATLAB's `sparse` command as follows.

```
k = [2:N 2:N];
j = [2:N 1:(N-1)];
c = [ones(1,N-1), -ones(1,N-1)];
C = sparse(k, j, c, N, N);
```

Alternatively one can use:

```
C = sparse(2:N, 2:N, ones(1,N-1), N, N) ...
 - sparse(2:N, 1:(N-1), ones(1,N-1), N, N);
```

or, for the most delightfully concise of all:

```
C = diff(speye(N));
```

For periodic boundary conditions in 1D, one can use the following concise command

```
C = speye(N) - circshift(speye(N), 1);
```

The sparse matrix form can be convenient for modest size experiments, but is inefficient computationally for large problems, particularly in higher dimensions, because a general sparse matrix data structure does not exploit the regularity of the pattern of nonzero elements in $\boldsymbol{C}$ and the fact that those nonzero elements are all $\pm 1$. Computing finite differences directly (with a compiled loop) is faster than using sparse matrix-vector multiplication.

MIRT `Cdiff1` *with* `'spmat'` *option generates this sparse matrix for non-periodic boundary conditions.*

t,Cx,ld,array indexing

#### 2.14.1.4 array indexing

Another option in MATLAB is to use array index operations to compute 1D first-order finite differences:

```
d = [0; x(2:end)-x(1:end-1)];
```

this indexing approach is portable but slow for large arrays.

If $\boldsymbol{C}$ is implemented as a matrix, then one can conveniently multiply $\boldsymbol{C}$ by several vectors stored in an array with a single multiplication operation, *e.g.*, `C * [x1 x2];` which appears similar to the mathematical expression $\boldsymbol{C}[\boldsymbol{x}_1 \ \boldsymbol{x}_2]$. The simple indexing command above works only for a single vector as written. To enable it to work with multiple column vectors stored in an array, we rewrite it as follows:

```
d = [zeros(1,size(x,2)); x(2:end,:)-x(1:end-1,:)];
```

MIRT `Cdiff1` *with* `'ind'` *option implements this approach.*

ular shift (circshift)

#### 2.14.1.5 circular shift (circshift)

MATLAB's `circshift` command offers another fast approach:

```
d = x - circshift(x, 1);
```

clearly this version uses periodic boundary conditions. This concise code also works when `x` is an array. This usually is the fastest non-`mex` approach.

MIRT `Cdiff1` *with* `'circshift'` *option implements this approach.*

,irt,Cx,ld,convolution

#### 2.14.1.6 convolution

Another approach is to use MATLAB's `convn` command:

```
d = convn(x, [0 1 -1]', 'same'); d(1,:) = 0;
```

For periodic boundary conditions, replace the last part with `d(1,:) = x(1,:) - x(end,:);`

MIRT `Cdiff1` *with* `'convn'` *option implements this approach.*

s,reg,irt,Cx,ld,filter

#### 2.14.1.7 filter

Another approach is to use MATLAB's `imfilter` command, which allows periodic boundary conditions easily:

```
d = imfilter(x, [0 1 -1]', 'circular', 'conv', 'same');
```

MIRT `Cdiff1` *with* `'imfilter'` *option implements this approach; however, it requires the Image Processing Toolbox.*

s,reg,irt,Cx,ld,diff

#### 2.14.1.8 diff

A final option is to compute finite differences using MATLAB's `diff` command:

```
d = [zeros(1,size(x,2)); diff(x, 1)];
```

MIRT `Cdiff1` *with* `'diff'` *option implements this approach.*

MIRT *There are many feasible approaches, and which one is fastest depends on computer hardware, image size, etc. The* `Cdiff1_tune` *command tries all of them and finds the fastest for a given image size.*

s,reg,irt,Cx',ld

### 2.14.2 Implementing $C'd$ in 1D

We also need the transpose (adjoint) operation:

$$\boldsymbol{C}' = \begin{bmatrix} 0 & -1 & 0 & 0 & \dots & 0 \\ 0 & 1 & -1 & 0 & \dots & 0 \\ & & \ddots & \ddots & & \\ 0 & \dots & 0 & 1 & -1 & 0 \\ 0 & \dots & 0 & 0 & 1 & -1 \\ 0 & \dots & 0 & 0 & 0 & 1 \end{bmatrix} \implies \boldsymbol{z} = \boldsymbol{C}'\boldsymbol{d} = \begin{bmatrix} -d_2 \\ d_2 - d_3 \\ \vdots \\ d_{N-1} - d_N \\ d_N \end{bmatrix}. \tag{2.14.2}$$

e,reg,C',N,N

The loop version is simple and if $\boldsymbol{C}$ is a matrix (full or sparse) then $\boldsymbol{C}'$ is built in to MATLAB.
For the `circshift` approach (with its periodic boundary conditions), the adjoint is

```
z = d - circshift(d, -1);
```

For the `convn` approach we must reverse the impulse response and handle the end conditions carefully:

```
tmp = d; tmp(1,:)  = 0; z = convn(tmp, [-1 1 0]', 'same');
```

For the `imfilter` approach with periodic boundary conditions, we simply reverse the impulse response:

```
z = imfilter(d, [-1 1 0]', 'circular', 'conv', 'same');
```

The `index` approach is based on (2.14.2):

```
z = zeros(1,size(d,2)); z = [z; d(2:end,:)]  - [d(2:end,:); z];
```

Finally, the `diff` approach also requires care with boundary conditions:

```
tmp = d; tmp([1 end+1],:)  = 0; z = -diff(tmp,1);
```

s,reg,irt,c2

### 2.14.3 Implementing 2D finite differences (s,reg,irt,c2)

As described in §1.10, regularizing 2D imaging problems with finite differences requires computing $\boldsymbol{d} = \boldsymbol{C}\boldsymbol{x}$ where in 2D (and higher), typically $\boldsymbol{C}$ is a "stack" of multiple finite differencing matrices. For the typical case of horizontal and vertical first-order finite differences described in (1.10.8), $\boldsymbol{C} = \begin{bmatrix} \boldsymbol{C}_1 \\ \boldsymbol{C}_2 \end{bmatrix}$. We focus in this section on this concrete case for illustration, but the ideas generalize to additional directions (*e.g.*, diagonals). Computing $\boldsymbol{d} = \boldsymbol{C}\boldsymbol{x}$ in 2D involves (at least) two separate matrix multiplications: $\boldsymbol{d}_1 = \boldsymbol{C}_1\boldsymbol{x}$ and $\boldsymbol{d}_2 = \boldsymbol{C}_2\boldsymbol{x}$, corresponding to horizontal and vertical finite differences respectively. (See §2.3 for generalizations.)

MIRT *The function* `Cdiff1` *generates such* $\boldsymbol{C}_l$ *objects that, when multiplied by* $\boldsymbol{x}$*, compute finite differences by any of several methods, described below.*

MIRT *The function* `Cdiffs` *represents* $\boldsymbol{C}$ *by stacking up objects generated by* `Cdiff1`*. (See (2.3.10).)*

Mathematically, we want to compute

$$d_l[m,n] = f[m,n] - f[m - m_l, n - n_l], \quad l = 1, 2$$

where $(m_1, n_1) = (1, 0)$ and $(m_2, n_2) = (0, 1)$.

s,reg,irt,c2,loop

#### 2.14.3.1 loop

If vector $\boldsymbol{x}$ corresponds to a 2D image $f[m,n]$ of size $M \times N$, then $\boldsymbol{d}_l = \boldsymbol{C}_l \boldsymbol{x}$ corresponds to the following loop.

```
for m=(1+ml):M
        for n=(1+nl):N
                d(m,n) = f(m,n) - f(m-ml, n-nl)
        end
end
```

Because 2D arrays are usually stored simply as one long vector, an alternative loop form is the following. This code assumes that $m$ varies fastest.

```
offset = ml + nl * M;
for j=(1+offset):(M*N)
        d(j) = f(j) - f(j - offset);
end
```

MIRT `Cdiff1` *with* `'for1'` *option uses this simpler single loop form; the* `'mex'` *option provides the same loop in compiled ANSI C which is usually the fastest option. This loop computes some extra finite differences that are usually unwanted and must be set to zero (by multiplying by 0) separately.* `Rweights` *provides a vector of with zeros in the appropriate locations.*

,irt,c2,array indexing

#### 2.14.3.2 array indexing

Using array indexing somewhat "hides" the loop.

```
d1 = [zeros(1,size(x,2)); f(2:end,:)  - f(1:end-1,:)];
d2 = [zeros(size(x,1),1), f(:,2:end) - f(:,1:end-1)];
```

s,reg,irt,c2,sparse

#### 2.14.3.3 sparse

Because $\boldsymbol{C}_l$ is **sparse**, one can store it even for quite large problem sizes. As shown in Problem 1.14, $\boldsymbol{C}_1 = \boldsymbol{I}_N \otimes \boldsymbol{D}_M$ and $\boldsymbol{C}_2 = \boldsymbol{D}_N \otimes \boldsymbol{I}_M$, which can be formed using the following simple commands for periodic boundary conditions:

```
C1 = kron(speye(N), speye(M) - circshift(speye(M), [1 0]));
C2 = kron(speye(N) - circshift(speye(N), [1 0]), speye(M));
```

For non-periodic boundary conditions, combine §2.14.1.3 with `kron`. For example, the following commands are particularly concise:

```
C1 = kron(speye(N), diff(speye(M)));
C2 = kron(diff(speye(N)), speye(M));
```

s,reg,irt,c2,convn

#### 2.14.3.4 convn

One can use convolution with `convn`

```
d1 = convn([0 1 -1]', f);
d2 = convn([0 1 -1], f);
```

Replacing `convn` with `imfilter` enables periodic boundary conditions.

s,reg,irt,c2,circshift

#### 2.14.3.5 circshift

Finally, for periodic boundary conditions, a particulary simple option is to use `circshift`:

```
d1 = f - circshift(f, [1 0]);
d2 = f - circshift(f, [0 1]);
```

### 2.14.4 Adjoint (transpose) in 2D

Implementing the adjoint (transpose) operation $\boldsymbol{z} = \boldsymbol{C}'\boldsymbol{d} = \boldsymbol{C}_1'\boldsymbol{d}_1 + \boldsymbol{C}_2'\boldsymbol{d}_2$ in 2D is similiary straightforward by any of the above methods. One must use addition.

MIRT *The function* `Cdiffs` *in the Michigan Image Reconstruction Toolbox generates matrix-like objects that perform the operations* $\boldsymbol{C}\boldsymbol{x}$ *and* $\boldsymbol{C}'\boldsymbol{d}$ *using the convenient syntax* `C * x` *and* `C' * d`. *Although this syntax is suggestive of matrix-vector multiplication, and indeed the operation that occurs is linear, the internal calculations are performed by one of several methods depending on which options are selected. One of the options is to use a sparse matrix, but this choice is available primarily for testing and completeness; it is not the most efficient in compute time or memory. The fastest choice is the* `'mex'` *option that invokes a call to a compiled C subroutine called* `penalty_mex`. *This* **MEX** *file computes the required finite differences directly. Because compiled* **MEX** *files are not portable,* `Cdiff1` *reverts to using the* `circshift` *option when the* **MEX** *file is unavailable.*

## 2.15 Problems (s,reg,prob)

s,reg,prob

p,reg,l0,poisson

**Problem 2.1** *Often it is assumed that the constrained minimization problem*

$$\hat{\boldsymbol{x}}_k \triangleq \arg\min_{\boldsymbol{x} \geq \boldsymbol{0}} \mathsf{L}(\boldsymbol{x}) \text{ sub. to } \mathsf{R}(\boldsymbol{x}) \leq k \tag{2.15.1}$$

e,reg,l0,poisson,con

*is equivalent, for some choice of regularization parameter* $\beta$*, to the following regularized problem:*

$$\hat{\boldsymbol{x}}_\beta \triangleq \arg\min_{\boldsymbol{x} \geq \boldsymbol{0}} \mathsf{L}(\boldsymbol{x}) + \beta\, \mathsf{R}(\boldsymbol{x}) \,. \tag{2.15.2}$$

e,reg,l0,poisson,reg

*Consider the Poisson denoising problem where* $\boldsymbol{y} \sim \mathsf{Poisson}\{\boldsymbol{x} + \boldsymbol{r}\}$*, where* $\boldsymbol{r}$ *is a known nonnegative vector, with* ***counting measure*** *regularizer* $\mathsf{R}(\boldsymbol{x}) = \|\boldsymbol{x}\|_0$ *. Find* analytical *solutions to* $\hat{\boldsymbol{x}}_k$ *and* $\hat{\boldsymbol{x}}_\beta$ *above and determine if they are equal for some choices of* $\beta$ *and* $k$ *[272, 273].*

p,reg,deriv1

**Problem 2.2** *Find a matrix* $\boldsymbol{C}$ *such that when* $f(t) = \sum_{j=1}^{n_\mathrm{p}} x_j \,\mathrm{tri}(t-j)$*, we get equivalent values for the following continuous-space and discrete-space roughness penalty functions:*

$$\int \left|\dot{f}\right|^2 \mathrm{d}t = \|\boldsymbol{C}\boldsymbol{x}\|^2 \,.$$

p,reg,limit,pl

**Problem 2.3** §2.6 *examined the properties of the QPWLS estimator* $\hat{\boldsymbol{x}}_\beta$ *as* $\beta \to \infty$ *for the case of a WLS data fit term and quadratic regularization. Find sufficient conditions that generalize the conclusions of that section to the case of penalized-likelihood estimators of the form*

$$\hat{\boldsymbol{x}}_\beta = \arg\min_{\boldsymbol{x}} \sum_{i=1}^{n_\mathrm{d}} \mathsf{h}_i([\boldsymbol{A}\boldsymbol{x}]_i) + \beta \sum_{k=1}^{K} \psi_k([\boldsymbol{C}\boldsymbol{x}]_k) \,.$$

p,reg,limit,time

**Problem 2.4** *Extend* §2.6 *to the case of dynamic image reconstruction with temporal regularization:*

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \sum_{m=1}^{M} \left( \|\boldsymbol{y}_m - \boldsymbol{A}_m \boldsymbol{x}_m\|^2_{\boldsymbol{W}_m^{1/2}} + \beta \|\boldsymbol{C}_s \boldsymbol{x}_m\|^2 \right) + \zeta \|\boldsymbol{C}_t \boldsymbol{x}\|^2 = \left[\mathsf{F} + \beta \mathsf{R}_s + \zeta \mathsf{R}_t\right]^{-1} \boldsymbol{A}'\boldsymbol{W}\boldsymbol{y}$$

*where* $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_M)$*,* $\mathsf{F} = \boldsymbol{A}'\boldsymbol{W}\boldsymbol{A}$*,* $\boldsymbol{A} = \mathsf{diag}\{\boldsymbol{A}_m\}$*,* $\boldsymbol{W} = \mathsf{diag}\{\boldsymbol{W}_m\}$*,* $\mathsf{R}_s = \boldsymbol{I}_M \otimes \boldsymbol{C}_s'\boldsymbol{C}_s$*, and*

$$\boldsymbol{C}_t = \boldsymbol{C}_0 \otimes \boldsymbol{I}_N$$

*where* $\boldsymbol{x}_m \in \mathbb{R}^N$ *and* $\boldsymbol{C}_0$ *denotes the* $M-1 \times M$ *1st-order differencing matrix defined in (1.8.4) or one of its variants [274].*

p,reg,var,tps

**Problem 2.5** *Use 2D FT properties to prove that the thin-plate regularizer (2.4.2) is rotation invariant.*

p,reg,hyper,pwls1

**Problem 2.6** *Derive the MSE expressions (2.5.9) and (2.5.10) using (2.5.8). Then find* $\beta_{\mathrm{MSE}}$*.*

p,reg,hyper,rss

**Problem 2.7** *Prove the* RSS *equalities (2.5.13), (2.5.15), (2.5.16), (2.5.19), (2.5.20), and (2.5.21).*

p,reg,hyper,rss,dp,i

**Problem 2.8** *Modify Example 2.5.7 using (2.5.21) to determine* $\beta_{\mathrm{DP}}$ *in the orthogonal case where* $\mathsf{F} = \sigma^{-2}\boldsymbol{I}$ *and* $\mathsf{R} = \boldsymbol{I}$*.*

p,reg,hyper,edf

**Problem 2.9** *Analyze* $\beta_{\mathrm{REDF}}$ *under the usual circulant approximation for the case where one uses (2.5.28) to define* REDF.

p,reg,hyper,cv,i

**Problem 2.10** *Use (2.5.37) to determine* $\beta_{\mathrm{CV}}$ *in the orthogonal white-noise case where* $\mathsf{F} = \sigma^{-2}\boldsymbol{I} = \boldsymbol{W}$ *and* $\mathsf{R} = \boldsymbol{I}$.

p,reg,hyper,gcv,i

**Problem 2.11** *Use (2.5.41) to determine* $\beta_{\mathrm{GCV}}$ *in the orthogonal white-noise case where* $\mathsf{F} = \sigma^{-2}\boldsymbol{I} = \boldsymbol{W}$ *and* $\mathsf{R} = \boldsymbol{I}$.

p,reg,hyper,wc

**Problem 2.12** *Use (2.5.9) to describe how to determine the value of* $\beta$ *that minimizes the* worst case *MSE over all signals with* $\|\boldsymbol{x}\| \leq 1$. *This is a* ***min-max regularization parameter*** *selection method.*

p,reg,hyper,mat1

**Problem 2.13** *Choose an image* $\boldsymbol{x}_{\mathrm{true}}$ *and a shift-invariant blur* $b[m,n]$ *with circulant end conditions and create a noisy, blurry image* $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{\varepsilon}$. *Apply the image restoration method of Example 2.5.1 with quadratic regularization based on 1st-order finite differences for a range of values of* $\beta$. *Plot* $\mathrm{MSE}_\beta$ *and locate* $\beta_{\mathrm{MSE}}$. *Plot at least one of* $|\mathsf{RSS}(\hat{\boldsymbol{x}}_\beta) - n_{\mathrm{d}}|$ *or* $|\mathsf{RSS}(\hat{\boldsymbol{x}}_\beta) - \mathsf{REDF}(\beta)|$ *or* $\Phi_{\mathrm{CV}}(\beta)$ *or* $\Phi_{\mathrm{GCV}}(\beta)$ *and indicate the corresponding "optimized"* $\beta$ *values to compare to* $\beta_{\mathrm{MSE}}$. *Examine the restored images* $\hat{\boldsymbol{x}}_\beta$ *at* $\beta_{\mathrm{MSE}}$ *and the optimized value of* $\beta$ *select by the criterion you chose. Hint: no iterations are needed; do this using FFT operations.*

p,reg,hyper,cv,ybi

**Problem 2.14** *Prove the equality (2.5.36) used for simplifying cross validation. Also show that* $\mathsf{M}_{ii}(\beta) < 1$ *for* $\beta > 0$, *so the ratio in (2.5.36) is well defined.*

p,reg,garrotte

**Problem 2.15** *Consider a modified soft thresholding function of the form*

$$\hat{x}(y) = \arg\min_x \frac{1}{2}|y-x|^2 + \beta\,\psi(x) = y\left[1 - \frac{\lambda}{|y|}\frac{\lambda+\alpha}{|y|+\alpha}\right]_+,$$

*for* $\lambda > 0$ *and* $\alpha > -\lambda$. *For the special case* $\alpha = 0$, *this is known as the* ***nonnegative garrotte*** *[275–277]. Determine the corresponding (nonconvex) potential function* $\psi$ *when* $\beta = 1$ *and* $\alpha = 0$.

p,reg,hyperbola,fast

**Problem 2.16** *The hyperbola potential (2.4.5) has a weighting function that involves a reciprocal square root. Suppose instead we use the fast approximation to the* ***inverse square root*** *developed in the graphics community [wiki]. Determine the corresponding potential function* $\psi$ *and compare its derivative* $\dot{\psi}$ *to that of the usual hyperbola. (Solve?)*

p,reg,shrink,gf

**Problem 2.17** *Extend Problem 1.12 to the case of the generalized Fair potential in §2.7.4.*

p,reg,gauss,mix

**Problem 2.18** *Refine the breakpoints of the piecewise quadratic regularizer of §2.11 so that it better matches the negative logarithm of a gaussian mixture.*

p,reg,multi,line

**Problem 2.19** *This problem generalizes (2.8.3) and outlines the derivation of (2.8.4). (It also relates to certain* ***half quadratic*** *methods in the literature.) Let* $\psi$ *be any differentiable, symmetric potential function for which (see Theorem 12.4.5) the* ***potential weighting function*** $\omega_\psi(z) = \dot{\psi}(z)/z$ *is finite at* $z = 0$ *and monotone decreasing for* $|z| > 0$. *Let* $g(l) \triangleq \omega_\psi^{-1}(l)$ *denote the inverse of* $\omega_\psi$ *and, motivated by (12.4.15), define the function*

$$u(l) = \psi(g(l)) - \frac{1}{2}lg^2(l). \tag{2.15.3}$$

e,reg,multi,line,ul,gen

*Show that minimizing (2.8.3) over* $l_k$ *yields* $l_k = \omega_\psi\left(\sqrt{\sum_{m=1}^M \left|\left[\boldsymbol{C}\boldsymbol{x}_m^{(n)}\right]_k\right|^2}\right)$. *Determine which potential function* $\psi$ *corresponds to (2.8.4).*

p,reg,tv,trap

**Problem 2.20** *Consider a trapezoid defined by* $f(x) = \begin{cases} h, & |x| < a \\ h\left(1 - \frac{|x|-a}{b-a}\right), & a \leq |x| < b \\ 0, & \text{otherwise,} \end{cases}$ *for* $0 \leq a \leq b$ *and* $h > 0$. *Solve the optimization problem* $\arg\min_{a,b,h} \mathrm{TV}(f)$ *subject to* $\int f(x)\,\mathrm{d}x = 1$ *and* $f(x_0) = 0$ *for a given* $x_0 > 0$.

## 2.16 Bibliography


phillips:62:atf

[1] D. L. Phillips. "A technique for the numerical solution of certain integral equations of the first kind." In: *J. Assoc. Comput. Mach.* 9.1 (Jan. 1962), 84–97. DOI: 10.1145/321105.321114 (cit. on pp. 2.2, 2.3, 2.18).

tikhonov:63:soi

[2] A. N. Tikhonov. "Solution of incorrectly formulated problems and the regularization method." In: *Soviet Math. Dokl.* 4 (1963). English translation of Dkl. Akad. Nauk. SSSR, 141:501-4, 1963., 1035–8 (cit. on p. 2.2).

miller:70:lsm

[3] K. Miller. "Least-squares methods for ill-posed problems with a prescribed bound." In: *SIAM J. Math. Anal.* 1.1 (Feb. 1970), 52–70. DOI: 10.1137/0501006 (cit. on p. 2.2).

engl:93:rmf

[4] H. W. Engl. "Regularization methods for the stable solution of inverse problems." In: *Surveys on Mathematics for Industry* 3 (1993), 71–143 (cit. on p. 2.2).

hanke:93:rmf [5] M. Hanke and P. C. Hansen. "Regularization methods for large-scale problems." In: *Surveys on Mathematics for Industry* 3.4 (1993), 253–315 (cit. on p. 2.2).

engl:96 [6] H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of inverse problems*. Dordrecht: Kluwer, 1996 (cit. on p. 2.2).

hansen:98 [7] P. C. Hansen. *Rank-deficient and discrete ill-posed problems : numerical aspects of linear inversion*. Philadelphia: Soc. Indust. Appl. Math., 1998 (cit. on p. 2.2).

groetsch:93 [8] C. W. Groetsch. *Inverse problems in the mathematical sciences*. Wiesbaden, Germany: Vieweg, 1993 (cit. on p. 2.2).

hansen:94:rta [9] P. C. Hansen. "Regularization tools: a Matlab package for analysis and solution of discrete ill-posed problems." In: *Numer. Algorithms* 6.1 (Mar. 1994), 1–35. DOI: 10.1007/BF02149761 (cit. on p. 2.2).

wahba:90 [10] G. Wahba. *Spline models for observational data*. CBMS-NSF. Philadelphia: Soc. Indust. Appl. Math., 1990 (cit. on pp. 2.3, 2.5, 2.21, 2.22).

green:94 [11] P. J. Green and B. W. Silverman. *Nonparametric regression and generalized linear models: a roughness penalty approach*. London: Chapman and Hall, 1994 (cit. on pp. 2.3, 2.4).

gelfand:63 [12] I. M. Gelfand and S. V. Fomin. *Calculus of variations*. Translation by R A Silverman. NJ: Prentice-Hall, 1963 (cit. on p. 2.3).

reinsch:67:sbs [13] C. H. Reinsch. "Smoothing by spline functions." In: *Numerische Mathematik* 10.3 (Oct. 1967), 177–83. DOI: 10.1007/BF02162161 (cit. on pp. 2.3, 2.5).

deboor:78:apg [14] C. de Boor. *A practical guide to splines*. New York: Springer Verlag, 1978 (cit. on pp. 2.3, 2.4).

kybic:00:uou [15] J. Kybic et al. "Unwarping of unidirectionally distorted EPI images." In: *IEEE Trans. Med. Imag.* 19.2 (Feb. 2000), 80–93. DOI: 10.1109/42.836368 (cit. on p. 2.4).

mammen:97:lar [16] E. Mammen and S. van de Geer. "Locally adaptive regression splines." In: *Ann. Stat.* 25.1 (1997), 387–413. URL: http://www.jstor.org/stable/2242726 (cit. on p. 2.4).

szeliski:90:fsi [17] R. Szeliski. "Fast surface interpolation using hierarchical basis functions." In: *IEEE Trans. Patt. Anal. Mach. Int.* 12.6 (June 1990), 513–28 (cit. on p. 2.5).

kim:09:etf [18] S. Kim et al. "$\ell_1$ trend filtering." In: *SIAM Review* 51.2 (June 2009), 339–60. DOI: 10.1137/070690274 (cit. on p. 2.5).

dehoog:87:aem [19] F. R. de Hoog and M. F. Hutchinson. "An efficient method for calculating smoothing splines using orthogonal transformations." In: *Numerische Mathematik* 50.3 (May 1987), 311–9. DOI: 10.1007/BF01390708 (cit. on p. 2.5).

akaike:74:anl [20] H. Akaike. "A new look at the statistical model identification." In: *IEEE Trans. Auto. Control* 19.6 (Dec. 1974), 716–23. DOI: 10.1109/TAC.1974.1100705 (cit. on p. 2.6).

rissanen:78:mbs [21] J. Rissanen. "Modeling by shortest data description." In: *Automatica* 14.5 (Sept. 1978), 465–71. DOI: 10.1016/0005-1098(78)90005-5 (cit. on p. 2.6).

schwarz:78:etd [22] G. Schwarz. "Estimating the dimension of a model." In: *Ann. Stat.* 6.2 (1978), 461–4. DOI: 10.1214/aos/1176344136 (cit. on p. 2.6).

rissanen:87:sc [23] J. Rissanen. "Stochastic complexity." In: *J. Royal Stat. Soc. Ser. B* 49.3 (1987), 223–39. URL: http://www.jstor.org/stable/2985991 (cit. on p. 2.6).

hansen:01:msa [24] M. H. Hansen and B. Yu. "Model selection and the principle of minimum description length." In: *J. Am. Stat. Assoc.* 96.454 (June 2001), 746–75. URL: http://proquest.umi.com/pqdlink?did=74293072&sid=1&Fmt=2&clientId=17822&RQT=309&VName=PQD (cit. on p. 2.6).

stoica:04:mos [25] P. Stoica and Y. Selen. "Model-order selection." In: *IEEE Sig. Proc. Mag.* 21.4 (July 2004), 36–47. DOI: 10.1109/MSP.2004.1311138 (cit. on p. 2.6).

kritchman:08:dtn [26] S. Kritchman and B. Nadler. "Determining the number of components in a factor model from limited noisy data." In: *Chemometrics and Intelligent Laboratory Systems* 94.1 (Nov. 2008), 19–32. DOI: 10.1016/j.chemolab.2008.06.002 (cit. on p. 2.6).

fessler:03:aat [27] J. A. Fessler. "Analytical approach to regularization design for isotropic spatial resolution." In: *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.* Vol. 3. 2003, 2022–6. DOI: 10.1109/NSSMIC.2003.1352277 (cit. on pp. 2.10, 2.11).

shi:09:qrd [28] H. R. Shi and J. A. Fessler. "Quadratic regularization design for 2D CT." In: *IEEE Trans. Med. Imag.* 28.5 (May 2009), 645–56. DOI: 10.1109/TMI.2008.2007366 (cit. on pp. 2.10, 2.11).

stayman:04:cfn [29] J. W. Stayman and J. A. Fessler. "Compensation for nonuniform resolution using penalized-likelihood reconstruction in space-variant imaging systems." In: *IEEE Trans. Med. Imag.* 23.3 (Mar. 2004), 269–84. DOI: 10.1109/TMI.2003.823063 (cit. on pp. 2.10, 2.11).

wang:13:ssw
[30] X. Wang, P. Du, and J. Shen. "Smoothing splines with varying smoothing parameter." In: *Biometrika* 100.4 (2013), 955–70. DOI: 10.1093/biomet/ast031 (cit. on p. 2.11).

fessler:96:srp
[31] J. A. Fessler and W. L. Rogers. "Spatial resolution properties of penalized-likelihood image reconstruction methods: Space-invariant tomographs." In: *IEEE Trans. Im. Proc.* 5.9 (Sept. 1996), 1346–58. DOI: 10.1109/83.535846 (cit. on p. 2.11).

fessler:95:a30
[32] J. A. Fessler. *ASPIRE 3.0 user's guide: A sparse iterative reconstruction library*. Tech. rep. 293. Available from web.eecs.umich.edu/~fessler. Univ. of Michigan, Ann Arbor, MI, 48109-2122: Comm. and Sign. Proc. Lab., Dept. of EECS, July 1995. URL: http://web.eecs.umich.edu/~fessler/papers/lists/files/tr/95,293,aspire3.pdf (cit. on p. 2.13).

geiger:91:pad
[33] D. Geiger and F. Girosi. "Parallel and deterministic algorithms from MRF's: Surface reconstruction." In: *IEEE Trans. Patt. Anal. Mach. Int.* 13.5 (May 1991), 401–12. DOI: 10.1109/34.134040 (cit. on p. 2.13).

lee:95:bir
[34] S-J. Lee, A. Rangarajan, and G. Gindi. "Bayesian image reconstruction in SPECT using higher order mechanical models as priors." In: *IEEE Trans. Med. Imag.* 14.4 (Dec. 1995), 669–80. DOI: 10.1109/42.476108 (cit. on p. 2.13).

lee:97:ttp
[35] S. J. Lee, I. T. Hsiao, and G. R. Gindi. "The thin plate as a regularizer in Bayesian SPECT reconstruction." In: *IEEE Trans. Nuc. Sci.* 44.3 (June 1997), 1381–7. DOI: 10.1109/23.597017 (cit. on p. 2.13).

grimson:82:act
[36] W. E. L. Grimson. "A computational theory of visual surface interpolation." In: *Phil. Trans. Roy. Soc. London Ser. B* 298.1092 (Sept. 1982), 395–427. URL: http://www.jstor.org/stable/2395803 (cit. on p. 2.14).

duchon:77:smr
[37] J. Duchon. "Splines minimizing rotation-invariant semi-norms in Sobolev spaces." In: *Constructive Theory of Functions of Several Variables*. Ed. by W Schempp and K Zeller. Berlin: Springer, 1977, pp. 85–100 (cit. on p. 2.14).

bookstein:89:pwt
[38] F. L. Bookstein. "Principal warps: thin-plate splines and the decomposition of deformations." In: *IEEE Trans. Patt. Anal. Mach. Int.* 11.6 (June 1989), 567–87. DOI: 10.1109/34.24792 (cit. on p. 2.14).

farsiu:06:mda
[39] S. Farsiu, M. Elad, and P. Milanfar. "Multiframe demosaicing and super-resolution of color images." In: *IEEE Trans. Im. Proc.* 15.1 (Jan. 2006), 141–59. DOI: 10.1109/TIP.2005.860336 (cit. on pp. 2.14, 2.40).

aubert:97:avm
[40] G. Aubert and L. Vese. "A variational method in image recovery." In: *SIAM J. Numer. Anal.* 34.5 (Oct. 1997), 1948–97. DOI: 10.1137/S003614299529230X (cit. on p. 2.14).

sochen:98:agf
[41] N. Sochen, R. Kimmel, and R. Malladi. "A general framework for low level vision." In: *IEEE Trans. Im. Proc.* 7.3 (Mar. 1998), 310–8. DOI: 10.1109/83.661181 (cit. on p. 2.14).

rudin:92:ntv
[42] L. I. Rudin, S. Osher, and E. Fatemi. "Nonlinear total variation based noise removal algorithm." In: *Physica D* 60.1-4 (Nov. 1992), 259–68. DOI: 10.1016/0167-2789(92)90242-F (cit. on p. 2.14).

alliney:94:aaf
[43] S. Alliney and S. A. Ruzinsky. "An algorithm for the minimization of mixed $l_1$ and $l_2$ norms with application to Bayesian estimation." In: *IEEE Trans. Sig. Proc.* 42.3 (Mar. 1994), 618–27. DOI: 10.1109/78.277854 (cit. on p. 2.14).

dobson:96:aor
[44] D. Dobson and O. Scherzer. "Analysis of regularized total variation penalty methods for denoising." In: *Inverse Prob.* 12.5 (Oct. 1996), 601–17. DOI: 10.1088/0266-5611/12/5/005 (cit. on p. 2.14).

li:96:aca
[45] Y. Li and F. Santosa. "A computational algorithm for minimizing total variation in image restoration." In: *IEEE Trans. Im. Proc.* 5.6 (June 1996), 987–95. DOI: 10.1109/83.503914 (cit. on p. 2.14).

vogel:96:imf
[46] C. R. Vogel and M. E. Oman. "Iterative methods for total variation denoising." In: *SIAM J. Sci. Comp.* 17.1 (Jan. 1996), 227–38. DOI: 10.1137/0917016 (cit. on p. 2.14).

dobson:97:coa
[47] D. C. Dobson and C. R. Vogel. "Convergence of an iterative method for total variation denoising." In: *SIAM J. Numer. Anal.* 34.5 (Oct. 1997), 1779–91. DOI: 10.1137/S003614299528701X (cit. on p. 2.14).

chan:99:anp
[48] T. F. Chan, G. H. Golub, and P. Mulet. "A nonlinear primal-dual method for total variation-based image restoration." In: *SIAM J. Sci. Comp.* 20.6 (1999), 1964–77. DOI: 10.1137/S1064827596299767 (cit. on p. 2.14).

candes:02:nmt
[49] E. J. Candès and F. Guo. "New multiscale transforms, minimum total variation synthesis: applications to edge-preserving image reconstruction." In: *sp* 82.11 (Nov. 2002), 1519–43. DOI: 10.1016/S0165-1684(02)00300-6 (cit. on p. 2.14).

strong:03:epa
[50] D. Strong and T. Chan. "Edge-preserving and scale-dependent properties of total variation regularization." In: *Inverse Prob.* 19.6 (Dec. 2003), S165–87. DOI: 10.1088/0266-5611/19/6/059 (cit. on p. 2.14).

hintermuller:06:aip
[51] M. Hintermüller and G. Stadler. "An infeasible primal-dual algorithm for total bounded variation–based inf-convolution-type image restoration." In: *SIAM J. Sci. Comp.* 28.1 (2006), 1–23. DOI: 10.1137/040613263 (cit. on p. 2.14).

michailovich:11:ais [52] O. V. Michailovich. "An iterative shrinkage approach to total-variation image restoration." In: *IEEE Trans. Im. Proc.* 20.5 (May 2011), 1281–99. DOI: 10.1109/TIP.2010.2090532 (cit. on p. 2.15).

hu:12:hdt [53] Y. Hu and M. Jacob. "Higher degree total variation (HDTV) regularization for image recovery." In: *IEEE Trans. Im. Proc.* 21.5 (May 2012), 2559–71. DOI: 10.1109/TIP.2012.2183143 (cit. on p. 2.15).

bredies:10:tgv [54] K. Bredies, K. Kunisch, and T. Pock. "Total generalized variation." In: *SIAM J. Imaging Sci.* 3 (2010), 492–526. DOI: 10.1137/090769521 (cit. on p. 2.15).

knoll:11:sot [55] F. Knoll et al. "Second order total generalized variation (TGV) for MRI." In: *Mag. Res. Med.* 65.2 (2011), 480–91. DOI: 10.1002/mrm.22595 (cit. on p. 2.15).

lefkimmiatis:12:hbn [56] S. Lefkimmiatis, A. Bourquard, and M. Unser. "Hessian-based norm regularization for image restoration with biomedical applications." In: *IEEE Trans. Im. Proc.* 21.3 (Mar. 2012), 983–5. DOI: 10.1109/TIP.2011.2168232 (cit. on p. 2.15).

huang:08:aft [57] Y. Huang, M. K. Ng, and Y-W. Wen. "A fast total variation minimization method for image restoration." In: *SIAM Multiscale Modeling and Simulation* 7.2 (2008), 774–95. DOI: 10.1137/070703533 (cit. on p. 2.15).

wang:08:ana [58] Y. Wang et al. "A new alternating minimization algorithm for total variation image reconstruction." In: *SIAM J. Imaging Sci.* 1.3 (2008), 248–72. DOI: 10.1137/080724265 (cit. on p. 2.16).

courant:1943:vmf [59] R. Courant. "Variational methods for the solution of problems of equilibrium and vibrations." In: *Bull. Amer. Math. Soc.* 49 (1943), 1–23. DOI: 10.1090/S0002-9904-1943-07818-4 (cit. on p. 2.16).

darbon:06:irw [60] J. Darbon and M. Sigelle. "Image restoration with discrete constrained total variation part I: Fast and exact optimization." In: *J. Math. Im. Vision* 26.3 (Dec. 2006), 261–76. DOI: 10.1007/s10851-006-8803-0 (cit. on p. 2.16).

tai:09:alm [61] X-C. Tai and C. Wu. "Augmented Lagrangian method, dual methods and split Bregman iteration for ROF model." In: *LNCS 5567*. Proc. of the Second International Conference on Scale Space and Variational Methods in Computer Vision. Section: Image Enhancement and Reconstruction. 2009, 502–13. DOI: 10.1007/978-3-642-02256-2_42 (cit. on p. 2.16).

figueiredo:10:rop [62] M. A. T. Figueiredo and José M Bioucas-Dias. "Restoration of Poissonian images using alternating direction optimization." In: *IEEE Trans. Im. Proc.* 19.12 (Dec. 2010), 3133–45. DOI: 10.1109/TIP.2010.2053941 (cit. on pp. 2.16, 2.35).

johnson:91:iru [63] V. E. Johnson et al. "Image restoration using Gibbs priors: Boundary modeling, treatment of blurring, and selection of hyperparameter." In: *IEEE Trans. Patt. Anal. Mach. Int.* 13.5 (May 1991), 413–25. DOI: 10.1109/34.134041 (cit. on p. 2.16).

hall:87:cso [64] P. Hall and D. M. Titterington. "Common structure of techniques for choosing smoothing parameters in regression problems." In: *J. Royal Stat. Soc. Ser. B* 49.2 (1987), 184–98. URL: http://www.jstor.org/stable/2345419 (cit. on pp. 2.16, 2.18, 2.19, 2.20).

thompson:91:aso [65] A. M. Thompson et al. "A study of methods or choosing the smoothing parameter in image restoration by regularization." In: *IEEE Trans. Patt. Anal. Mach. Int.* 13.4 (Apr. 1991), 326–39. DOI: 10.1109/34.88568 (cit. on pp. 2.16, 2.17, 2.20).

galatsanos:92:mfc [66] N. P. Galatsanos and A. K. Katsaggelos. "Methods for choosing the regularization parameter and estimating the noise variance in image restoration and their relation." In: *IEEE Trans. Im. Proc.* 1.3 (July 1992), 322–336. DOI: 10.1109/83.148606 (cit. on pp. 2.16, 2.22, 2.25).

thompson:93:osb [67] A. M. Thompson and J. Kay. "On some Bayesian choices of regularization parameter in image restoration." In: *Inverse Prob.* 9.6 (Dec. 1993), 749–61. DOI: 10.1088/0266-5611/9/6/011 (cit. on pp. 2.16, 2.23).

archer:95:osb [68] G. Archer and D. M. Titterington. "On some Bayesian/regularization methods for image restoration." In: *IEEE Trans. Im. Proc.* 4.7 (July 1995), 989–95. DOI: 10.1109/83.392339 (cit. on pp. 2.16, 2.23).

jones:96:abs [69] M. C. Jones, J. S. Marron, and S. J. Sheather. "A brief survey of bandwidth selection for density estimation." In: *J. Am. Stat. Assoc.* 91.433 (Mar. 1996), 401–7. URL: http://www.jstor.org/stable/2291420 (cit. on p. 2.16).

gu:98:mia [70] C. Gu. "Model indexing and smoothing parameter selection in nonparametric function estimation." In: *Statistica Sinica* 8.3 (July 1998), 607–46. URL: http://www3.stat.sinica.edu.tw/statistica/j8n3/j8n31/j8n31.htm (cit. on p. 2.16).

lukas:98:cop [71] M. A. Lukas. "Comparisons of parameter choice methods for regularization with discrete noisy data." In: *Inverse Prob.* 14.1 (Feb. 1998), 161–84. DOI: 10.1088/0266-5611/14/1/014 (cit. on p. 2.16).

james:61:ewq [72] W. James and C. Stein. "Estimation with quadratic loss." In: *Proc. Fourth Berkeley Symp. Math. Statist. Prob.* 1 (1961), 361–79. URL: http://projecteuclid.org/euclid.bsmsp/1200512173 (cit. on p. 2.17).

vogel:02
[73] C. R. Vogel. *Computational methods for inverse problems*. Soc. Indust. Appl. Math., 2002. DOI: 10.1137/1.9780898717570 (cit. on pp. 2.17, 2.20).

desbat:95:tmr
[74] L. Desbat and D. Girard. "The "minimum reconstruction error" choice of regularization parameters: Some effective methods and their application to deconvolution problems." In: *SIAM J. Sci. Comp.* 16.6 (Nov. 1995), 1387–403. DOI: 10.1137/0916080 (cit. on pp. 2.17, 2.24).

vainikko:82:tdp
[75] G. M. Vainikko. "The discrepancy principle for a class of regularization methods." In: *USSR Comp. Math. and Math. Phys.* 22.3 (1982), 1–19. DOI: 10.1016/0041-5553(82)90120-3 (cit. on p. 2.18).

osher:05:air
[76] S. Osher et al. "An iterative regularization method for total variation based image restoration." In: *SIAM Multiscale Modeling and Simulation* 4.2 (2005), 460–89. DOI: 10.1137/040605412 (cit. on p. 2.19).

hebert:92:sbm
[77] T. J. Hebert and R. Leahy. "Statistic-based MAP image reconstruction from Poisson data using Gibbs priors." In: *IEEE Trans. Sig. Proc.* 40.9 (Sept. 1992), 2290–303. DOI: 10.1109/78.157228 (cit. on pp. 2.19, 2.27).

zanella:09:egp
[78] R. Zanella et al. "Efficient gradient projection methods for edge-preserving removal of Poisson noise." In: *Inverse Prob.* 25.4 (Apr. 2009), p. 045010. DOI: 10.1088/0266-5611/25/4/045010 (cit. on p. 2.19).

teuber:13:map
[79] T. Teuber, G. Steidl, and R. H. Chan. "Minimization and parameter estimation for seminorm regularization models with I -divergence constraints." In: *Inverse Prob.* 29.3 (Mar. 2013), p. 035007. DOI: 10.1088/0266-5611/29/3/035007 (cit. on p. 2.19).

stagliano:11:aoa
[80] A. Staglianò, P. Boccacci, and M. Bertero. "Analysis of an approximate model for Poisson data reconstruction and a related discrepancy principle." In: *Inverse Prob.* 27.12 (Dec. 2011), p. 125003. DOI: 10.1088/0266-5611/27/12/125003 (cit. on p. 2.19).

wahba:83:bci
[81] G. Wahba. "Bayesian "confidence intervals" for the cross-validated smoothing spline." In: *J. Royal Stat. Soc. Ser. B* 45.1 (1983), 133–50. URL: http://www.jstor.org/stable/2345632 (cit. on p. 2.19).

janson:15:edo
[82] L. Janson, W. Fithian, and T. J. Hastie. "Effective degrees of freedom: a flawed metaphor." In: *Biometrika* 102.2 (2015), 479–85. DOI: 10.1093/biomet/asv019 (cit. on p. 2.19).

allen:74:trb
[83] D. M. Allen. "The relationship between variable selection and data agumentation and a method for prediction." In: *Technometrics* 16.1 (Feb. 1974), 125–7. URL: http://www.jstor.org/stable/1267500 (cit. on p. 2.21).

wahba:75:aca
[84] G. Wahba and S. Wold. "A completely automatic French curve: Fitting spline functions by cross validation." In: *Comm. in Statistics—Theory and Methods* 4.1 (1975), 1–17. DOI: 10.1080/03610927508827223 (cit. on p. 2.21).

craven:79:snd
[85] P. Craven and G. Wahba. "Smoothing noisy data with spline functions." In: *Numerische Mathematik* 31.4 (Dec. 1979), 377–403. DOI: 10.1007/BF01404567 (cit. on p. 2.21).

hall:09:rvo
[86] P. Hall and A. P. Robinson. "Reducing variability of crossvalidation for smoothing-parameter choice." In: *Biometrika* 96 (2009), 175–86. DOI: 10.1093/biomet/asn068 (cit. on p. 2.21).

lim:13:esw
[87] C. Lim and B. Yu. *Estimation stability with cross validation (ESCV)*. arxiv 1303.3128. 2013. URL: http://arxiv.org/abs/1303.3128 (cit. on p. 2.21).

golub:79:gcv
[88] G. H. Golub, M. Heath, and G. Wahba. "Generalized cross-validation as a method for choosing a good ridge parameter." In: *Technometrics* 21.2 (May 1979), 215–23. URL: http://www.jstor.org/stable/1268518 (cit. on p. 2.21).

utreras:81:oso
[89] F. Utreras. "Optimal smoothing of noisy data using spline functions." In: *SIAM J. Sci. Stat. Comp.* 2.3 (Sept. 1981), 349–62. DOI: 10.1137/0902028 (cit. on p. 2.22).

reeves:90:oeo
[90] S. J. Reeves and R. M. Mersereau. "Optimal estimation of the regularization parameter and stabilizing functional for regularized image restoration." In: *Optical Engineering* 29.5 (May 1990), 446–54. DOI: 10.1117/12.55613 (cit. on p. 2.22).

reeves:92:bib
[91] S. J. Reeves and R. M. Mersereau. "Blur identification by the method of generalized cross-validation." In: *IEEE Trans. Im. Proc.* 1.3 (July 1992), 301–11. DOI: 10.1109/83.148604 (cit. on p. 2.22).

hutchinson:90:ase
[92] M. F. Hutchinson. "A stochastic estimator for the trace of the influence matrix for Laplacian smoothing splines." In: *Comm. in Statistics - Simulation and Computation* 19.2 (1990), 433–50. DOI: 10.1080/03610919008812866 (cit. on p. 2.22).

deshpande:91:fco
[93] L. N. Deshpande and D. A. Girard. "Fast computation of cross-validated robust splines and other non-linear smoothing splines." In: *Curves and Surfaces*. Ed. by Larry L Schumaker Pierre-Jean Laurent Alain Le Méhauté. Boston, MA: Academic, 1991, pp. 143–8 (cit. on p. 2.22).

girard:95:tfm
[94] D. A. Girard. "The fast Monte-Carlo cross-validation and CL procedures: Comments, new results and application to image recovery problems." In: *Comput. Statist.* 10.3 (1995), 205–31 (cit. on p. 2.22).

golub:97:gcv
[95] G. H. Golub and U. von Matt. "Generalized cross-validation for large-scale problems." In: *J. Computational and Graphical Stat.* 6.1 (Mar. 1997), 1–34. URL: http://www.jstor.org/stable/1390722 (cit. on p. 2.22).

reeves:95:gcv
[96] S. J. Reeves. "Generalized cross-validation as a stopping rule for the Richardson-Lucy algorithm." In: *Intl. J. Imaging Sys. and Tech.* 6.4 (1995), 387–91. DOI: 10.1002/ima.1850060412 (cit. on pp. 2.22, 2.27).

ramani:08:mcs
[97] S. Ramani, T. Blu, and M. Unser. "Monte-Carlo SURE: A black-box optimization of regularization parameters for general denoising algorithms." In: *IEEE Trans. Im. Proc.* 17.9 (Sept. 2008), 1540–54. DOI: 10.1109/TIP.2008.2001404 (cit. on pp. 2.22, 2.23, 2.24, 2.25, 2.27).

dong:94:sew
[98] S. Dong and K. Liu. "Stochastic estimation with z2 noise." In: *Phys. Lett. B* 328.1-2 (May 1994), 130–6. DOI: '10.1016/0370-2693(94)90440-5' (cit. on p. 2.22).

ramani:12:rps
[99] S. Ramani et al. "Regularization parameter selection for nonlinear iterative image restoration and MRI reconstruction using GCV and SURE-based methods." In: *IEEE Trans. Im. Proc.* 21.8 (Aug. 2012), 3659–72. DOI: 10.1109/TIP.2012.2195015 (cit. on p. 2.22).

osullivan:85:acv
[100] F. O'Sullivan and G. Wahba. "A cross validated Bayesian retrieval algorithm for nonlinear remote sensing experiments." In: *J. Comp. Phys.* 59.3 (July 1985), 441–55. DOI: 10.1016/0021-9991(85)90121-4 (cit. on p. 2.22).

haber:00:agb
[101] E. Haber and D. Oldenburg. "A GCV based method for nonlinear ill-posed problems." In: *Comput. Geosci.* 4.1 (2000), 41–63. DOI: 10.1023/A:1011599530422 (cit. on p. 2.22).

fu:05:nga
[102] W. J. Fu. "Nonlinear GCV and quasi-GCV for shrinkage models." In: *J. Statist. Plann. Inference* 131.2 (2005), 333–47. DOI: 10.1016/j.jspi.2004.03.001 (cit. on p. 2.22).

giryes:11:tpg
[103] R. Giryes, M. Elad, and Y. C. Eldar. "The projected GSURE for automatic parameter tuning in iterative shrinkage methods." In: *Applied and Computational Harmonic Analysis* 30.3 (May 2011), 407–22. DOI: 10.1016/j.acha.2010.11.005 (cit. on pp. 2.22, 2.23, 2.27).

saquib:98:mpe
[104] S. S. Saquib, C. A. Bouman, and K. Sauer. "ML parameter estimation for Markov random fields, with applications to Bayesian tomography." In: *IEEE Trans. Im. Proc.* 7.7 (July 1998), 1029–44. DOI: 10.1109/83.701163 (cit. on p. 2.23).

higdon:97:fbe
[105] D. M. Higdon et al. "Fully Bayesian estimation of Gibbs hyperparameters for emission computed tomography data." In: *IEEE Trans. Med. Imag.* 16.5 (Oct. 1997), 516–26. DOI: 10.1109/42.640741 (cit. on p. 2.23).

keren:99:afb
[106] D. Keren and M. Werman. "A full Bayesian approach to curve and surface reconstruction." In: *J. Math. Im. Vision* 11.1 (Sept. 1999), 27–43. DOI: 10.1023/A:1008317210576 (cit. on p. 2.23).

ying:08:asa
[107] L. Ying et al. "A statistical approach to SENSE regularization with arbitrary k-space trajectories." In: *Mag. Res. Med.* 60.2 (Aug. 2008), 414–21. DOI: 10.1002/mrm.21665 (cit. on p. 2.23).

bertero:10:adp
[108] M. Bertero et al. "A discrepancy principle for Poisson data." In: *Inverse Prob.* 26.10 (Oct. 2010), p. 105004. DOI: 10.1088/0266-5611/26/10/105004 (cit. on p. 2.24).

hansen:92:aod
[109] P. C. Hansen. "Analysis of discrete ill-posed problems by means of the L-curve." In: *SIAM Review* 34.4 (Dec. 1992), 561–580. DOI: 10.1137/1034115 (cit. on p. 2.24).

hansen:93:tuo
[110] P. C. Hansen and D. P. O'Leary. "The use of the L-curve in the regularization of discrete ill-posed problems." In: *SIAM J. Sci. Comp.* 14.6 (1993), 1487–506. DOI: 10.1137/0914086 (cit. on p. 2.24).

reginska:96:arp
[111] T. Regińska. "A regularization parameter in discrete ill-posed problems." In: *SIAM J. Sci. Comp.* 17.3 (May 1996), 740–9. DOI: 10.1137/S1064827593252672 (cit. on p. 2.24).

kaufman:96:prb
[112] L. Kaufman and A. Neuman. "PET regularization by envelope guided conjugate gradients." In: *IEEE Trans. Med. Imag.* 15.3 (June 1996), 385–6. DOI: 10.1109/42.500147 (cit. on p. 2.24).

belge:02:edo
[113] M. Belge, M. E. Kilmer, and E. L. Miller. "Efficient determination of multiple regularization parameters in a generalized L-curve framework." In: *Inverse Prob.* 18.4 (Aug. 2002), 1161–83. DOI: 10.1088/0266-5611/18/4/314 (cit. on p. 2.24).

vogel:96:nco
[114] C. R. Vogel. "Non-convergence of the L-curve regularization parameter selection method." In: *Inverse Prob.* 12.4 (Aug. 1996), 535–47. DOI: 10.1088/0266-5611/12/4/013 (cit. on p. 2.24).

stein:81:eot
[115] C. Stein. "Estimation of the mean of a multivariate normal distribution." In: *Ann. Stat.* 9.6 (Nov. 1981), 1135–51. DOI: 10.1214/aos/1176345632. URL: http://www.jstor.org/stable/2240405 (cit. on pp. 2.24, 2.25).

rice:86:cos
[116] J. A. Rice. "Choice of smoothing parameter in deconvolution problems." In: *Contemporary Mathematics* 59 (1986), 137–51. DOI: 10.1090/conm/059/10 (cit. on p. 2.24).

solo:96:asf
[117] V. Solo. "A sure-fired way to choose smoothing parameters in ill-conditioned inverse problems." In: *Proc. IEEE Intl. Conf. on Image Processing*. Vol. 3. 1996, 89–92. DOI: 10.1109/ICIP.1996.560376 (cit. on p. 2.24).

eldar:08:rbe
[118] Y. C. Eldar. "Rethinking biased estimation: Improving maximum likelihood and the Cramer-Rao bound." In: *Found. & Trends in Sig. Pro.* 1.4 (2008), 305–449. DOI: 10.1561/2000000008 (cit. on pp. 2.24, 2.25, 2.27).

eldar:09:gsf
[119] Y. C. Eldar. "Generalized SURE for exponential families: applications to regularization." In: *IEEE Trans. Sig. Proc.* 57.2 (Feb. 2009), 471–81. DOI: 10.1109/TSP.2008.2008212 (cit. on pp. 2.24, 2.25, 2.27).

pesquet:09:asa
[120] J-C. Pesquet, A. Benazza-Benyahia, and C. Chaux. "A SURE approach for digital signal/image deconvolution problems." In: *IEEE Trans. Sig. Proc.* 57.12 (Dec. 2009), 4616–32. DOI: 10.1109/TSP.2009.2026077 (cit. on p. 2.24).

ramani:13:ncm
[121] S. Ramani et al. "Non-Cartesian MRI reconstruction with automatic regularization via Monte-Carlo SURE." In: *IEEE Trans. Med. Imag.* 32.8 (Aug. 2013), 1411–22. DOI: 10.1109/TMI.2013.2257829 (cit. on pp. 2.24, 2.27).

deledalle:14:sug
[122] C-A. Deledalle et al. "Stein unbiased grAdient estimator of the risk (SUGAR) for multiple parameter selection." In: *SIAM J. Imaging Sci.* 7.4 (2014), 2448–87. DOI: 10.1137/140968045 (cit. on p. 2.24).

weller:14:mcs
[123] D. S. Weller et al. "Monte Carlo SURE-based parameter selection for parallel magnetic resonance imaging reconstruction." In: *Mag. Res. Med.* 71.5 (May 2014), 1760–70. DOI: 10.1002/mrm.24840 (cit. on pp. 2.24, 2.27).

lucka:17:ref
[124] F. Lucka et al. *Risk estimators for choosing regularization parameters in ill-posed problems - properties and limitations*. arxiv 1701.04970. 2017. URL: http://arxiv.org/abs/1701.04970 (cit. on p. 2.24).

blu:07:tsl
[125] T. Blu and F. Luisier. "The SURE-LET approach to image denoising." In: *IEEE Trans. Im. Proc.* 16.11 (Nov. 2007), 2778–86. DOI: 10.1109/TIP.2007.906002 (cit. on p. 2.25).

nowak:97:ose
[126] R. D. Nowak. "Optimal signal estimation using cross-validation." In: *IEEE Signal Proc. Letters* 4.1 (Jan. 1997), 23–5. DOI: 10.1109/97.551692 (cit. on p. 2.27).

nowak:99:wdf
[127] R. D. Nowak and R. G. Baraniuk. "Wavelet-domain filtering for photon imaging systems." In: *IEEE Trans. Im. Proc.* 8.5 (May 1999), 666–78. DOI: 10.1109/83.760334 (cit. on p. 2.27).

liang:15:rpt
[128] H. Liang and D. S. Weller. "Regularization parameter trimming for iterative image reconstruction." In: *Proc., IEEE Asilomar Conf. on Signals, Systems, and Comp.* 2015, 755–9. DOI: 10.1109/ACSSC.2015.7421235 (cit. on p. 2.27).

frommer:99:fcb
[129] A. Frommer and P. Maass. "Fast CG-based methods for Tikhonov-Phillips regularization." In: *SIAM J. Sci. Comp.* 20.5 (1999), 1831–50. DOI: 10.1137/S1064827596313310 (cit. on p. 2.27).

veklerov:87:srf
[130] E. Veklerov and J. Llacer. "Stopping rule for the MLE algorithm based on statistical hypothesis testing." In: *IEEE Trans. Med. Imag.* 6.4 (Dec. 1987), 313–9. DOI: 10.1109/TMI.1987.4307849 (cit. on p. 2.27).

llacer:89:fia
[131] J. Llacer and E. Veklerov. "Feasible images and practical stopping rules for iterative algorithms in emission tomography." In: *IEEE Trans. Med. Imag.* 8.2 (June 1989). Corrections, 9(1), Mar 1990, 186–93. DOI: 10.1109/42.24867 (cit. on p. 2.27).

johnson:94:ano
[132] V. E. Johnson. "A note on stopping rules in EM-ML reconstructions of ECT images." In: *IEEE Trans. Med. Imag.* 13.3 (Sept. 1994), 569–71. DOI: 10.1109/42.310891 (cit. on p. 2.27).

perry:94:aps
[133] K. M. Perry and S. J. Reeves. "A practical stopping rule for iterative signal restoration." In: *IEEE Trans. Sig. Proc.* 42.7 (July 1994), 1829–32. DOI: 10.1109/78.298292 (cit. on p. 2.27).

selivanov:01:cvs
[134] V. V. Selivanov et al. "Cross-validation stopping rule for ML-EM reconstruction of dynamic PET series: effect on image quality and quantitative accuracy." In: *IEEE Trans. Nuc. Sci.* 48.3 (June 2001), 883–9. DOI: 10.1109/NSSMIC.1999.842828 (cit. on p. 2.27).

bauer:05:alt
[135] F. Bauer and T. Hohage. "A Lepskij-type stopping rule for regularized Newton methods." In: *Inverse Prob.* 21.6 (Dec. 2005), 1975–92. DOI: 10.1088/0266-5611/21/6/011 (cit. on p. 2.27).

blanchard:12:dpf
[136] G. Blanchard and P. Mathé. "Discrepancy principle for statistical inverse problems with application to conjugate gradient iteration." In: *Inverse Prob.* 28.11 (Nov. 2012), p. 115011. DOI: 10.1088/0266-5611/28/11/115011 (cit. on p. 2.27).

fan:95:ddb
[137] J. Fan and I. Gijbels. "Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation." In: *J. Royal Stat. Soc. Ser. B* 57.2 (1995), 371–94. URL: http://www.jstor.org/stable/2345968 (cit. on p. 2.27).

charbonnier:94:tdh
[138] P. Charbonnier et al. "Two deterministic half-quadratic regularization algorithms for computed imaging." In: *Proc. IEEE Intl. Conf. on Image Processing*. Vol. 2. 1994, 168–71. DOI: 10.1109/ICIP.1994.413553 (cit. on p. 2.28).

panin:99:tvr
[139] V. Y. Panin, G. L. Zeng, and G. T. Gullberg. "Total variation regulated EM algorithm." In: *IEEE Trans. Nuc. Sci.* 46.6 (Dec. 1999), 2202–10. DOI: 10.1109/23.819305 (cit. on p. 2.28).

kisilev:01:wra
[140] P. Kisilev, M. Zibulevsky, and Y. Zeevi. "Wavelet representation and total variation regularization in emission tomography." In: *Proc. IEEE Intl. Conf. on Image Processing*. Vol. 1. 2001, 702–5. DOI: 10.1109/ICIP.2001.959142 (cit. on p. 2.28).

green:90:ouo
[141] P. J. Green. "On use of the EM algorithm for penalized likelihood estimation." In: *J. Royal Stat. Soc. Ser. B* 52.3 (1990), 443–52. URL: http://www.jstor.org/stable/2345668 (cit. on p. 2.28).

green:90:brf
[142] P. J. Green. "Bayesian reconstructions from emission tomography data using a modified EM algorithm." In: *IEEE Trans. Med. Imag.* 9.1 (Mar. 1990), 84–93. DOI: 10.1109/42.52985 (cit. on p. 2.28).

lange:90:coe
[143] K. Lange. "Convergence of EM image reconstruction algorithms with Gibbs smoothing." In: *IEEE Trans. Med. Imag.* 9.4 (Dec. 1990). Corrections, T-MI, 10:2(288), June 1991., 439–46. DOI: 10.1109/42.61759 (cit. on pp. 2.28, 2.31, 2.32, 2.33).

fair:74:otr
[144] R. C. Fair. "On the robust estimation of econometric models." In: *Ann. Econ. Social Measurement* 2 (Oct. 1974), 667–77. URL: http://fairmodel.econ.yale.edu/rayfair/pdf/1974D.HTM (cit. on pp. 2.28, 2.31, 2.32, 2.33).

holland:77:rru
[145] P. W. Holland and R. E. Welsch. "Robust regression using iteratively reweighted least-squares." In: *Comm. in Statistics—Theory and Methods* 6.9 (1977), 813–27. DOI: 10.1080/03610927708827533 (cit. on pp. 2.28, 2.29, 2.31, 2.32, 2.33).

rey:83
[146] W. J. J. Rey. *Introduction to robust and quasi-robust statistical methods*. Berlin: Springer, 1983 (cit. on pp. 2.28, 2.31, 2.32, 2.33).

li:98:cfs
[147] S. Z. Li. "Close-form solution and parameter selection for convex minimization-based edge-preserving smoothing." In: *IEEE Trans. Patt. Anal. Mach. Int.* 20.9 (Sept. 1998), 916–32. DOI: 10.1109/34.713359 (cit. on pp. 2.28, 2.29).

dax:92:orl
[148] A. Dax. "On regularized least norm problems." In: *SIAM J. Optim.* 2.4 (1992), 602–18. DOI: 10.1137/0802029 (cit. on pp. 2.28, 2.29).

bouman:93:agg
[149] C. Bouman and K. Sauer. "A generalized Gaussian image model for edge-preserving MAP estimation." In: *IEEE Trans. Im. Proc.* 2.3 (July 1993), 296–310. DOI: 10.1109/83.236536 (cit. on pp. 2.28, 2.29).

zervakis:95:aco
[150] M. E. Zervakis, A. K. Katsaggelos, and T. M. Kwon. "A class of robust entropic functionals for image restoration." In: *IEEE Trans. Im. Proc.* 4.6 (June 1995), 752–73. DOI: 10.1109/83.388078 (cit. on pp. 2.28, 2.29).

orchard:03:sra
[151] J. Orchard et al. "Simultaneous registration and activation detection for fMRI." In: *IEEE Trans. Med. Imag.* 22.11 (Nov. 2003), 1427–35. DOI: 10.1109/TMI.2003.819294 (cit. on p. 2.29).

beaton:74:tfo
[152] A. E. Beaton and J. W. Tukey. "The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data." In: *Technometrics* 16.2 (May 1974), 147–85. URL: http://www.jstor.org/stable/1267936 (cit. on p. 2.29).

hebert:89:age
[153] T. Hebert and R. Leahy. "A generalized EM algorithm for 3-D Bayesian reconstruction from Poisson data using Gibbs priors." In: *IEEE Trans. Med. Imag.* 8.2 (June 1989), 194–202. DOI: 10.1109/42.24868 (cit. on pp. 2.29, 2.33).

delaney:98:gce
[154] A. H. Delaney and Y. Bresler. "Globally convergent edge-preserving regularized reconstruction: an application to limited-angle tomography." In: *IEEE Trans. Im. Proc.* 7.2 (Feb. 1998), 204–21. DOI: 10.1109/83.660997 (cit. on pp. 2.29, 2.33).

bourgeois:01:rom
[155] M. Bourgeois et al. "Reconstruction of MRI images from non-uniform sampling and application to intrascan motion correction in functional MRI." In: *Modern Sampling Theory: Mathematics and Applications*. Ed. by J J Benedetto and P Ferreira. Boston: Birkhauser, 2001, pp. 343–63 (cit. on pp. 2.29, 2.33).

wajer:00:nsi
[156] F. T. A. W. Wajer et al. "Nonuniform sampling in magnetic resonance imaging." In: *Proc. IEEE Conf. Acoust. Speech Sig. Proc.* Vol. 6. 2000, 3846–9. DOI: 10.1109/ICASSP.2000.860242 (cit. on p. 2.29).

candes:08:esb
[157] E. J. Candes, M. B. Wakin, and S. Boyd. "Enhancing sparsity by reweighted l1 minimization." In: *J. Fourier Anal. and Appl.* 14.5 (Dec. 2008), 877–905. DOI: 10.1007/s00041-008-9045-x (cit. on p. 2.29).

ramirez:12:urf
[158] I. Ramirez and G. Sapiro. "Universal regularizers for robust sparse coding and modeling." In: *IEEE Trans. Im. Proc.* 21.9 (Sept. 2012), 3850–64. DOI: 10.1109/TIP.2012.2197006 (cit. on p. 2.29).

geman:85:bia
[159] S. Geman and D. E. McClure. "Bayesian image analysis: an application to single photon emission tomography." In: *Proc. of Stat. Comp. Sect. of Amer. Stat. Assoc.* 1985, 12–8. URL: http://www.dam.brown.edu/people/geman/Homepage/Image%20processing,%20image%20analysis,%20Markov%20random%20fields,%20and%20MCMC/1985GemanMcClureASA.pdf (cit. on pp. 2.29, 2.30, 2.31, 2.34).

geman:92:cra
[160] D. Geman and G. Reynolds. "Constrained restoration and the recovery of discontinuities." In: *IEEE Trans. Patt. Anal. Mach. Int.* 14.3 (Mar. 1992), 367–83. DOI: 10.1109/34.120331 (cit. on pp. 2.29, 2.30, 2.31).

boykov:01:fae [161] Y. Boykov, O. Veksler, and R. Zabih. "Fast approximate energy minimization via graph cuts." In: *IEEE Trans. Patt. Anal. Mach. Int.* 23.11 (Nov. 2001), 1222–39. DOI: 10.1109/34.969114 (cit. on p. 2.29).

soubies:15:ace [162] E. Soubies, L. Blanc-Féraud, and G. Aubert. "A continuous exact $\ell_0$ penalty (CEL0) for least squares regularized problem." In: *SIAM J. Imaging Sci.* 8.3 (2015), 1607–39. DOI: 10.1137/151003714 (cit. on p. 2.29).

rivera:03:ehq [163] M. Rivera and J. L. Marroquin. "Efficient half-quadratic regularization with granularity control." In: *Im. and Vision Computing* 21.4 (Apr. 2003), 345–57. DOI: 10.1016/S0262-8856(03)00005-2 (cit. on p. 2.29).

lalush:93:agg [164] D. S. Lalush and B. M. W. Tsui. "A generalized Gibbs prior for maximum a posteriori reconstruction in SPECT." In: *Phys. Med. Biol.* 38.6 (June 1993), 729–41. DOI: 10.1088/0031-9155/38/6/007 (cit. on p. 2.29).

stevenson:94:dpr [165] R. L. Stevenson, B. E. Schmitz, and E. J. Delp. "Discontinuity preserving regularization of inverse visual problems." In: *ieee-smc* 24.3 (Mar. 1994), 455–69. DOI: 10.1109/21.278994 (cit. on p. 2.30).

chartrand:12:nsf [166] R. Chartrand. "Nonconvex splitting for regularized low-rank + sparse decomposition." In: *IEEE Trans. Sig. Proc.* 60.11 (Nov. 2012), 5810–9. DOI: 10.1109/TSP.2012.2208955 (cit. on p. 2.30).

huber:64:reo [167] P. J. Huber. "Robust estimation of a location parameter." In: *Ann. Math. Stat.* 35.1 (Mar. 1964), 73–101. URL: http://www.jstor.org/stable/2238020 (cit. on p. 2.30).

mehranian:13:aos [168] A. Mehranian et al. "An ordered-subsets proximal preconditioned gradient algorithm for edge-preserving PET image reconstruction." In: *Med. Phys.* 40.5 (2013), p. 052503. DOI: 10.1118/1.4801898 (cit. on p. 2.30).

deman:05:ggp [169] B. De Man and S. Basu. "Generalized Geman prior for iterative reconstruction." In: *14th Intl. Conf. Medical Physics, Nuremberg, Germany*. 2005 (cit. on pp. 2.30, 2.31).

iatrou:06:acb [170] M. Iatrou, B. De Man, and S. Basu. "A comparison between filtered backprojection, post-smoothed weighted least squares, and penalized weighted least squares for CT reconstruction." In: *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.* Vol. 5. 2006, 2845–50. DOI: 10.1109/NSSMIC.2006.356470 (cit. on p. 2.30).

iatrou:07:a3s [171] M. Iatrou et al. "A 3D study comparing filtered backprojection, weighted least squares, and penalized weighted least squares for CT reconstruction." In: *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.* Vol. 4. 2007, 2639–43. DOI: 10.1109/NSSMIC.2007.4436689 (cit. on p. 2.30).

geman:87:smf [172] S. Geman and D. E. McClure. "Statistical methods for tomographic image reconstruction." In: *Proc. 46 Sect. ISI, Bull. ISI* 52.4 (1987), 5–21. URL: http://www.dam.brown.edu/people/geman/Homepage/Image%20processing,%20image%20analysis,%20Markov%20random%20fields,%20and%20MCMC/1987GemanMcClureBulletinISI.pdf (cit. on pp. 2.30, 2.31, 2.34).

thibault:07:atd [173] J-B. Thibault et al. "A three-dimensional statistical approach to improved image quality for multi-slice helical CT." In: *Med. Phys.* 34.11 (Nov. 2007), 4526–44. DOI: 10.1118/1.2789499 (cit. on pp. 2.30, 2.31).

geman:86:bia [174] D. Geman and S. Geman. "Bayesian image analysis." In: *Disorded systems and biological organization*. Ed. by G. Wiesbuch E Bienenstock F. Fogelman. ? 1986, F20 (cit. on p. 2.31).

abramowitz:64 [175] M. Abramowitz and I. A. Stegun. *Handbook of mathematical functions*. New York: Dover, 1964 (cit. on p. 2.33).

nikolova:98:iol [176] M. Nikolova, J. Idier, and A. Mohammad-Djafari. "Inversion of large-support ill-posed linear operators using a piecewise Gaussian MRF." In: *IEEE Trans. Im. Proc.* 7.4 (Apr. 1998), 571–85. DOI: 10.1109/83.663502 (cit. on p. 2.40).

nikolova:99:mru [177] M. Nikolova. "Markovian reconstruction using a GNC approach." In: *IEEE Trans. Im. Proc.* 8.9 (Sept. 1999), 1204–20. DOI: 10.1109/83.784433 (cit. on p. 2.40).

nikolova:00:lsh [178] M. Nikolova. "Local strong homogeneity of a regularized estimator." In: *SIAM J. Appl. Math.* 61.2 (2000), 633–58. DOI: 10.1137/S0036139997327794 (cit. on p. 2.40).

nikolova:00:tib [179] M. Nikolova. "Thresholding implied by truncated quadratic regularization." In: *IEEE Trans. Sig. Proc.* 48.12 (Dec. 2000), 3437–50. DOI: 10.1109/78.887035 (cit. on p. 2.40).

nikolova:10:fnn [180] M. Nikolova, M. K. Ng, and C-P. Tam. "Fast nonconvex nonsmooth minimization methods for image restoration and reconstruction." In: *IEEE Trans. Im. Proc.* 19.12 (Dec. 2010), 3073–88. DOI: 10.1109/TIP.2010.2052275 (cit. on p. 2.40).

valenzuela:08:reo [181] J. Valenzuela and J. A. Fessler. "Regularized estimation of Stokes images from polarimetric measurements." In: *Proc. SPIE 6814 Computational Imaging VI*. 2008, 681403:1–10. DOI: 10.1117/12.777882 (cit. on p. 2.40).

valenzuela:09:jro [182] J. Valenzuela and J. A. Fessler. "Joint reconstruction of Stokes images from polarimetric measurements." In: *J. Opt. Soc. Am. A* 26.4 (Apr. 2009), 962–8. DOI: 10.1364/JOSAA.26.000962 (cit. on p. 2.40).

he:08:rra
[183] X. He, J. A. Fessler, and E. C. Frey. "Regularized reconstruction algorithms for dual-isotope myocardial perfusion SPECT (MPS) imaging using a cross-tracer edge-preserving prior." In: *J. Nuc. Med. (Abs. Book)* 49.s1 (2008), p. 152. URL: http://jnumedmtg.snmjournals.org/cgi/content/meeting_abstract/49/MeetingAbstracts_1/152P-a (cit. on p. 2.40).

he:11:rir
[184] X. He et al. "Regularized image reconstruction algorithms for dual-isotope myocardial perfusion SPECT (MPS) imaging using a cross-tracer edge-preserving prior." In: *IEEE Trans. Med. Imag.* 30.6 (June 2011), 1169–83. DOI: 10.1109/TMI.2010.2087031 (cit. on p. 2.40).

weisenseel:02:sbf
[185] R. A. Weisenseel, W. C. Karl, and R. C. Chan. "Shared-boundary fusion for estimation of noisy multi-modality atherosclerotic plaque imagery." In: *Proc. IEEE Intl. Conf. on Image Processing*. Vol. 3. 2002, 157–60. DOI: 10.1109/ICIP.2002.1038929 (cit. on pp. 2.40, 2.41, 2.42).

holt:14:tnv
[186] K. M. Holt. "Total nuclear variation and Jacobian extensions of total variation for vector fields." In: *IEEE Trans. Im. Proc.* 23.9 (Sept. 2014), 3975–89. DOI: 10.1109/TIP.2014.2332397 (cit. on p. 2.41).

rigie:14:agv
[187] D. S. Rigie and P. J. La Rivière. "A generalized vectorial total-variation for spectral CT reconstruction." In: *Proc. 3rd Intl. Mtg. on image formation in X-ray CT*. 2014, 9–12 (cit. on p. 2.41).

blomgren:98:ctt
[188] P. Blomgren and T. F. Chan. "Color TV: total variation methods for restoration of vector-valued images." In: *IEEE Trans. Im. Proc.* 7.3 (Mar. 1998), 304–9. DOI: 10.1109/83.661180 (cit. on p. 2.41).

keren:98:dci
[189] D. Keren and A. Gotlib. "Denoising color images using regularization and correlation terms." In: *J. Visual Comm. Im. Rep.* 9.4 (Dec. 1998), 352–65. DOI: 10.1006/jvci.1998.0392 (cit. on p. 2.41).

wu:10:alm
[190] C. Wu and X-C. Tai. "Augmented Lagrangian method, dual methods, and split Bregman iteration for ROF, vectorial TV, and high order models." In: *SIAM J. Imaging Sci.* 3.3 (2010), 300–39. DOI: 10.1137/090767558 (cit. on p. 2.41).

holt:11:aro
[191] K. M. Holt. "Angular regularization of vector-valued signals." In: *Proc. IEEE Conf. Acoust. Speech Sig. Proc.* 2011, 1105–08. DOI: 10.1109/ICASSP.2011.5946601 (cit. on p. 2.41).

goldluecke:12:tnv
[192] B. Goldluecke, E. Strekalovskiy, and D. Cremers. "The natural vectorial total variation which arises from geometric measure theory." In: *SIAM J. Imaging Sci.* 5.2 (2012), 537–63. DOI: 10.1137/110823766 (cit. on p. 2.41).

strekalovskiy:14:cro
[193] E. Strekalovskiy, A. Chambolle, and D. Cremers. "Convex relaxation of vectorial problems with coupled regularization." In: *SIAM J. Imaging Sci.* 7.1 (2014), 294–336. DOI: 10.1137/130908348 (cit. on p. 2.41).

cotter:05:sst
[194] S. F. Cotter et al. "Sparse solutions to linear inverse problems with multiple measurement vectors." In: *IEEE Trans. Sig. Proc.* 53.7 (July 2005), 2477–88. DOI: 10.1109/TSP.2005.849172 (cit. on p. 2.41).

malioutov:05:ass
[195] D. Malioutov, M. Cetin, and A. S. Willsky. "A sparse signal reconstruction perspective for source localization with sensor arrays." In: *IEEE Trans. Sig. Proc.* 53.8 (Aug. 2005), 3010–3022. DOI: 10.1109/TSP.2005.850882 (cit. on p. 2.41).

duarte:05:dcs
[196] M. F. Duarte et al. "Distributed compressed sensing of jointly sparse signals." In: *Proc., IEEE Asilomar Conf. on Signals, Systems, and Comp.* 2005, 1537–41. DOI: 10.1109/ACSSC.2005.1600024 (cit. on p. 2.41).

tropp:05:ssa
[197] J. A. Tropp, A. C. Gilbert, and M. J. Strauss. "Simultaneous sparse approximation via greedy pursuit." In: *Proc. IEEE Conf. Acoust. Speech Sig. Proc.* Vol. 5. 2005, 721–4. DOI: 10.1109/ICASSP.2005.1416405 (cit. on p. 2.41).

tropp:06:afs-1
[198] J. A. Tropp, A. C. Gilbert, and M. J. Strauss. "Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit." In: *Signal Processing* 86.3 (Mar. 2006), 572–88. DOI: 10.1016/j.sigpro.2005.05.030 (cit. on p. 2.41).

tropp:06:afs-2
[199] J. A. Tropp. "Algorithms for simultaneous sparse approximation, Part II: convex relaxation." In: *Signal Processing* 86.3 (Mar. 2006), 589–602. DOI: 10.1016/j.sigpro.2005.05.031 (cit. on p. 2.41).

bach:08:cot
[200] F. R. Bach. "Consistency of the group lasso and multiple kernel learning." In: *J. Mach. Learning Res.* 9 (June 2008), 1179–225. URL: http://www.jmlr.org/papers/v9/bach08b.html (cit. on p. 2.41).

ji:09:mcs
[201] S. Ji, D. Dunson, and L. Carin. "Multitask compressive sensing." In: *IEEE Trans. Sig. Proc.* 57.1 (Jan. 2009), 92–106. DOI: 10.1109/TSP.2008.2005866 (cit. on p. 2.41).

vandenberg:10:tae
[202] E. van den Berg and M. P. Friedlander. "Theoretical and empirical results for recovery from multiple measurements." In: *IEEE Trans. Info. Theory* 56.5 (May 2010), 2516–27. DOI: 10.1109/TIT.2010.2043876 (cit. on p. 2.41).

ziniel:13:ehd
[203] J. Ziniel and P. Schniter. "Efficient high-dimensional inference in the multiple measurement vector problem." In: *IEEE Trans. Sig. Proc.* 61.2 (Jan. 2013), 340–54. DOI: 10.1109/TSP.2012.2222382 (cit. on p. 2.41).

baron:05:dcs
[204] D. Baron et al. *Distributed compressed sensing*. 2005. URL: http://www.dsp.ece.rice.edu/cs/DCS112005.pdf (cit. on p. 2.41).

baron:09:dcs [205] D. Baron et al. *Distributed compressive sensing*. arxiv 0901.3403. 2009. URL: http://arxiv.org/abs/0901.3403 (cit. on p. 2.41).

sebastiani:97:otu [206] G. Sebastiani and F. Godtliebsen. "On the use of Gibbs priors for Bayesian image restoration." In: *Signal Processing* 56.1 (Jan. 1997), 111–8. DOI: 10.1016/S0165-1684(97)00002-9 (cit. on p. 2.42).

olafsson:06:sra [207] V. Olafsson, J. A. Fessler, and D. C. Noll. "Spatial resolution analysis of iterative image reconstruction with separate regularization of real and imaginary parts." In: *Proc. IEEE Intl. Symp. Biomed. Imag.* 2006, 5–8. DOI: 10.1109/ISBI.2006.1624838 (cit. on p. 2.42).

hoge:07:frr [208] W. S. Hoge et al. "Fast regularized reconstruction of non-uniformly subsampled partial-Fourier parallel MRI data." In: *Proc. IEEE Intl. Symp. Biomed. Imag.* 2007, 1012–5. DOI: 10.1109/ISBI.2007.357026 (cit. on p. 2.42).

chen:10:rod [209] L. Chen, M. C. Schabel, and E. V. R. DiBella. "Reconstruction of dynamic contrast enhanced magnetic resonance imaging of the breast with temporal constraints." In: *Mag. Res. Im.* 28.5 (June 2010), 637–45. DOI: 10.1016/j.mri.2010.03.001 (cit. on p. 2.42).

cetin:01:fes [210] M. Cetin and W. C. Karl. "Feature-enhanced synthetic aperture radar image formation based on nonquadratic regularization." In: *IEEE Trans. Im. Proc.* 10.4 (Apr. 2001), 623–31. DOI: 10.1109/83.913596 (cit. on p. 2.42).

fessler:04:iir [211] J. A. Fessler and D. C. Noll. "Iterative image reconstruction in MRI with separate magnitude and phase regularization." In: *Proc. IEEE Intl. Symp. Biomed. Imag.* 2004, 209–12. DOI: 10.1109/ISBI.2004.1398511 (cit. on p. 2.42).

zibetti:10:sma [212] M. V. W. Zibetti and A. R. D. Pierro. "Separate magnitude and phase regularization in MRI with incomplete data: preliminary results." In: *Proc. IEEE Intl. Symp. Biomed. Imag.* 2010, 0736–9. DOI: 10.1109/ISBI.2010.5490069 (cit. on p. 2.42).

zhao:11:sma [213] F. Zhao et al. "Separate magnitude and phase regularization via compressed sensing." In: *Proc. Intl. Soc. Mag. Res. Med.* 2011, p. 2841. URL: http://cds.ismrm.org/protected/11MProceedings/files/2841.pdf (cit. on p. 2.42).

zhao:12:sma [214] F. Zhao et al. "Separate magnitude and phase regularization via compressed sensing." In: *IEEE Trans. Med. Imag.* 31.9 (Sept. 2012), 1713–23. DOI: 10.1109/TMI.2012.2196707 (cit. on p. 2.42).

funai:08:rfm [215] A. K. Funai et al. "Regularized field map estimation in MRI." In: *IEEE Trans. Med. Imag.* 27.10 (Oct. 2008), 1484–94. DOI: 10.1109/TMI.2008.923956 (cit. on p. 2.42).

lu:98:crw [216] H. H-S. Lu, C-M. Chen, and I-H. Yang. "Cross-reference weighted least square estimates for positron emission tomography." In: *IEEE Trans. Med. Imag.* 17.1 (Feb. 1998), 1–8. DOI: 10.1109/42.668690 (cit. on p. 2.42).

zeng:07:dta [217] K. Zeng et al. "Digital tomosynthesis aided by low-resolution exact computed tomography." In: *J. Comp. Assisted Tomo.* 31.6 (Nov. 2007), 976–83. DOI: 10.1097/rct.0b013e31803e8c1f. URL: http://gateway.ovid.com/ovidweb.cgi?T=JS&MODE=ovid&NEWS=n&PAGE=toc&D=ovft&AN=00004728-200711000-00024 (cit. on p. 2.42).

tu:01:eso [218] K-Y. Tu et al. "Empirical studies of cross-reference maximum likelihood estimate reconstruction for positron emission tomography." In: *Biomed. Engin. - Appl., Basis and Commun.* 13.1 (Feb. 2001), 1–7. DOI: 10.4015/S1016237201000029 (cit. on p. 2.42).

chen:91:iop [219] C. T. Chen et al. "Improvement of PET image reconstruction using high-resolution anatomic images." In: *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.* Vol. 3. (Abstract.) 1991, p. 2062. DOI: 10.1109/NSSMIC.1991.259320 (cit. on p. 2.42).

chen:91:ios [220] C. T. Chen et al. "Incorporation of structural CT and MR images in PET image reconstruction." In: *Proc. SPIE 1445 Med. Im. V: Im. Proc.* 1991, 222–5. DOI: 10.1117/12.45219 (cit. on p. 2.42).

chen:91:sfi [221] C. T. Chen et al. "Sensor fusion in image reconstruction." In: *IEEE Trans. Nuc. Sci.* 38.2 (Apr. 1991), 687–92. DOI: 10.1109/23.289375 (cit. on p. 2.42).

leahy:91:sma [222] R. Leahy and X. H. Yan. "Statistical models and methods for PET image reconstruction." In: *Proc. of Stat. Comp. Sect. of Amer. Stat. Assoc.* 1991, 1–10 (cit. on p. 2.42).

leahy:91:ioa [223] R. Leahy and X. H. Yan. "Incorporation of anatomical MR data for improved functional imaging with PET." In: *Information Processing in Medical Im.* LNCS 511. 1991, 105–20. DOI: 10.1007/BFb0033746 (cit. on p. 2.42).

yan:91:mir [224] X. H. Yan and R. Leahy. "MAP image reconstruction using intensity and line processes for emission tomography data." In: *Proc. SPIE 1452 Im. Proc. Alg. and Tech. II*. 1991, 158–69. DOI: 10.1117/12.45380 (cit. on p. 2.42).

yan:92:meo [225] X. H. Yan et al. "MAP Estimation of PET Images using prior anatomical information from MR scans." In: *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.* Vol. 2. 1992, 1201–3. DOI: 10.1109/NSSMIC.1992.301033 (cit. on p. 2.42).

gindi:93:bro
[226] G. Gindi et al. "Bayesian reconstruction of functional images using anatomical information as priors." In: *IEEE Trans. Med. Imag.* 12.4 (Dec. 1993), 670–80. DOI: 10.1109/42.251117 (cit. on p. 2.42).

johnson:93:aff
[227] V. Johnson. "A framework for incorporating structural prior information into the estimation of medical images." In: *Information Processing in Medical Im.* Ed. by H H Barrett and A F Gmitro. Berlin: Springer Verlag, 1993, pp. 307–21 (cit. on p. 2.42).

zhou:93:acs
[228] Z. Zhou, R. M. Leahy, and E. U. Mumcuoglu. "A comparative study of the effect of using anatomical priors in PET reconstruction." In: *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.* Vol. 3. 1993, 1749–53. DOI: 10.1109/NSSMIC.1993.373592 (cit. on p. 2.42).

bowsher:96:bra
[229] J. E. Bowsher et al. "Bayesian reconstruction and use of anatomical a priori information for emission tomography." In: *IEEE Trans. Med. Imag.* 15.5 (Oct. 1996), 673–86. DOI: 10.1109/42.538945 (cit. on p. 2.42).

wang:04:dae
[230] C-H. Wang, J-C. Chen, and R-S. Liu. "Development and evaluation of MRI-based Bayesian image reconstruction methods for PET." In: *Computerized Medical Imaging and Graphics* 28.4 (June 2004), 177–84. DOI: 10.1016/j.compmedimag.2003.11.005 (cit. on p. 2.42).

fessler:92:rei
[231] J. A. Fessler, N. H. Clinthorne, and W. L. Rogers. "Regularized emission image reconstruction using imperfect side information." In: *IEEE Trans. Nuc. Sci.* 39.5 (Oct. 1992), 1464–71. DOI: 10.1109/23.173225 (cit. on p. 2.42).

zubal:92:bro
[232] I. G. Zubal et al. "Bayesian reconstruction of SPECT images using registered anatomical images as priors." In: *J. Nuc. Med. (Abs. Book)* 33.5 (May 1992), p. 963 (cit. on p. 2.42).

lipinski:97:emr
[233] B. Lipinski et al. "Expectation maximization reconstruction of positron emission tomography images using anatomical magnetic resonance information." In: *IEEE Trans. Med. Imag.* 16.2 (Apr. 1997), 129–36. DOI: 10.1109/42.563658 (cit. on p. 2.42).

piramuthu:98:sia
[234] R. Piramuthu and A. O. Hero. "Side information averaging method for PML emission tomography." In: *Proc. IEEE Intl. Conf. on Image Processing*. Vol. 2. 1998, 671–5. DOI: 10.1109/ICIP.1998.723614 (cit. on p. 2.42).

hero:99:mec
[235] A. O. Hero et al. "Minimax emission computed tomography using high resolution anatomical side information and B-spline models." In: *IEEE Trans. Info. Theory* 45.3 (Apr. 1999), 920–38. DOI: 10.1109/18.761333 (cit. on p. 2.42).

comtat:02:cfr
[236] C. Comtat et al. "Clinically feasible reconstruction of 3d whole-body PET/CT data using blurred anatomical labels." In: *Phys. Med. Biol.* 47.1 (Jan. 2002), 1–20. DOI: 10.1088/0031-9155/47/1/301 (cit. on p. 2.42).

maddjafari:02:fox-icip
[237] A. Mohammad-Djafari. "Fusion of x-ray radiographic data and anatomical data in computed tomography." In: *Proc. IEEE Intl. Conf. on Image Processing*. Vol. 2. 2002, 461–64. DOI: 10.1109/ICIP.2002.1039987 (cit. on p. 2.42).

guven:05:dot
[238] M. Guven et al. "Diffuse optical tomography with a priori anatomical information." In: *Phys. Med. Biol.* 50.12 (June 2005), 2837–2858. DOI: 10.1088/0031-9155/50/12/008 (cit. on p. 2.42).

nuyts:05:cbm
[239] J. Nuyts et al. "Comparison between MAP and postprocessed ML for image reconstruction in emission tomography when anatomical knowledge is available." In: *IEEE Trans. Med. Imag.* 24.5 (May 2005), 667–75. DOI: 10.1109/TMI.2005.846850 (cit. on p. 2.42).

alessio:06:iqf
[240] A. M. Alessio and P. E. Kinahan. "Improved quantitation for PET/CT image reconstruction with system modeling and anatomical priors." In: *Med. Phys.* 33.11 (Nov. 2006), 4095–103. DOI: 10.1118/1.2358198 (cit. on p. 2.42).

boussion:06:ami
[241] N. Boussion et al. "A multiresolution image based approach for correction of partial volume effects in emission tomography." In: *Phys. Med. Biol.* 51.7 (Apr. 2006), 1857–76. DOI: 10.1088/0031-9155/51/7/016 (cit. on p. 2.42).

sastry:97:mba
[242] S. Sastry and R. E. Carson. "Multimodality Bayesian algorithm for image reconstruction in positron emission tomography: a tissue composition model." In: *IEEE Trans. Med. Imag.* 16.6 (Dec. 1997), 750–61. DOI: 10.1109/42.650872 (cit. on p. 2.42).

hsu:00:bef
[243] C-H. Hsu. "Bayesian estimator for positron emission tomography imaging using a prior image model with mixed continuity constraints." In: *J. Electronic Imaging* 9.3 (July 2000), 260–8. DOI: 10.1117/1.482744 (cit. on p. 2.42).

rangarajan:00:abj
[244] A. Rangarajan, I-T. Hsiao, and G. Gindi. "A Bayesian joint mixture framework for the integration of anatomical information in functional image reconstruction." In: *J. Math. Im. Vision* 12.3 (June 2000), 199–217. DOI: 10.1023/A:1008314015446 (cit. on p. 2.42).

chiao:94:mbe-2
[245] P. C. Chiao et al. "Model-based estimation with boundary side information or boundary regularization." In: *IEEE Trans. Med. Imag.* 13.2 (June 1994), 227–34. DOI: 10.1109/42.293915 (cit. on p. 2.42).

bowsher:06:aet
[246] J. E. Bowsher et al. "Aligning emission tomography and MRI images by optimizing the emission-tomography image reconstruction objective function." In: *IEEE Trans. Nuc. Sci.* 53.3 (June 2006), 1248–58. DOI: 10.1109/TNS.2006.875467 (cit. on p. 2.42).

snyder:84:usi
[247] D. L. Snyder. "Utilizing side information in emission tomography." In: *IEEE Trans. Nuc. Sci.* 31.1 (Feb. 1984), 533–7. DOI: 10.1109/TNS.1984.4333313 (cit. on p. 2.42).

carson:85:aml
[248] R. E. Carson, M. V. Green, and S. M. Larson. "A maximum likelihood method for calculation of tomographic region-of-interest (ROI) values." In: *J. Nuc. Med. (Abs. Book)* 26.5 (1985), P20:71. URL: http://www.osti.gov/scitech/biblio/6957603 (cit. on p. 2.42).

carson:86:aml
[249] R. E. Carson. "A maximum likelihood method for region-of-interest evaluation in emission tomography." In: *J. Comp. Assisted Tomo.* 10.4 (July 1986), 654–63. URL: http://gateway.ovid.com/ovidweb.cgi?T=JS&MODE=ovid&NEWS=n&PAGE=toc&D=ovft&AN=00004728-198607000-00021 (cit. on p. 2.42).

glidewell:95:ace
[250] M. Glidewell and K. T. Ng. "Anatomically constrained electrical impedance tomography for anisotropic bodies via a two-step approach." In: *IEEE Trans. Med. Imag.* 14.3 (Sept. 1995), 498–503. DOI: 10.1109/42.414615 (cit. on p. 2.42).

ardekani:96:mce
[251] B. A. Ardekani et al. "Minimum cross-entropy reconstruction of PET images using prior anatomical information." In: *Phys. Med. Biol.* 41.11 (Nov. 1996), 2497–517. DOI: 10.1088/0031-9155/41/11/018 (cit. on p. 2.42).

som:98:pom
[252] S. Som, B. F. Hutton, and M. Braun. "Properties of minimum cross-entropy reconstruction of emission tomography with anatomically based prior." In: *IEEE Trans. Nuc. Sci.* 45.6 (Dec. 1998), 3014–21. DOI: 10.1109/23.737658 (cit. on p. 2.42).

somayajula:05:pir
[253] S. Somayajula, E. Asma, and R. M. Leahy. "PET image reconstruction using anatomical information through mutual information based priors." In: *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.* 2005, 2722–26. DOI: 10.1109/NSSMIC.2005.1596899 (cit. on p. 2.42).

somayajula:07:pir
[254] S. Somayajula, A. Rangarajan, and R. M. Leahy. "PET image reconstruction using anatomical information through mutual information based priors: A scale space approach." In: *Proc. IEEE Intl. Symp. Biomed. Imag.* 2007, 165–8. DOI: 10.1109/ISBI.2007.356814 (cit. on p. 2.42).

nuyts:07:tuo
[255] J. Nuyts. "The use of mutual information and joint entropy for anatomical priors in emission tomography." In: *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.* Vol. 6. 2007, 4194–54. DOI: 10.1109/NSSMIC.2007.4437034 (cit. on p. 2.42).

tang:08:bpi
[256] J. Tang, B. M. W. Tsui, and A. Rahmim. "Bayesian PET image reconstruction incorporating anato-functional joint entropy." In: *Proc. IEEE Intl. Symp. Biomed. Imag.* 2008, 1043–6. DOI: 10.1109/ISBI.2008.4541178 (cit. on p. 2.42).

kirov:08:pve
[257] A. S. Kirov, J. Z. Piao, and C. R. Schmidtlein. "Partial volume effect correction in PET using regularized iterative deconvolution with variance control based on local topology." In: *Phys. Med. Biol.* 53.10 (May 2008), 2577–92. DOI: 10.1088/0031-9155/53/10/009 (cit. on p. 2.42).

bruyant:04:nos
[258] P. P. Bruyant et al. "Numerical observer study of MAP-OSEM regularization methods with anatomical priors for lesion detection in $^{67}$Ga images." In: *IEEE Trans. Nuc. Sci.* 51.1 (Feb. 2004), 193–7. DOI: 10.1109/TNS.2003.823050 (cit. on p. 2.42).

hart:87:bip
[259] H. Hart and Z. Liang. "Bayesian image processing in two dimensions." In: *IEEE Trans. Med. Imag.* 6.3 (Sept. 1987), 201–8. DOI: 10.1109/TMI.1987.4307828 (cit. on p. 2.43).

liang:91:srs
[260] Z. Liang et al. "Simultaneous reconstruction, segmentation, and edge enhancement of relatively piecewise continuous images with intensity-level information." In: *Med. Phys.* 18.3 (May 1991), 394–401. DOI: 10.1118/1.596685 (cit. on p. 2.43).

choi:91:pvt
[261] H. S. Choi, D. R. Haynor, and Y. Kim. "Partial volume tissue classification of multichannel magnetic resonance images—A mixel model." In: *IEEE Trans. Med. Imag.* 10.3 (Sept. 1991), 395–407. DOI: 10.1109/42.97590 (cit. on p. 2.43).

nuyts:99:sma
[262] J. Nuyts et al. "Simultaneous maximum a-posteriori reconstruction of attenuation and activity distributions from emission sinograms." In: *IEEE Trans. Med. Imag.* 18.5 (May 1999), 393–403. DOI: 10.1109/42.774167 (cit. on p. 2.43).

lemmens:09:som
[263] C. Lemmens, D. Faul, and J. Nuyts. "Suppression of metal artifacts in CT using a reconstruction procedure that combines MAP and projection completion." In: *IEEE Trans. Med. Imag.* 28.2 (Feb. 2009), 250–60. DOI: 10.1109/TMI.2008.929103 (cit. on p. 2.43).

buades:05:aro
[264] A. Buades, B. Coll, and J. M. Morel. "A review of image denoising methods, with a new one." In: *SIAM Multiscale Modeling and Simulation* 4.2 (2005), 490–530. DOI: 10.1137/040616024 (cit. on p. 2.43).

mignotte:08:anl
[265] M. Mignotte. "A non-local regularization strategy for image deconvolution." In: *Pattern Recognition Letters* 29.16 (Dec. 2008), 2206–12. DOI: 10.1016/j.patrec.2008.08.004 (cit. on p. 2.44).

rousseau:10:anl
[266] F. Rousseau. "A non-local approach for image super-resolution using intermodality priors." In: *Med. Im. Anal.* 14.4 (Aug. 2010), 594–605. DOI: 10.1016/j.media.2010.04.005 (cit. on p. 2.44).

kindermann:05:dad
[267] S. Kindermann, S. Osher, and P. W. Jones. "Deblurring and denoising of images by nonlocal functionals." In: *SIAM Multiscale Modeling and Simulation* 4.4 (2005), 1091–115. DOI: 10.1137/050622249 (cit. on p. 2.44).

bougleux:08:nlr
[268] S. Bougleux, G. Peyré, and L. Cohen. "Non-local regularization of inverse problems." In: *ECCV*. Vol. III. LNCS 5304, Springer-Verlag. 2008, 57–68. DOI: 10.1007/978-3-540-88690-7_5 (cit. on p. 2.44).

protter:09:gtn
[269] M. Protter et al. "Generalizing the nonlocal-means to super-resolution reconstruction." In: *IEEE Trans. Im. Proc.* 18.1 (Jan. 2009), 36–51. DOI: 10.1109/TIP.2008.2008067 (cit. on p. 2.44).

vaksman:16:poa
[270] G. Vaksman, M. Zibulevsky, and M. Elad. "Patch ordering as a regularization for inverse problems in image processing." In: *SIAM J. Imaging Sci.* 9.1 (2016), 287–319. DOI: 10.1137/15M1038074 (cit. on p. 2.44).

zhang:17:aon
[271] H. Zhang et al. "Applications of nonlocal means algorithm in low-dose X-ray CT image processing and reconstruction: a review." In: *Med. Phys.* (2017). DOI: 10.1002/mp.12097 (cit. on p. 2.44).

lingenfelter:09:srf
[272] D. J. Lingenfelter, J. A. Fessler, and Z. He. "Sparsity regularization for image reconstruction with Poisson data." In: *Proc. SPIE 7246 Computational Imaging VII*. 2009, 72460F. DOI: 10.1117/12.816961 (cit. on p. 2.48).

nikolova:13:dot
[273] M. Nikolova. "Description of the minimizers of least squares regularized with $\ell_0$-norm. uniqueness of the global minimizer." In: *SIAM J. Imaging Sci.* 6.2 (2013), 904–37. DOI: 10.1137/11085476X (cit. on p. 2.48).

chun:13:npo
[274] S. Y. Chun and J. A. Fessler. "Noise properties of motion-compensated tomographic image reconstruction methods." In: *IEEE Trans. Med. Imag.* 32.2 (Feb. 2013), 141–52. DOI: 10.1109/TMI.2012.2206604 (cit. on p. 2.48).

breiman:95:bsr
[275] L. Breiman. "Better subset regression using the nonnegative garrote." In: *Technometrics* 37.4 (Nov. 1995), 373–84. URL: http://www.jstor.org/stable/1269730 (cit. on p. 2.49).

gao:98:wsd
[276] H. Gao. "Wavelet shrinkage denoising using the nonnegative garrote." In: *J. Computational and Graphical Stat.* 7 (1998), 469–88. URL: http://www.jstor.org/stable/1390677 (cit. on p. 2.49).

yuan:06:msa
[277] M. Yuan and Y. Lin. "Model selection and estimation in regression with grouped variables." In: *J. Royal Stat. Soc. Ser. B* 68.1 (2006), 49–67. DOI: 10.1111/j.1467-9868.2005.00532.x (cit. on p. 2.49).