Chapter 5

Analytical Regularization Design for Tomography

ch,tomo-design

Contents

s,tomo,design,intro

5.1 Introduction (s,tomo,design,intro)

This chapter describes some specialized methods for regularization design for tomographic image reconstruction. The methods in this chapter are based on the "analytical" approach to tomography described in Chapter 3 and Chapter 4. See Chapter 22 for further discussion of how regularization design affects spatial resolution.

s,tomo,design,2

5.2 Regularization design for isotropic spatial resolution (s,tomo,design,2)

In Chapter 4, §4.3.8 and §4.6.4 described the frequency responses of some "ideal" quadratic regularization methods that would lead to isotropic spatial resolution. However, the target frequency responses (4.3.23) and (4.6.23) are difficult to achieve using practical regularization methods based on the differences between neighboring pixel values. This section describes practical methods for designing regularizers that achieve approximately isotropic spatial resolution.

s,tomo,design,2d

5.2.1 2D case

As described in §1.10, for a discrete-space 2D object $f[m,n]$, a typical quadratic roughness penalty has the following form:

$$\mathsf{R}(f) = \sum_{n,m} \sum_{l=1}^{L} r_l \frac{1}{2} \left| (c_l ** f)[n,m] \right|^2,$$

where the r_l values are design parameters that affect the directionality of the regularization and hence the shape of the PSF, and each $c_l[m,n]$ is a (typically) high-pass filter. This section describes a method for choosing the r_l values to achieve approximately isotropic spatial resolution in 2D [1, 2].

For regularizers using **first-order finite differences**:

$$c_l[m,n] = \frac{1}{\sqrt{n_l^2 + m_l^2}} \left(\delta_2[m,n] - \delta_2[m - m_l, n - n_l] \right),$$

whereas for **2nd-order finite differences**:

$$c_l[m,n] = \left(\frac{1}{\sqrt{n_l^2 + m_l^2}} \right)^2 \left(\delta_2[m,n] - \delta_2[m - m_l, n - n_l] \right) ** \left(\delta_2[m,n] - \delta_2[m + m_l, n + n_l] \right).$$

For example, for first-order differences using the nearest horizontal, vertical, and diagonal neighbors, we usually use[1]

$$\begin{aligned}
c_1[m,n] &= \delta_2[m,n] - \delta_2[m-1,n]] \\
c_2[m,n] &= \delta_2[m,n] - \delta_2[m,n-1] \\
c_3[m,n] &= \frac{1}{\sqrt{2}} (\delta_2[m,n] - \delta_2[m-1,n-1]) \\
c_4[m,n] &= \frac{1}{\sqrt{2}} (\delta_2[m,n] - \delta_2[m-1,n+1]) \,.
\end{aligned} \tag{5.2.1}$$

e,tomo,blur,iso,p4

To generalize (4.3.17) for such regularizers, we move to the frequency domain by applying **Parseval's theorem**:

$$\mathsf{R}(f) = \sum_{l=1}^{L} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \frac{1}{2} r_l \left| C_l(\Omega_1, \Omega_2) \, F(\Omega_1, \Omega_2) \right|^2 \frac{\mathrm{d}\Omega_1}{2\pi} \frac{\mathrm{d}\Omega_2}{2\pi},$$

where the Fourier transform of a first-order $c_l[m,n]$ has the following approximation:

$$\begin{aligned}
|C_l(\Omega_1, \Omega_2)|^2 &= \frac{1}{n_l^2 + m_l^2} \left| 1 - \mathrm{e}^{-\imath(\Omega_1 n_l + \Omega_2 m_l)} \right|^2 = \frac{1}{n_l^2 + m_l^2} 4 \sin^2 \left(\frac{\Omega_1 n_l + \Omega_2 m_l}{2} \right) \\
&\approx \frac{1}{n_l^2 + m_l^2} \left(\Omega_1 n_l + \Omega_2 m_l \right)^2 .
\end{aligned}$$

More generally, if the lth term is a M_lth order difference, where $M_l \in \mathbb{N}$, then

$$|C_l(\Omega_1, \Omega_2)|^2 \approx \left(\frac{1}{n_l^2 + m_l^2} \right)^{M_l} \left(\Omega_1 n_l + \Omega_2 m_l \right)^{2M_l} .$$

By the frequency relationships associated with **sampling**, $\Omega_1 = 2\pi \triangle_{\mathrm{x}} u = 2\pi \triangle_{\mathrm{x}} \rho \cos \Phi$ and $\Omega_2 = 2\pi \triangle_{\mathrm{Y}} \rho \sin \Phi$. Thus

$$\begin{aligned}
|C_l(\Omega_1, \Omega_2)|^2 &\approx \left(\frac{1}{n_l^2 + m_l^2} \right)^{M_l} (2\pi \triangle_{\mathrm{x}} \rho n_l \cos \Phi + 2\pi \triangle_{\mathrm{Y}} \rho m_l \sin \Phi)^{2M_l} \\
&= \left(\frac{1}{n_l^2 + m_l^2} \right)^{M_l} (2\pi\rho)^{2M_l} (n_l \cos \Phi + m_l \sin \Phi)^{2M_l} \\
&= (2\pi\rho)^{2M_l} \cos^{2M_l}(\Phi - \varphi_l),
\end{aligned}$$

where

$$\varphi_l \triangleq \tan^{-1} \frac{m_l}{n_l}$$

and we assume $\triangle_{\mathrm{Y}} = \triangle_{\mathrm{x}} = 1$ hereafter for simplicity. For such practical regularization methods, the frequency domain expression (4.3.10) for the regularization operator becomes $\mathsf{R}(f) = \frac{1}{2} \langle \boldsymbol{\mathcal{R}} f, \, f \rangle$, with

$$\boldsymbol{\mathcal{R}} = \boldsymbol{\mathcal{F}}_2^{-1} \boldsymbol{\mathcal{D}}(R(\rho, \Phi; \boldsymbol{r})) \, \boldsymbol{\mathcal{F}}_2$$

where $\boldsymbol{r} = (r_1, \ldots, r_L)$ and

$$R(\rho, \Phi; \boldsymbol{r}) \approx \sum_{l=1}^{L} r_l (2\pi\rho)^{2M_l} \cos^{2M_l}(\Phi - \varphi_l) \,. \tag{5.2.2}$$

e,tomo,design,R(p,q,r)

Similarly, the frequency response expression (4.3.17) for regularized WLS tomographic image reconstruction becomes

$$L(\rho, \Phi) \approx L(\rho, \Phi; \boldsymbol{r}) \triangleq \frac{w(\Phi) \left| B_\Phi(\rho) \right|^2}{w(\Phi) \left| B_\Phi(\rho) \right|^2 + \beta \left| \rho \right| R(\rho, \Phi; \boldsymbol{r})}.$$

To achieve isotropic spatial resolution, we want to choose the regularization coefficients $\boldsymbol{r}$ so that the above expression is approximately independent of Φ. We could even go one step further and try to choose $\boldsymbol{r}$ to so that $L(\rho, \Phi; \boldsymbol{r})$ approximates some desired frequency response shape, such as that corresponding to a gaussian bell PSF. Our practical experience is that arbitrary isotropic PSF shapes can be difficult to achieve with practical penalty neighborhood sizes. A realistic goal is to approximate the shape of the frequency response associated with *unweighted* regularized least-squares reconstruction. So our goal is to choose $\boldsymbol{r}$ such that

$$L(\rho, \Phi; \boldsymbol{r}) = \frac{w(\Phi) \left| B_\Phi(\rho) \right|^2}{w(\Phi) \left| B_\Phi(\rho) \right|^2 + \beta \left| \rho \right| R(\rho, \Phi; \boldsymbol{r})} \approx L_0(\rho, \Phi) \triangleq \frac{|B(\rho)|^2}{|B(\rho)|^2 + \beta \left| \rho \right| |2\pi\rho|^{2M_{\mathrm{R}}}},$$

footnote,sqrt2

[1] Note that the diagonal differences are divided by $\sqrt{2}$, so the penalty includes terms of the form $\left| \frac{f[m,n] - f[m-1,n-1]}{\sqrt{2}} \right|^2$. This differs from the common practice [3, 4] of dividing the *square* of the diagonal differences by $\sqrt{2}$, *i.e.*, $\frac{1}{\sqrt{2}} |f[m,n] - f[m-1,n-1]|^2$. The analysis of this section suggests that the common practice is suboptimal in terms of producing isotropic spatial resolution.

where $B(\rho)$ is some Φ-independent detector response, such as the angular average response

$$B(\rho) = \frac{1}{\pi} \int_0^\pi B_\Phi(\rho) \, \mathrm{d}\Phi \,.$$

There are many possible approaches to choosing $\boldsymbol{r}$. A particularly simple approach is based on a continuous-space extension of [5]. Cross multiplying, we want

$$w(\Phi) \left|B_\Phi(\rho)\right|^2 \left(\left|B(\rho)\right|^2 + \beta \left|\rho\right| \left|2\pi\rho\right|^{2M_\mathrm{R}} \right) \approx \left|B(\rho)\right|^2 \left(w(\Phi) \left|B_\Phi(\rho)\right|^2 + \beta \left|\rho\right| R(\rho, \Phi; \boldsymbol{r}) \right),$$

or equivalently

$$w(\Phi) \left|2\pi\rho\right|^{2M_\mathrm{R}} \left|B_\Phi(\rho)\right|^2 \approx \left|B(\rho)\right|^2 R(\rho, \Phi; \boldsymbol{r}).$$

Hereafter we choose $M_l = M_\mathrm{R}$, in which case our goal becomes to choose $\boldsymbol{r}$ such that

$$w(\varphi) \left|B_\varphi(\rho)\right|^2 \approx \left|B(\rho)\right|^2 \sum_{l=1}^L r_l \cos^{2M_\mathrm{R}}(\varphi - \varphi_l) \,.$$

A natural approach is nonnegative weighted least-squares fitting of $\boldsymbol{r}$:

$$\hat{\boldsymbol{r}} = \arg\min_{\boldsymbol{r} \succeq \boldsymbol{0}} \frac{1}{\pi} \int_0^\pi \int_{-\infty}^\infty v(\rho) \frac{1}{2} \left(w(\varphi) \left|B_\varphi(\rho)\right|^2 - \left|B(\rho)\right|^2 \sum_{l=1}^L r_l \cos^{2M_\mathrm{R}}(\varphi - \varphi_l) \right)^2 \mathrm{d}\rho \, \mathrm{d}\varphi, \tag{5.2.3}$$

e,tomo,design,rh,int

where $v(\rho)$ is a user-selected weighting function. We allow these weights to depend on spatial frequency ρ because it is unclear whether high or low spatial frequencies will be more important for achieving isotropy. However, we do not allow v to depend on angle φ because intuition suggests that all angles should be equally important when isotropy is the goal. The nonnegativity constraint ensures that $\boldsymbol{\mathcal{R}}$ is positive-semidefinite.

Expanding the quadratic, integrating over ρ, and completing the square leads to the following simplified expression:

$$\hat{\boldsymbol{r}} = \arg\min_{\boldsymbol{r} \succeq \boldsymbol{0}} \frac{1}{\pi} \int_0^\pi \frac{1}{2} \left(\bar{w}(\varphi) - \sum_{l=1}^L r_l \cos^{2M_\mathrm{R}}(\varphi - \varphi_l) \right)^2 \mathrm{d}\varphi,$$

where

$$\bar{w}(\varphi) \triangleq w(\varphi) \frac{\int_{-\infty}^\infty v(\rho) \left|B(\rho)\right|^2 \left|B_\varphi(\rho)\right|^2 \mathrm{d}\rho}{\int_{-\infty}^\infty v(\rho) \left|B(\rho)\right|^4 \mathrm{d}\rho}.$$

In the case of angularly-independent blur with $B_\varphi(\rho) = B(\rho)$, then we simply have $\bar{w}(\varphi) = w(\varphi)$.

A convenient choice for the weights is $v(\rho) = 1/\left|B(\rho)\right|^2$, in which case it follows from **Parseval's theorem** that

$$\bar{w}(\varphi) = w(\varphi) \frac{\int_{-\infty}^\infty \left|B_\varphi(\rho)\right|^2 \mathrm{d}\rho}{\int_{-\infty}^\infty \left|B(\rho)\right|^2 \mathrm{d}\rho} = w(\varphi) \frac{\int_{-\infty}^\infty \left|b_\varphi(r)\right|^2 \mathrm{d}r}{\int_{-\infty}^\infty \left|b(r)\right|^2 \mathrm{d}r}. \tag{5.2.4}$$

e,tomo,design2,wangb

We can approximate these integrals easily by finite sums in the realistic case of discrete radial samples. The denominator is just a constant so it is essentially irrelevant and can be absorbed into β.

For a shift-varying detector response and radial-dependent weighting, it follows from (4.4.3) and (4.4.8) that $\bar{w}(\varphi)$ in (5.2.4) becomes

$$\bar{w}(\varphi) = \frac{\int \left|b(r, \varphi; x_0, y_0)\right|^2 w_\varphi(r) \, \mathrm{d}r}{\int \left|b(r)\right|^2 \mathrm{d}r}. \tag{5.2.5}$$

e,tomo,design2,wangb,nuyts

This expression is remarkably similar to the weighting proposed by Nuyts in [4] using a quite different derivation.

Hereafter we focus on the first-order difference case where $M_\mathrm{R} = 1$, and expand the cosines in terms of a 3-term basis that is orthonormal with respect to the inner product $\langle f, g \rangle = \frac{1}{\pi} \int_0^\pi f(\varphi) g(\varphi) \, \mathrm{d}\varphi$ as follows:

$$\begin{aligned} \cos^2(\varphi - \varphi_l) &= \frac{1}{2} + \frac{1}{2} \cos(2\varphi - 2\varphi_l) = \frac{1}{2} + \frac{1}{2} \left[\cos(2\varphi_l) \cos(2\varphi) + \sin(2\varphi_l) \sin(2\varphi) \right] \\ &= \frac{1}{2} \cdot 1 + \frac{\cos(2\varphi_l)}{2\sqrt{2}} \left[\sqrt{2} \cos(2\varphi) \right] + \frac{\sin(2\varphi_l)}{2\sqrt{2}} \left[\sqrt{2} \sin(2\varphi) \right]. \end{aligned}$$

Using the second-order neighborhood (5.2.1), we find $\varphi_1 = 0, \ \varphi_2 = \pi/2, \ \varphi_3 = \pi/4, \ \varphi_4 = -\pi/4$, so the minimization problem (5.2.3) simplifies to

$$\hat{\boldsymbol{r}} = \arg\min_{\boldsymbol{r} \succeq \boldsymbol{0}} \Psi(\boldsymbol{r}), \qquad \Psi(\boldsymbol{r}) = \frac{1}{2} \left\| \boldsymbol{A} \boldsymbol{r} - \boldsymbol{w} \right\|_2^2, \tag{5.2.6}$$

e,tomo,design,rh

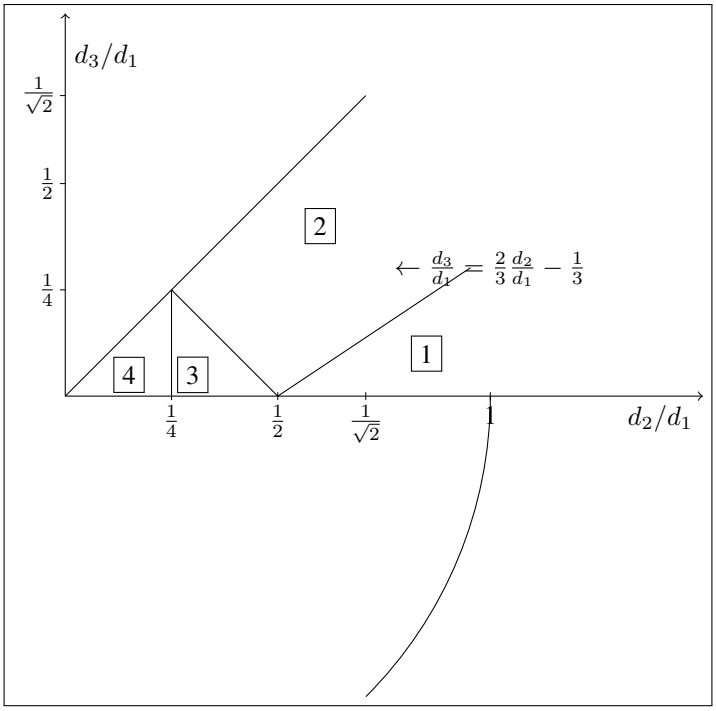

Figure 5.2.1: First octant of quadratic penalty design space showing the four regions where different constraints are active. fig,tomo,design,octant

where

$$\boldsymbol{A} = \frac{1}{2}\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1/\sqrt{2} & -1/\sqrt{2} & 0 & 0 \\ 0 & 0 & 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix}$$
$$\boldsymbol{w} = \begin{bmatrix} d_1 \\ \sqrt{2}d_2 \\ \sqrt{2}d_3 \end{bmatrix}, \quad \boldsymbol{d} = \begin{bmatrix} \frac{1}{\pi}\int_0^\pi \bar{w}(\varphi)\,\mathrm{d}\varphi \\ \frac{1}{\pi}\int_0^\pi \bar{w}(\varphi)\cos(2\varphi)\,\mathrm{d}\varphi \\ \frac{1}{\pi}\int_0^\pi \bar{w}(\varphi)\sin(2\varphi)\,\mathrm{d}\varphi \end{bmatrix}. \tag{5.2.7}$$

e,tomo,design,w

The **null space** of $\boldsymbol{A}$ is spanned by the vector $(1, 1, -1, -1)$. When $\bar{w}(\varphi)$ is the inverse of the variance of the projection data at angle φ, one can think of d_1 as quantifying the overall "certainty," d_2 is related to the horizontal and vertical directions, and d_3 is related to the diagonal directions.

This under-determined 3×4 system could be solved by the iterative **NNLS algorithm** [6, p. 158]. However, using the properties of $\boldsymbol{A}$ and $\boldsymbol{d}$, we can avoid iterations entirely by solving (5.2.6) analytically. Any solution $\boldsymbol{r}$ of (5.2.6) must satisfy the following **KKT conditions**: $r_l > 0 \implies \frac{\partial}{\partial r_l}\Psi(\boldsymbol{r}) = 0$ and $r_l \le 0 \implies \frac{\partial}{\partial r_l}\Psi(\boldsymbol{r}) \ge 0$, where

$$\nabla\Psi(\boldsymbol{r}) = \boldsymbol{A}'\boldsymbol{A}\,\boldsymbol{r} - \boldsymbol{A}'\boldsymbol{w} = \frac{1}{8}\begin{bmatrix} 3 & 1 & 2 & 2 \\ 1 & 3 & 2 & 2 \\ 2 & 2 & 3 & 1 \\ 2 & 2 & 1 & 3 \end{bmatrix}\boldsymbol{r} - \frac{1}{2}\begin{bmatrix} 1 & 1 & 0 \\ 1 & -1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & -1 \end{bmatrix}\boldsymbol{d}.$$

We assume hereafter that $\bar{w}(\varphi) \ge 0$. It follows from (5.2.7) that $\sqrt{d_2^2 + d_3^2} \le d_1$. The structure of $\boldsymbol{A}$ leads to eight-fold symmetry that simplifies analysis.

- If $d_2 < 0$ we can solve for $\boldsymbol{r}$ using $|d_2|$ and then swap r_1 with r_2.
- If $d_3 < 0$ we can solve for $\boldsymbol{r}$ using $|d_3|$ and then swap r_3 with r_4.
- If $d_3 > d_2$ we can solve for $\boldsymbol{r}$ with d_2 and d_3 interchanged, and then swap r_1 with r_3 and r_2 with r_4.

Therefore, hereafter we focus on cases where $0 \le d_3 \le d_2 \le d_1$. Fig. 5.2.1 shows these first octant cases, where $0 \le d_3/d_1 \le d_2/d_1 \le 1$. The cases are numbered according to the number of nonzero elements of $\boldsymbol{r}$. We used MATLAB to "discover" the approximate region boundaries and then verified them analytically as follows.

1 If $d_2 \ge \frac{1}{2}d_1$ and $d_3 \le \frac{2}{3}d_2 - \frac{1}{3}d_1$, then

$$r_1 = \frac{4}{3}(d_1 + d_2), \qquad r_2 = r_3 = r_4 = 0$$

because $\nabla\Psi(\boldsymbol{r}) = \begin{bmatrix} 0 & \frac{2}{3}(d_2 - \frac{1}{2}d_1) & \frac{1}{2}\left(\frac{2}{3}d_2 - \frac{1}{3}d_1 - d_3\right) & \frac{1}{3}(d_2 - \frac{1}{2}d_1) + \frac{1}{2}d_3 \end{bmatrix}$ satisfies the KKT conditions. That solution has only one nonzero penalty coefficient (r_1), so it alone does not ensure that $\mathbf{R}$ is positive-semidefinite.

2 If $d_3 \geq \frac{2}{3} d_2 - \frac{1}{3} d_1$ and $d_3 + d_2 \geq \frac{1}{2} d_1$, then

$$r_1 = \frac{8}{5} \left(\frac{1}{2} d_1 + \frac{3}{2} d_2 - d_3 \right), \quad r_3 = \frac{12}{5} \left[d_3 - \left(\frac{2}{3} d_2 - \frac{1}{3} d_1 \right) \right], \quad r_2 = r_4 = 0$$

because $\nabla \Psi(\boldsymbol{r}) = \left[\begin{array}{cccc} 0 & \frac{2}{5} \left(d_2 + d_3 - \frac{1}{2} d_1 \right) & 0 & \frac{2}{5} \left(d_2 + d_3 - \frac{1}{2} d_1 \right) \end{array} \right]$ satisfies the KKT conditions.

3 If $d_3 + d_2 \leq \frac{1}{2} d_1$ and $d_2 \geq \frac{1}{4} d_1$, then there are multiple nonnegative $\boldsymbol{r}$ choices that are exact solutions satisfying $\nabla \Psi(\boldsymbol{r}) = \boldsymbol{0}$. The solution with minimum norm is

$$r_1 = 4 d_2, \quad r_2 = 0, \quad r_3 = d_1 - 2 d_2 + 2 d_3, \quad r_4 = 2 \left[\frac{1}{2} d_1 - (d_2 + d_3) \right]. \tag{5.2.8}$$

e,tomo,design,case3

4 If $d_2 \leq \frac{1}{4} d_1$, then there are multiple nonnegative $\boldsymbol{r}$ that are exact solutions satisfying $\nabla \Psi(\boldsymbol{r}) = \boldsymbol{0}$. The natural choice is the minimum-norm $\boldsymbol{r}$ given by the pseudo-inverse solution $\boldsymbol{r} = \boldsymbol{A}^\dagger \boldsymbol{w}$, where

$$r_1 = 2 \left(\frac{1}{4} d_1 + d_2 \right), \quad r_2 = 2 \left(\frac{1}{4} d_1 - d_2 \right), \quad r_3 = 2 \left(\frac{1}{4} d_1 + d_3 \right), \quad r_4 = 2 \left(\frac{1}{4} d_1 - d_3 \right).$$

As a sanity check, consider what happens in the unweighted case where $\bar{w}(\varphi) = 1$. Then $\boldsymbol{d} = (1, 0, 0)$ so the 4th case is applicable, so $r_l = 1/2$, $l = 1, \ldots, 4$ and $R(\rho, \Phi; \boldsymbol{r})$ in (5.2.2) simplifies to $(2\pi\rho)^2$, yielding isotropic spatial resolution (provided the $\sqrt{2}$ factors are handled appropriately as discussed in footnote 1).

The analytical solution presented above is for the usual first-order differences with the second-order neighborhood (5.2.1). For higher-order differences or neighborhoods, it would appear to become increasingly cumbersome to solve (5.2.3) analytically, so the iterative NNLS approach may be more appealing.

The analytical solution above is a continuous function of $\boldsymbol{d}$, which in turn is a continuous function of $\bar{w}(\varphi)$. (In contrast, MATLAB's NNLS solution is discontinuous in $\boldsymbol{d}$, presumably because it does not use the minimum norm condition in regions where there are multiple solutions.) This continuity property would seem to be desirable for avoiding artifacts in the reconstructed images.

MIRT *This analytical solution is in* `penalty2_design.m`*, which includes a built-in comparison to the NNLS algorithm, called* `lsqnonlin` *in* MATLAB.

Up to this point, everything has been solved analytically. For practical implementation, only the three simple integrals in (5.2.7) need discretization.

Using the **local shift invariance** concept and the local angular weighting concept described in Chapter 22, this approach can be extended to be a simple and practical design method even for systems with more realistic noise models than (4.3.19).

s,tomo,design,3

5.2.2 3D case (s,tomo,design,3)

A 3D quadratic roughness penalty has the form

$$\mathsf{R}(f) = \sum_{\vec{n} \in \mathbb{Z}^3} \sum_{l=1}^{L} \frac{1}{2} \left| (c_l \mathbin{***} f)[\vec{n}] \right|^2,$$

where $***$ denotes 3D convolution and $c_l[\vec{n}]$ is a 3D high-pass filter. For the usual first-order differences,

$$c_l[\vec{n}] = \frac{1}{\|\vec{n}_l\|} \left(\delta[\vec{n}] - \delta[\vec{n} - \vec{n}_l] \right),$$

where $\vec{n}_l \in \{-1, 0, 1\}^3$. The corresponding magnitude response is

$$|C_l(\vec{\omega})|^2 = \frac{1}{\|\vec{n}_l\|^2} \left| 1 - \mathrm{e}^{-\imath \vec{\omega} \cdot \vec{n}} \right|^2 \approx \frac{1}{\|\vec{n}_l\|^2} (\vec{\omega} \cdot \vec{n}_l)^2 \tag{5.2.9}$$

e,tomo,design,3,Cl,approx

for $\vec{\omega} \in [-\pi, \pi]^3$. Now we convert to spherical coordinates with unit vector

$$\vec{e}(\Phi, \Theta) = (\cos(\Theta)\cos(\Phi), \ \cos(\Theta)\sin(\Phi), \ \sin(\Theta)).$$

Then $\vec{\omega} = 2\pi\varrho\, \vec{e}(\Phi, \Theta)$ and $\vec{n}_l = \|\vec{n}_l\| \ \vec{e}(\Phi_l, \Theta_l)$ for suitable (Φ_l, Θ_l). Then the magnitude response is

$$|C_l(\vec{\omega})|^2 \approx (2\pi\varrho)^2 \left(\vec{e}(\Phi, \Theta) \cdot \vec{e}(\Phi_l, \Theta_l) \right)^2.$$

Thus the regularization operator (*cf.* (4.3.12)) is

$$\boldsymbol{\mathcal{R}} = \boldsymbol{\mathcal{F}}_3^{-1} \boldsymbol{\mathcal{D}}(R(\varrho, \Phi, \Theta; \boldsymbol{r})) \boldsymbol{\mathcal{F}}_3$$

where $\boldsymbol{r} = (r_1, \ldots, r_L)$ and

$$R(\varrho, \Phi, \Theta; \boldsymbol{r}) \approx (2\pi\varrho)^2 \sum_{l=1}^{L} \left(\vec{e}(\Phi, \Theta) \cdot \vec{e}(\Phi_l, \Theta_l)\right)^2 r_l. \tag{5.2.10}$$

e,tomo,design,R(p,q,r),3d

Considering (4.6.23), we would like to design the penalty coefficients $\boldsymbol{r}$ using something like the following minimization criterion:

$$\hat{\boldsymbol{r}} = \arg\min_{\boldsymbol{r} \succeq \boldsymbol{0}} \frac{1}{\pi^2} \int_0^{\pi} \int_{-\pi/2}^{\pi/2} v(\Phi, \Theta) \left| T(\Phi, \Theta) - \sum_{l=1}^{L} \left(\vec{e}(\Phi, \Theta) \cdot \vec{e}(\Phi_l, \Theta_l)\right)^2 r_l \right|^2 \mathrm{d}\Theta \, \mathrm{d}\Phi, \tag{5.2.11}$$

e,tomo,design,rh,int,3d

where $T(\Phi, \Theta)$ denotes a target frequency response for the regularizer and $v(\Phi, \Theta)$ denotes some user-selected weighting function.

The natural choices of weighting functions have the form $v(\Phi, \Theta) = \cos^m(\Theta)$ for $m = 0$ or $m = 1$; the latter is natural for spherical coordinates [7].

Letting $\vec{e}(\Phi_l, \Theta_l) = (x_l, y_l, z_l)$, we can expand the inner squared terms in (5.2.11) in terms of 6 functions defined for $(\Phi, \Theta) \in [0, \pi] \times [-\pi/2, \pi/2]$ as follows:

$$\begin{aligned}
\left(\vec{e}(\Phi, \Theta) \cdot \vec{e}(\Phi_l, \Theta_l)\right)^2 &= \left(x_l \cos(\Theta)\cos(\Phi) + y_l \cos(\Theta)\sin(\Phi) + z_l \sin(\Theta)\right)^2 = \\
&= \sum_{k=1}^{6} b_{kl}\, g_k(\Theta, \Phi) \\
\boldsymbol{b}_l &= \left(x_l^2, y_l^2, z_l^2, x_l y_l, x_l z_l, y_l z_l\right) \\
g_1(\Theta, \Phi) &= \cos^2(\Theta)\cos^2(\Phi) \\
g_2(\Theta, \Phi) &= \cos^2(\Theta)\sin^2(\Phi) \\
g_3(\Theta, \Phi) &= \sin^2(\Theta) \\
g_4(\Theta, \Phi) &= \cos^2(\Theta)\sin(2\Phi) \\
g_5(\Theta, \Phi) &= \sin(2\Theta)\cos(\Phi) \\
g_6(\Theta, \Phi) &= \sin(2\Theta)\sin(\Phi)\,.
\end{aligned}$$

It is somewhat disheartening that, at least to within the approximation (5.2.9), no matter how many neighbors we use, the local frequency response of $\boldsymbol{\mathcal{R}}$ is spanned by just 6 functions. Even if we use all $L = 13$ nearest neighbors in 3D, the regularizer's local frequency response can provide only modest angular variations.

In light of the above expansion, we rewrite the NNLS problem (5.2.11) as follows

$$\hat{\boldsymbol{r}} = \arg\min_{\boldsymbol{r} \succeq \boldsymbol{0}} \Psi(\boldsymbol{r}), \quad \Psi(\boldsymbol{r}) \triangleq \frac{1}{\pi^2} \int_0^{\pi} \int_{-\pi/2}^{\pi/2} \cos^m(\Theta) \left| T(\Phi, \Theta) - (\boldsymbol{\mathcal{B}}\,\boldsymbol{r})(\Phi, \Theta) \right|^2 \mathrm{d}\Theta \, \mathrm{d}\Phi, \tag{5.2.12}$$

e,tomo,design,rh,int,3d,new

where the operator $\boldsymbol{\mathcal{B}} : \mathbb{R}^L \mapsto \mathcal{L}_2([0, \pi] \times [-\pi/2, \pi/2])$ is defined by

$$(\boldsymbol{\mathcal{B}}\,\boldsymbol{r})(\Phi, \Theta) = \sum_{l=1}^{L} \left(\vec{e}(\Phi, \Theta) \cdot \vec{e}(\Phi_l, \Theta_l)\right)^2 r_l = \sum_{l=1}^{L} \sum_{k=1}^{6} b_{kl}\, g_k(\Theta, \Phi)\, r_l = \boldsymbol{\mathcal{G}}\boldsymbol{B}\,\boldsymbol{r},$$

where

$$(\boldsymbol{\mathcal{G}}\,\boldsymbol{u})(\Phi, \Theta) = \sum_{k=1}^{6} g_k(\Theta, \Phi)\, u_k$$

and $\boldsymbol{B}$ is the $6 \times L$ matrix whose lth column is given by $\boldsymbol{b}_l$ above.

It will be convenient to find a QR decomposition for $\boldsymbol{\mathcal{B}}$, *i.e.*, $\boldsymbol{\mathcal{B}} = \boldsymbol{\mathcal{Q}}\boldsymbol{A}$ where $\boldsymbol{A} \in \mathbb{R}^{6 \times L}$ and the 6 "columns" of $\boldsymbol{\mathcal{Q}}$, *i.e.*, $\{q_k(\Phi, \Theta)\}_{k=1}^{6}$, are an orthonormal basis for the span of $\boldsymbol{\mathcal{B}}$. Note that by construction, $\boldsymbol{A} = \boldsymbol{\mathcal{Q}}^*\boldsymbol{\mathcal{B}} = (\boldsymbol{\mathcal{Q}}^*\boldsymbol{\mathcal{G}})\,\boldsymbol{B}$, where $(\boldsymbol{\mathcal{Q}}^*\boldsymbol{\mathcal{G}})$ is a 6×6 matrix.

Defining an orthonormal basis requires a suitable inner product. The two natural inner products for (5.2.12) are

$$\langle f,\, g \rangle_m \triangleq \frac{1}{\pi^2} \int_0^{\pi} \int_{-\pi/2}^{\pi/2} \cos^m(\Theta)\, f(\Phi, \Theta)\, g^*(\Phi, \Theta) \, \mathrm{d}\Theta \, \mathrm{d}\Phi, \quad m \in \{0, 1\}\,. \tag{5.2.13}$$

e,tomo,design,3,inprod

With either of these two inner products, we can rewrite (5.2.12) as

$$\Psi(\boldsymbol{r}) = \|T - \boldsymbol{\mathcal{B}}\,\boldsymbol{r}\|_m^2 = \|T - \boldsymbol{\mathcal{Q}}\boldsymbol{A}\,\boldsymbol{r}\|_m^2 = \|T\|_m^2 - 2\,\langle T,\, \boldsymbol{\mathcal{Q}}\boldsymbol{A}\,\boldsymbol{r} \rangle_m + \|\boldsymbol{\mathcal{Q}}\boldsymbol{A}\,\boldsymbol{r}\|_m^2 \tag{5.2.14}$$

$$= \|T\|_m^2 - 2\,\langle \boldsymbol{\mathcal{Q}}^* T,\, \boldsymbol{A}\,\boldsymbol{r} \rangle_2 + \|\boldsymbol{A}\,\boldsymbol{r}\|_2^2 \equiv \|\boldsymbol{t} - \boldsymbol{A}\,\boldsymbol{r}\|_2^2\,, \tag{5.2.15}$$

e,tomo,design,rh,int,3d,vr

where

$$\boldsymbol{t} \triangleq \boldsymbol{\mathcal{Q}}^* T \in \mathbb{R}^6, \quad t_k = [\boldsymbol{\mathcal{Q}}^* T]_k = \frac{1}{\pi^2} \int_0^\pi \int_{-\pi/2}^{\pi/2} \cos^m(\Theta)\, T(\Phi, \Theta)\, q_k(\Phi, \Theta)\, \mathrm{d}\Theta\, \mathrm{d}\Phi, \quad k = 1, \dots, 6.$$

For $m = 0$, the following 6 functions $\{q_k(\Phi, \Theta)\}$ span the range of $\boldsymbol{\mathcal{B}}$ and are orthonormal w.r.t. (5.2.13) [8, p. 44]:

$$\begin{array}{c} 1 \\ \sqrt{2}\cos(2\Theta) \\ \frac{4}{\sqrt{3}}\cos^2(\Theta)\cos(2\Phi) \\ \frac{4}{\sqrt{3}}\cos^2(\Theta)\sin(2\Phi) \\ 2\sin(2\Theta)\cos(\Phi) \\ 2\sin(2\Theta)\sin(\Phi)\,. \end{array}$$

For this choice of $\boldsymbol{\mathcal{Q}}$, symbolic integration determines the elements of $\boldsymbol{A} = \boldsymbol{\mathcal{Q}}^* \boldsymbol{\mathcal{G}} \boldsymbol{B}$ using

$$\boldsymbol{\mathcal{Q}}^* \boldsymbol{\mathcal{G}} = \begin{bmatrix} 2 & 2 & 4 & 0 & 0 & 0 \\ \sqrt{2} & \sqrt{2} & -2\sqrt{2} & 0 & 0 & 0 \\ \sqrt{3} & -\sqrt{3} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}. \tag{5.2.16}$$

e,tomo,design,3,qq0

For $m = 1$, the following 6 functions $\{q_k(\Phi, \Theta)\}$ span the range of $\boldsymbol{\mathcal{B}}$ and are orthonormal[2] w.r.t. (5.2.13):

$$\sqrt{15\pi} \begin{bmatrix} \sqrt{\frac{1}{30}} \\ \sqrt{\frac{3}{32}}\left(\cos(2\Theta) - \frac{1}{3}\right) \\ \sqrt{\frac{15\pi}{8}}\cos^2(\Theta)\cos(2\Phi) \\ \sqrt{\frac{1}{8}}\cos^2(\Theta)\sin(2\Phi) \\ \sqrt{\frac{1}{8}}\sin(2\Theta)\cos(\Phi) \\ \sqrt{\frac{1}{8}}\sin(2\Theta)\sin(\Phi) \end{bmatrix}.$$

For this choice of $\boldsymbol{\mathcal{Q}}$, symbolic integration determines the elements of $\boldsymbol{A} = \boldsymbol{\mathcal{Q}}^* \boldsymbol{\mathcal{G}} \boldsymbol{B}$. The final expressions are messy but have the same sparsity pattern as (5.2.16).

For a 3D penalty, typically $L = 3$ or $L = 13$. In either case, it would seem to be tedious to try to solve (5.2.15) analytically, so we resort to the iterative nonnegative least-squares (NNLS) method. Because L is small, the number of iterations required is modest.

In the practical use of this method, one would need to consider the fact that $T(\Phi, \Theta)$ is space variant. In particular, slices near the end of a cylindrical PET scanner have a smaller acceptance angle $\theta_{\max}$ than slices in the middle.

At first glance there appears to be a significant difference between the 2D case and the 3D case. In the 2D case, if we use uniform weighting, $w(\varphi) = 1$, then we can easily "design" a 2D penalty function to achieve isotropic spatial resolution. However, in the 3D case, even if we use $w(\varphi) = 1$, the function $T(\Phi, \Theta)$ is nonuniform (see (4.6.21)), and cannot be expressed exactly as a linear combination of the 6 basis functions above, except in the unrealistic case where $\theta_{\max} = \pi/2$, corresponding to a complete spherical detector. This situation provides further rationale for adopting the "hybrid" regularization/smoothing approach proposed in [9], where one uses a modest amount of regularization (designed to be nearly isotropic) for iterative reconstruction, followed by post-filtering with an isotropic filter such as a separable gaussian. With a closer look, perhaps the 2D and 3D problems are not really so different. The difficulty in the 3D case arises due to the finite limit $\theta_{\max}$. A more realistic 2D model would include a finite maximum radius $r_{\max}$, and such a limit would also introduce a space-varying anisotropy that could not be overcome perfectly with practical regularizers.

Given that there are only 6 basis functions, it could be somewhat wasteful to solve the highly under-determined NNLS problem (5.2.15), where $\boldsymbol{r}$ might have 13 elements in the 3D case. It seems plausible that most of the solutions lie in or close to a lower-dimensional subspace (or manifold) and we could parameterize that subspace, *e.g.*, $\hat{\boldsymbol{r}} = \boldsymbol{S}\hat{\boldsymbol{\theta}}$ for some matrix $\boldsymbol{S}$. We could then search over

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\,:\,\boldsymbol{S}\boldsymbol{\theta} \succeq \boldsymbol{0}}{\arg\min} \|\boldsymbol{A}\boldsymbol{S}\boldsymbol{\theta} - \boldsymbol{w}\|_2^2\,,$$

and store only $\hat{\boldsymbol{\theta}}$ for each voxel rather than $\hat{\boldsymbol{r}}$. A natural choice would be $\boldsymbol{S} = \boldsymbol{A}^\dagger$. An *open problem* is to determine whether this would give the same family solutions, or almost the same.

[2] The version in [8, p. 44] is not quite correct. That version is essentially $m = 2$ not $m = 1$.

s,tomo,design,summ

5.3 Summary (s,tomo,design,summ)

This short chapter just scratches the surface of methods for regularization design. See also Chapter 22 and [7].

s,tomo,design,prob

5.4 Problems (s,tomo,design,prob)

Problem 5.1 *Verify that (5.2.8) is indeed a minimum norm solution for Case 3.*

Problem 5.2 *Use appropriate KKT conditions to show that if* $L = 2$, i.e.*, only horizontal and vertical neighbors are used, then the solution to (5.2.3) is*

$$r_1 = \begin{cases} d_1 + 2d_2, & |d_2| < d_1/2 \\ \frac{4}{3}(d_1 + d_2), & d_2 > d_1/2 \\ 0, & \text{otherwise,} \end{cases} \qquad r_2 = \begin{cases} d_1 - 2d_2, & |d_2| < d_1/2 \\ \frac{4}{3}(d_1 - d_2), & d_2 < -d_1/2 \\ 0, & \text{otherwise.} \end{cases}$$

5.5 Bibliography

fessler:97:srp

[1] J. A. Fessler. *Spatial resolution properties of penalized weighted least-squares image reconstruction with model mismatch*. Tech. rep. 308. Available from `web.eecs.umich.edu/~fessler`. Univ. of Michigan, Ann Arbor, MI, 48109-2122: Comm. and Sign. Proc. Lab., Dept. of EECS, Mar. 1997. URL: http://web.eecs.umich.edu/~fessler/papers/lists/files/tr/97,308,srp.pdf (cit. on p. 5.1).

fessler:03:aat

[2] 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 p. 5.1).

fessler:94:pwl

[3] J. A. Fessler. "Penalized weighted least-squares image reconstruction for positron emission tomography." In: *IEEE Trans. Med. Imag.* 13.2 (June 1994), 290–300. DOI: 10.1109/42.293921 (cit. on p. 5.2).

nuyts:03:apl

[4] J. Nuyts and J. A. Fessler. "A penalized-likelihood image reconstruction method for emission tomography, compared to post-smoothed maximum-likelihood with matched spatial resolution." In: *IEEE Trans. Med. Imag.* 22.9 (Sept. 2003), 1042–52. DOI: 10.1109/TMI.2003.816960 (cit. on pp. 5.2, 5.3).

stayman:00:rfu

[5] J. W. Stayman and J. A. Fessler. "Regularization for uniform spatial resolution properties in penalized-likelihood image reconstruction." In: *IEEE Trans. Med. Imag.* 19.6 (June 2000), 601–15. DOI: 10.1109/42.870666 (cit. on p. 5.3).

lawson:74

[6] C. L. Lawson and R. J. Hanson. *Solving least squares problems*. Prentice-Hall, 1974 (cit. on p. 5.4).

shi:06:qrd

[7] H. Shi and J. A. Fessler. "Quadratic regularization design for 3D axial CT." In: *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.* 2006, 2834–6. DOI: 10.1109/NSSMIC.2006.356467 (cit. on pp. 5.6, 5.8).

shi:08:phd

[8] H. Shi. "Fast regularization design for tomographic image reconstruction for uniform and isotropic spatial resolution." PhD thesis. Ann Arbor, MI: Univ. of Michigan, Ann Arbor, MI, 48109-2122, 2008. URL: https://hdl.handle.net/2027.42/61554 (cit. on p. 5.7).

stayman:04:cfn

[9] 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 p. 5.7).