

Mathematical Foundations of Data Sciences



Gabriel Peyré
CNRS & DMA
École Normale Supérieure
gabriel.peyre@ens.fr
<https://mathematical-tours.github.io>
www.numerical-tours.com

December 25, 2017

Presentation

This book draft presents an overview of important mathematical and numerical foundations for modern data sciences. It covers in particulars the basics of signal and image processing (Fourier, Wavelets, and their applications to denoising and compression), imaging sciences (inverse problems, sparsity, compressed sensing) and machine learning (linear regression, logistic classification, deep learning). The focus is on the mathematically-sounded exposition of the methodological tools (in particular linear operators, non-linear approximation, convex optimization, optimal transport) and how they can be mapped to efficient computational algorithms. These course notes are also intended to be the theoretical companion for the Numerical Tours¹ web site, which presents Matlab/Python/Julia/R detailed implementations of all the concepts covered here.

¹www.numerical-tours.com

Contents

1	Shannon Theory	11
1.1	Analog vs. Discrete Signals	11
1.1.1	Acquisition and Sampling	12
1.1.2	Linear Translation Invariant Sampler	12
1.2	Shannon Sampling Theorem	13
1.3	Shannon Source Coding Theorem	15
2	Fourier Transforms	17
2.1	Hilbert spaces and Fourier Transforms	17
2.1.1	Hilbertian bases	17
2.1.2	Fourier basis on $\mathbb{R}/2\pi\mathbb{Z}$	17
2.2	Convolution on \mathbb{R} and \mathbb{T}	18
2.2.1	Convolution	18
2.2.2	Translation Invariant Operators	19
2.2.3	Revisiting Poisson formula using distributions.	21
2.3	Finite Fourier Transform and Convolution	22
2.3.1	Discrete Ortho-bases	22
2.3.2	Discrete Fourier transform	23
2.3.3	Fast Fourier transform	23
2.3.4	Finite convolution	24
2.4	Discretisation Issues	25
2.4.1	Fourier approximation via spatial zero padding.	25
2.4.2	Fourier approximation via spatial zero padding.	26
2.5	Fourier in Multiple Dimensions	26
2.5.1	On Continuous Domains	26
2.5.2	On Discrete Domains	28
2.5.3	Shannon sampling theorem.	29
2.5.4	Convolution in higher dimension.	29
2.6	Application to ODEs and PDEs	29
2.6.1	On Continuous Domains	29
2.6.2	Finite Domain and Discretization	30
2.7	A Bit of Group Theory	30
2.7.1	Characters	31
2.7.2	More General cases	32
2.8	A Bit of Spectral Theory	33
2.8.1	On a Surface or a Manifold	33
2.8.2	Spherical Harmonics	33
2.8.3	On a Graph	33

3 Linear Mesh Processing	35
3.1 Surface Discretization with Triangulated Mesh	35
3.1.1 Continuous Geometry of Surfaces	35
3.1.2 Discretization of Surfaces with Triangulations	36
3.2 Linear Mesh Processing	37
3.2.1 Functions on a Mesh	38
3.2.2 Local Operators	38
3.2.3 Approximating Integrals on a Mesh	39
3.2.4 Example on a Regular Grid	41
3.2.5 Gradients and Laplacians on Meshes	42
3.2.6 Examples in 1D and 2D	43
3.2.7 Example of a Parametric Surface	44
3.3 Diffusion and Regularization on Surfaces	44
3.3.1 Heat Diffusion	44
3.3.2 Spectral Decomposition	46
3.3.3 Spectral Theory on a Regular Grid	48
3.3.4 Spectral Resolution of the Heat Diffusion	49
3.3.5 Quadratic Regularization	50
3.3.6 Application to Mesh Compression	50
3.3.7 Application to Mesh Parameterization	53
3.3.8 Application to Mesh Flattening	53
4 Wavelets	57
4.1 Multi-resolution Approximation Spaces	57
4.2 Multi-resolution Details Spaces	59
4.3 On Bounded Domains	60
4.4 Fast Wavelet Transform	61
4.4.1 Discretization	61
4.4.2 Forward Fast Wavelet Transform (FWT)	62
4.4.3 Inverse Fast Transform (iFWT)	64
4.5 2-D Wavelets	65
4.5.1 Anisotropic Wavelets	65
4.5.2 Isotropic Wavelets	66
4.6 Wavelet Design	71
4.6.1 Low-pass Filter Constraints	71
4.6.2 High-pass Filter Constraints	73
4.6.3 Wavelet Design Constraints	74
4.6.4 Daubechies Wavelets	76
5 Multiresolution Mesh Processing	79
5.1 Semi-regular Meshes	79
5.1.1 Nested Multiscale Grids.	79
5.1.2 Semi-regular Triangulation.	79
5.1.3 Spherical Geometry Images	81
5.2 Subdivision Curves	82
5.3 Subdivision Surfaces	84
5.3.1 Interpolation Operators	84
5.3.2 Some Classical Subdivision Stencils	85
5.3.3 Invariant Neighborhoods	86
5.3.4 Convergence of Subdivisions	87
5.4 Wavelets on Meshes	90
5.4.1 Multiscale Biorthogonal Bases on Meshes	90

5.4.2	The Lifting Scheme	90
5.4.3	Imposing vanishing moments.	92
5.4.4	Lifted Wavelets on Meshes	93
5.4.5	Non-linear Mesh Compression	94
6	Linear and Non-linear Approximation	97
6.1	Approximation	97
6.1.1	Approximation in an Ortho-basis	97
6.1.2	Linear Approximation	97
6.1.3	Non-linear Approximation	98
6.2	Signal and Image Modeling	99
6.2.1	Uniformly Smooth Signals and Images	99
6.2.2	Piecewise Regular Signals and Images	101
6.2.3	Bounded Variation Signals and Images	101
6.2.4	Cartoon Images	102
6.3	Efficient approximation	102
6.3.1	Decay of Approximation Error	102
6.3.2	Comparison of bases.	103
6.4	Fourier Linear Approximation of Smooth Functions	104
6.4.1	1-D Fourier Approximation	104
6.4.2	Sobolev Images	107
6.5	Wavelet Approximation of Piecewise Smooth Functions	107
6.5.1	Decay of Wavelet Coefficients	107
6.5.2	1-D Piecewise Smooth Approximation	108
6.5.3	2-D Piecewise Smooth Approximation	110
6.6	Cartoon Images Approximation	112
6.6.1	Wavelet Approximation of Cartoon Images	112
6.6.2	Finite Element Approximation	112
6.6.3	Curvelets Approximation	113
7	Compression	117
7.1	Transform Coding	117
7.1.1	Coding	117
7.1.2	De-coding	118
7.1.3	Support Coding	118
7.2	Entropic Coding	120
7.3	JPEG-2000	121
8	Denoising	125
8.1	Noise Modeling	125
8.1.1	Noise in Images	125
8.1.2	Image Formation	126
8.1.3	Denoiser	127
8.2	Linear Denoising using Filtering	127
8.2.1	Translation Invariant Estimators	127
8.2.2	Optimal Filter Selection	128
8.2.3	Wiener Filter	128
8.2.4	Denoising and Linear Approximation	129
8.3	Non-linear Denoising using Thresholding	132
8.3.1	Hard Thresholding	132
8.3.2	Soft Thresholding	133
8.3.3	Minimax Optimality of Thresholding	134

8.3.4	Translation Invariant Thresholding Estimators	136
8.3.5	Exotic Thresholdings	138
8.3.6	Block Thresholding	139
8.4	Data-dependant Noises	141
8.4.1	Poisson Noise	142
8.4.2	Multiplicative Noise	145
9	Variational Priors and Regularization	149
9.1	Sobolev and Total Variation Priors	149
9.1.1	Continuous Priors	149
9.1.2	Discrete Priors	149
9.2	PDE and Energy Minimization	152
9.2.1	General Flows	152
9.2.2	Heat Flow	152
9.2.3	Total Variation Flows	153
9.2.4	PDE Flows for Denoising	155
9.3	Regularization for Denoising	156
9.3.1	Regularization	156
9.3.2	Sobolev Regularization	157
9.3.3	TV Regularization	158
10	Inverse Problems	161
10.1	Inverse Problems Regularization	161
10.2	Theoretical Study of Quadratic Regularization	162
10.2.1	Singular Value Decomposition	163
10.2.2	Tikonov Regularization	164
10.3	Quadratic Regularization	167
10.3.1	Solving Linear System	168
10.4	Non-Quadratic Regularization	169
10.4.1	Total Variation Regularization	169
10.4.2	Gradient Descent Method	171
10.4.3	Examples of Gradient Computation	171
10.5	Examples of Inverse Problems	172
10.5.1	Deconvolution	172
10.5.2	Inpainting	172
10.5.3	Tomography Inversion	173
11	Sparse Regularization	177
11.1	Sparsity Priors	177
11.1.1	Ideal sparsity prior	177
11.1.2	Convex relaxation	177
11.1.3	Sparse Regularization and Thresholding	178
11.2	Sparse Regularization of Inverse Problems	180
11.3	Iterative Soft Thresholding Algorithm	180
11.3.1	Noiseless Recovery as a Linear Program	181
11.3.2	Projected Gradient Descent for ℓ^1	181
11.3.3	Iterative Soft Thresholding and Forward Backward	182
11.4	Example: Sparse Deconvolution	183
11.4.1	Sparse Spikes Deconvolution	183
11.4.2	Sparse Wavelets Deconvolution	183
11.4.3	Sparse Inpainting	184

12 Convex Analysis	189
12.1 Basics of Convex Analysis	189
12.1.1 Convex Sets and Functions	189
12.1.2 First Order Conditions	190
12.2 Convex Duality	192
12.2.1 Lagrange Duality	192
12.2.2 Legendre-Fenchel Transform	194
12.2.3 Fenchel-Rockafellar Duality	196
13 Convex Optimization	197
13.1 Gradient Descent Methods	197
13.1.1 Gradient Descent	197
13.1.2 Sub-gradient Descent	201
13.1.3 Projected Gradient Descent	201
13.2 Proximal Algorithm	201
13.2.1 Proximal Map	201
13.2.2 Basic Properties	202
13.2.3 Related Concepts	203
13.3 Primal Algorithms	204
13.3.1 Proximal Point Algorithm	204
13.3.2 Forward-Backward	204
13.3.3 Douglas-Rachford	206
13.4 Dual and Primal-Dual Algorithms	207
13.4.1 Forward-backward on the Dual	207
13.4.2 Primal-Dual Splitting	208
14 Theory of Sparse Regularization	209
14.1 Existence and Uniqueness	209
14.1.1 Existence	209
14.1.2 Polytope Projection for the Constraint Problem	209
14.1.3 Optimality Conditions	211
14.1.4 Uniqueness	212
14.1.5 Duality	213
14.2 Consistency and Sparsistency	214
14.2.1 Bregman Divergence Rates for General Regularizations	214
14.2.2 Linear Rates in Norms for ℓ^1 Regularization	216
14.2.3 Sparsistency	217
14.3 Sparse Deconvolution Case Study	220
15 Compressed Sensing	223
15.1 Motivation and Potential Applications	223
15.1.1 Single Pixel Camera	223
15.1.2 Sparse Recovery	224
15.2 Dual Certificate Theory and Non-Uniform Guarantees	225
15.2.1 Random Projection of Polytopes	225
15.2.2 Random Matrices	225
15.2.3 Dual Certificates	227
15.3 RIP Theory for Uniform Guarantees	230
15.3.1 Restricted Isometry Constants	230
15.3.2 RIP implies dual certificates	231
15.3.3 RIP implies stable recovery	233
15.3.4 Fourier sampling RIP	234

16 Machine Learning	237
16.1 Unsupervised Learning	237
16.1.1 Dimensionality Reduction and PCA	237
16.1.2 Clustering and k -means	239
16.2 Empirical Risk Minimization	241
16.2.1 Empirical Risk	241
16.2.2 Prediction and Consistency	241
16.2.3 Parametric Approaches and Regularization	242
16.2.4 Testing Set and Cross-validation	242
16.3 Supervised Learning: Regression	243
16.3.1 Linear Regression	243
16.4 Supervised Learning: Classification	245
16.4.1 Nearest Neighbors Classification	245
16.4.2 Two Classes Logistic Classification	246
16.4.3 Multi-Classes Logistic Classification	248
16.5 Kernel Methods	250
16.5.1 Reproducing Kernel Hilbert Space	250
16.5.2 Examples of Kernelized Algorithms	252
17 Deep Learning	255
17.1 Stochastic Optimization	255
17.1.1 Minimizing Sums and Expectation	255
17.1.2 Batch Gradient Descent (BGD)	256
17.1.3 Stochastic Gradient Descent (SGD)	256
17.1.4 Stochastic Gradient Descent with Averaging (SGA)	259
17.1.5 Stochastic Averaged Gradient Descent (SAG)	259
17.2 Automatic Differentiation	260
17.2.1 Reverse Differentiation on a Feedforward Graph	260
17.2.2 Reverse Differentiation on a Generic Computational Graph	261
17.3 Deep Discriminative Models	263
17.3.1 Deep Network Structure	263
17.3.2 Perceptron and Shallow Models	265
17.3.3 Convolutional Neural Networks	265
17.3.4 Scattering Transform	266
18 Optimal Transport	267
18.1 Radon Measures	267
18.2 Monge Problem	270
18.3 Kantorovitch Problem	271
18.4 Duality	276
18.5 Sinkhorn	282
18.6 Extensions	289

Chapter 1

Shannon Theory

The main reference is [28].

1.1 Analog vs. Discrete Signals

To develop numerical tools and analyze their performances, the mathematical modeling is usually done over a continuous setting. An analog signal is a 1D function $f_0 \in L^2([0, 1])$ where $[0, 1]$ denotes the domain of acquisition, which might for instance be time. An analog image is a 2D function $f_0 \in L^2([0, 1]^2)$ where the unit square $[0, 1]^2$ is the image domain.

Although these notes are focussed on the processing of sounds and natural images, most of the methods extend to multi-dimensional datasets, which are higher dimensional mappings

$$f_0 : [0, 1]^d \rightarrow [0, 1]^s$$

where d is the dimensionality of the input space ($d = 1$ for sound and $d = 2$ for images) whereas s is the dimensionality of the feature space. For instance, gray scale images corresponds to $(d = 2, s = 1)$, videos to $(d = 3, s = 1)$, color images to $(d = 2, s = 3)$ where one has three channels (R, G, B). One can even consider multi-spectral images where $(d = 2, s \gg 3)$ that is made of a large number of channels for different light wavelengths. Figures 1.1 and 1.2 show examples of such data.

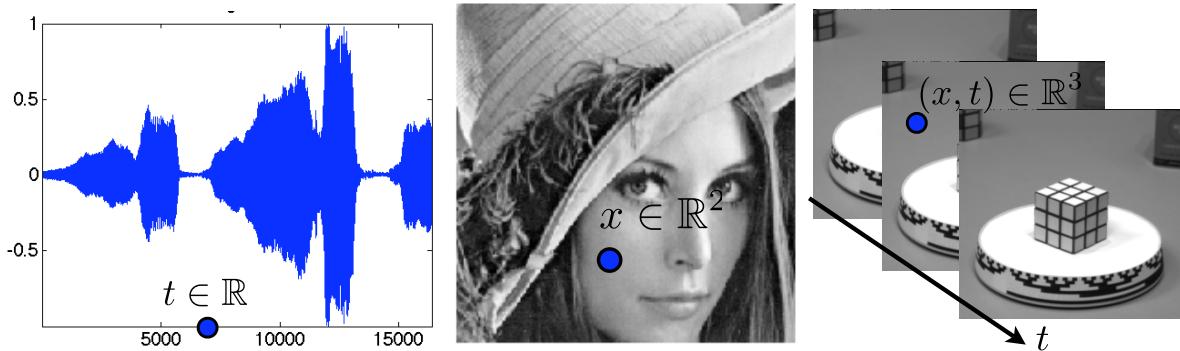


Figure 1.1: Examples of sounds ($d = 1$), image ($d = 2$) and videos ($d = 3$).

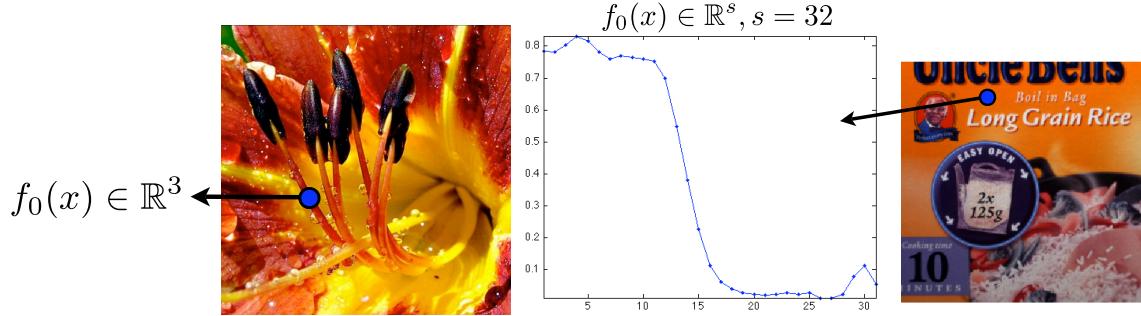


Figure 1.2: Example of color image $s = 3$ and multispectral image ($s = 32$).

1.1.1 Acquisition and Sampling

Signal acquisition is a low dimensional projection of the continuous signal performed by some hardware device. This is for instance the case for a microphone that acquires 1D samples or a digital camera that acquires 2D pixel samples. The sampling operation thus corresponds to mapping from the set of continuous functions to a discrete finite dimensional vector with N entries.

$$f_0 \in L^2([0, 1]^d) \mapsto f \in \mathbb{C}^N$$

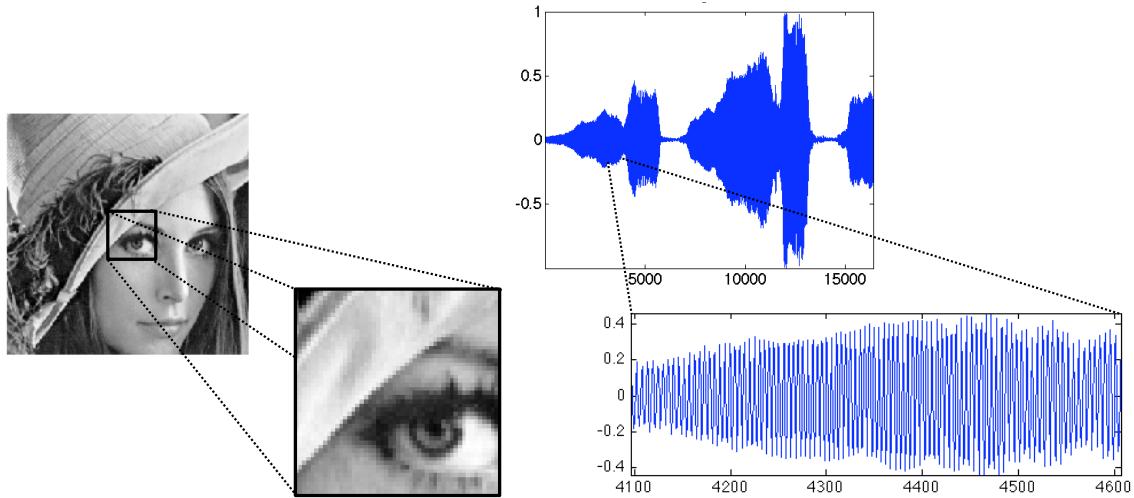


Figure 1.3: Image and sound discretization.

Figure 1.3 shows examples of discretized signals.

1.1.2 Linear Translation Invariant Sampler

A translation invariant sampler performs the acquisition as an inner product between the continuous signal and a constant impulse response h translated at the sample location

$$f[n] = \int_{-S/2}^{S/2} f_0(x)h(n/N - x)dx = f_0 \star h(n/N). \quad (1.1)$$

The precise shape of $h(x)$ depends on the sampling device, and is usually a smooth low pass function that is maximal around $x = 0$. The size S of the sampler determines the precision of the sampling device, and is usually of the order of $1/N$ to avoid blurring (if S is too large) or aliasing (if S is too small).

Section ?? details how to reverse the sampling operation in the case where the function is smooth.

1.2 Shannon Sampling Theorem

Reminders about Fourier transform. For $f \in L^1(\mathbb{R})$, its Fourier transform is defined as

$$\forall \omega \in \mathbb{R}, \quad \hat{f}(\omega) \stackrel{\text{def.}}{=} \int_{\mathbb{R}} f(x) e^{-ix\omega} dx. \quad (1.2)$$

One has $\|\hat{f}\|^2 = (2\pi)^{-1} \|f\|^2$, so that $f \mapsto \hat{f}$ can be extended by continuity to $L^2(\mathbb{R})$, which corresponds to computing \hat{f} as a limit when $T \rightarrow +\infty$ of $\int_{-T}^T f(x) e^{-ix\omega} dx$. When $\hat{f} \in L^1(\mathbb{R})$, one can invert the Fourier transform so that

$$f(x) = \int_{\mathbb{R}} \hat{f}(\omega) e^{ix\omega} d\omega, \quad (1.3)$$

which shows in particular that f is continuous with vanishing limits at $\pm\infty$.

The Fourier transform $\mathcal{F} : f \mapsto \hat{f}$ exchanges regularity and decay. For instance, if $f \in C^p(\mathbb{R})$ with an integrable Fourier transform, then $\mathcal{F}(f^{(p)})(\omega) = (i\omega)^{-p} \hat{f}(\omega)$ so that $|\hat{f}(\omega)| = O(1/|\omega|^p)$. Conversely,

$$\int_{\mathbb{R}} (1 + |\omega|)^{-p} |\hat{f}(\omega)| d\omega < +\infty \implies f \in C^p(\mathbb{R}). \quad (1.4)$$

Reminders about Fourier series. We denote $\mathbb{T} = \mathbb{R}/2\pi\mathbb{Z}$ the torus. A function $f \in L^2(\mathbb{T})$ is 2π -periodic, and can be viewed as a function $f \in L^2([0, 1])$ (beware that this means that the boundary points are glued together), and its Fourier coefficients are

$$\forall n \in \mathbb{Z}, \quad \hat{f}_n \stackrel{\text{def.}}{=} \frac{1}{2\pi} \int_0^{2\pi} f(x) e^{-ixn} dx.$$

This formula is equivalent to the computation of an inner-product $\hat{f}_n = \langle f, e_n \rangle$ for the inner-product $\langle f, g \rangle \stackrel{\text{def.}}{=} \frac{1}{2\pi} \int_{\mathbb{T}} f(x) \bar{g}(x) dx$. For this inner product, $(e_n)_n$ is orthonormal and is actually an Hilbert basis, meaning that one reconstruct with the following converging series

$$f = \sum_{n \in \mathbb{Z}} \langle f, e_n \rangle e_n \quad (1.5)$$

which means $\|f - \sum_{n=-N}^N \langle f, e_n \rangle e_n\|_{L^2(\mathbb{T})} \rightarrow 0$ for $N \rightarrow +\infty$. The pointwise convergence of (1.5), and is ensured (and there is normal convergence) when for instance $f \in C^3(\mathbb{T})$.

Poisson formula. The poisson formula connects the Fourier transform and the Fourier series to sampling and periodization operators. For some function $\hat{f}(\omega)$ defined on \mathbb{R} , its periodization reads

$$\hat{f}_P(\omega) \stackrel{\text{def.}}{=} \sum_n f(\omega - 2\pi n). \quad (1.6)$$

This formula makes sense if $\hat{f} \in L^1(\mathbb{R})$, and in this case $\|\hat{f}_P\|_{L^1(\mathbb{T})} \leq \|\hat{f}\|_1$. The Poisson formula, state in Proposition 1 bellow, corresponds to proving that the following diagram

$$\begin{array}{ccc} f(x) & \xrightarrow{\mathcal{F}} & \hat{f}(\omega) \\ \downarrow & & \downarrow \\ (\mathbf{sampling}) & \xrightarrow{\text{Fourier serie}} & \sum_n f(n) e^{-i\omega n} \\ & & \xrightarrow{\text{periodization}} \end{array}$$

is actually commutative.

Proposition 1 (Poisson formula). *Assume that \hat{f} has compact support and that $|f(x)| \leq C(1 + |x|)^{-3}$ for some C . Then one has*

$$\forall \omega \in \mathbb{R}, \quad \sum_n f(n) e^{-i\omega n} = \hat{f}_P(\omega). \quad (1.7)$$

Proof. Since \hat{f} is compactly supported, \hat{f}_P is well defined (it involves only a finite sum) and since f has fast decay, using (1.4), \hat{f}_P is C^1 . It is thus the sum of its Fourier transform

$$\hat{f}_P(\omega) = \sum_k c_k e^{ik\omega}, \quad (1.8)$$

where

$$c_k = \frac{1}{2\pi} \int_0^{2\pi} \hat{f}_P(\omega) e^{-ik\omega} d\omega = \frac{1}{2\pi} \int_0^{2\pi} \sum_n f(x - 2\pi n) e^{-ik\omega} d\omega.$$

One has

$$\int_0^{2\pi} \sum_n |f(x - 2\pi n) e^{-ik\omega}| d\omega = \int_{\mathbb{R}} |f|$$

which is bounded because $\hat{f} \in L^1(\mathbb{R})$ (it has a compact support and is C^1), so one can exchange the sum and integral

$$c_k = \sum_n \frac{1}{2\pi} \int_0^{2\pi} f(x - 2\pi n) e^{-ik\omega} d\omega = \frac{1}{2\pi} \int_{\mathbb{R}} f(x) e^{-ik\omega} d\omega = f(-k)$$

where we used the inverse Fourier transform formula (1.3), which is legit because $\hat{f} \in L^1(\mathbb{R})$. \square

Shannon theorem. Shannon sampling theorem state a sufficient condition ensuring that the sampling operator $f \mapsto (f(ns))_n$ is invertible for some sampling step size $s > 0$. It require that $\text{supp}(\hat{f}) \subset [-\pi/s, \pi/s]$, which, thanks to formula (1.3), implies that \hat{f} is C^∞ (in fact it is even analytic).

Theorem 1. *If $|f(x)| \leq C(1 + |x|)^{-3}$ for some C and $\text{supp}(\hat{f}) \subset [-\pi/s, \pi/s]$, then one has*

$$\forall x \in \mathbb{R}, \quad f(x) = \sum_n f(ns) \text{sinc}(x/s - n) \quad \text{where} \quad \text{sinc}(u) = \frac{\sin(\pi u)}{\pi u} \quad (1.9)$$

with uniform convergence.

Proof. The change of variable $g = f(s \cdot)$ results in $\hat{g} = s\hat{f}(s \cdot)$ so that we can restrict our attention to $s = 1$. The compact support hypothesis implies $\hat{f}(\omega) = 1_{[-\pi, \pi]}(\omega) \hat{f}_P(\omega)$. Combining the inversion formula (1.3) with Poisson formula (1.8)

$$f(x) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \hat{f}_P(\omega) e^{i\omega x} d\omega = \frac{1}{2\pi} \int_{-\pi}^{\pi} \sum_n f(n) e^{i\omega(x-n)} d\omega.$$

Since f has fast decay, $\int_{-\pi}^{\pi} \sum_n |f(n) e^{i\omega(x-n)}| d\omega = \sum_n |f(n)| < +\infty$, so that one can exchange summation and integration and obtain

$$f(x) = \sum_n f(n) \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{i\omega(x-n)} d\omega = \sum_n f(n) \text{sinc}(x - n).$$

\square

1.3 Shannon Source Coding Theorem

We consider an alphabet (x_1, \dots, x_K) of K symbols, and assume at our disposal some probability distribution over this alphabet, which is just an histogram $p = (p_1, \dots, p_K) \in \mathbb{R}_+^K$ in the simplex, i.e. $\sum_k p_k = 1$.

The entropy of such an histogram is

$$H(p) \stackrel{\text{def.}}{=} - \sum_k p_k \log_2(p_k)$$

with the convention $O \log_2(0) = 0$.

Lemma 1. *One has*

$$0 \leq H(p) \leq \log_2(K).$$

Proof. We consider the following constrained optimization problem

$$\min_p \left\{ f(p) ; g(p) = \sum_k p_k = 1 \right\}$$

where $f = -H$. According to the linked extrema theorem, at an optimum p^* , $\nabla f(p^*) = \lambda \nabla g(p^*)$ for some $\lambda \in \mathbb{R}$, so that here $\log(p_k^*) + 1 = \lambda$, i.e. $p_k^* = c$ is constant, and since $\sum_k p_k^* = 1$, one has $p_k^* = 1/K$ and thus $H(p) = \log_2(K)$. \square

A code $c_k = c(x_k)$ associate to each symbol x_k a code word $c_k \in \{0, 1\}^{\mathbb{N}}$ with a varying length $|c_k| \in \mathbb{N}^*$. We denote the average length associated to this code as

$$L(c) \stackrel{\text{def.}}{=} \sum_k p_k |c_k|.$$

A prefix code $c_k = c(x_k)$ is such that no word c_k is the beginning of another word c'_k . This is equivalent to be able to embed the $(c_k)_k$ as leaves of a binary tree T , with the code being output of a traversal from root to leaves (with a convention that going to a left (resp. right) child output a 0 (resp. a 1)). We denote $c = \text{Leaves}(T)$ such prefix property. The following fundamental lemma describes the set of prefix code using an inequality.

Lemma 2 (Kraft inequality). *(i) For a code c , if there exists a tree T such that $c = \text{Leaves}(T)$ then*

$$\sum_k 2^{-|c_k|} \leq 1. \quad (1.10)$$

(ii) Conversely, if $(\ell_k)_k$ are such that

$$\sum_k 2^{-\ell_k} \leq 1 \quad (1.11)$$

then there exists a code $c = \text{Leaves}(T)$ such that $|c_k| = \ell_k$.

Proof. \Rightarrow We suppose $c = \text{Leaves}(T)$. We denote $m = \max_k |c_k|$ and consider the full binary tree. Below each c_k , one has a sub-tree of height $m - |c_k|$. This sub-tree has $2^{m-|c_k|}$ leaves. Since all these sub-trees do not overlap, the total number of leaf do not exceed the total number of leaves 2^m of the full binary tree, hence

$$\sum_k 2^{m-|c_k|} \leq 2^m,$$

hence (1.10).

\Leftarrow Conversely, we assume (1.10) holds. Without loss of generality, we assume that $|c_1| \leq \dots \leq |c_K|$. We start by putting a sub-tree of height $2^{m-|c_1|}$. Since the second tree is smaller, one can put it immediately aside, and continue this way. Since $\sum_k 2^{m-|c_k|} \leq 2^m$, this ensure that we can stack side-by-side all these sub-tree, and this defines a proper sub-tree of the full binary tree. \square

We now are ready to state and prove Shannon theory for entropic coding.

Theorem 2. (i) If $c = \text{Leaves}(T)$ for some tree T , then

$$L(c) \geq H(p).$$

(ii) Conversely, there exists a code c with $c = \text{Leaves}(T)$ such that

$$L(c) \leq H(p) + 1.$$

Proof. First, we consider the following optimization problem

$$\min_{\ell=(\ell_k)_k} \left\{ f(\ell) \stackrel{\text{def.}}{=} \sum_k \ell_k p_k ; g(\ell) \stackrel{\text{def.}}{=} \sum_k 2^{-\ell_k} \leq 1 \right\}. \quad (1.12)$$

We first show that at an optimal ℓ^* , the constraint is saturated, i.g. $g(\ell^*) = 1$. Indeed, if $g(\ell^*) = 2^{-u} < 1$, with $u > 0$, we define $\ell'_k \stackrel{\text{def.}}{=} \ell_k^* - u$, which satisfies $g(\ell') = 1$ and also $f(\ell') = \sum_k (\ell_k - u)p_k < f(\ell^*)$, which is a contradiction. So we can restrict in (1.12) the constraint to $g(\ell) = 1$ and apply the linked extra theorem, which shows that necessarily, there exists $\lambda \in \mathbb{R}$ with $\nabla f(\ell^*) = \nabla g(\ell^*)$, i.e. $(p_k)_k = -\lambda \ln(2)(2^{-\ell_k^*})_k$. Since $\sum_k p_k = \sum_k 2^{-\ell_k^*} = 1$, we deduce that $\ell_k^* = -\log(p_k)$.

(i) If $c = \text{Leave}(T)$, the by Kraft inequality (1.10), necessarily $\ell_k = |c_k|$ satisfy the constraints of (1.12), and thus $H(p) = f(\ell^*) \leq f(\ell) = L(\ell)$.

(ii) We define $\ell_k \stackrel{\text{def.}}{=} \lceil -\log_2(p_k) \rceil \in \mathbb{N}^*$. Then $\sum_k 2^{-\ell_k} \leq \sum_k 2^{\log_2(p_k)} = 1$, so that these lengths satisfy (1.11). Thanks to Proposition 2 (ii), there thus exists a prefix code c with $|c_k| = \lceil -\log_2(p_k) \rceil$. Furthermore

$$L(c) = \sum_k p_k \lceil -\log_2(p_k) \rceil \leq \sum_k p_k (-\log_2(p_k) + 1) = H(p) + 1.$$

□

Chapter 2

Fourier Transforms

The main references for this chapter is [28]. The Fourier transforms offers a perfect blend of analysis (solution of PDEs, approximation of functions), algebra (characters of groups, representation theory) and computer science (the FFT). This chapter offers a glimpse of all these different facets.

2.1 Hilbert spaces and Fourier Transforms

2.1.1 Hilbertian bases.

An Hilbert space $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ is complete. If it is separable, it can be equipped with an Hilbertian orthogonal basis $(\varphi_n)_{n \in \mathbb{N}}$, which means that one can expand any $f \in \mathcal{H}$ as

$$f = \sum_n \langle f, \varphi_n \rangle \varphi_n$$

where the convergence is in the sense of $\|f\|^2 \stackrel{\text{def.}}{=} \langle f, f \rangle$, i.e. $\|f - \sum_{n=0}^N \langle f, \varphi_n \rangle \varphi_n\| \rightarrow 0$ as $N \rightarrow +\infty$. One also have the conservation of energy

$$\|f\|^2 = \sum_n \langle f, \varphi_n \rangle^2.$$

A way to construct such an ortho-basis is using Gram-Schmidt orthogonalization procedure. On $L^2([0, 1])$ equipped with the usual inner product, orthogonalization of monomials defines the Legendre polynomials. On $L^2(\mathbb{R})$ equipped with a Gaussian measure, this leads to functions of the form $P_n(x)e^{-x^2}$ where P_n are Laguerre polynomials. Intuitively, orthogonality forces φ_n to have n “oscillations”, e.g. orthogonal polynomials have exactly n zeros.

Figure 2.1, left, shows examples of the real part of Fourier atoms.

2.1.2 Fourier basis on $\mathbb{R}/2\pi\mathbb{Z}$.

On $L^2(\mathbb{T})$ where $\mathbb{T} = \mathbb{R}/2\pi\mathbb{Z}$, equipped with $\langle f, g \rangle \stackrel{\text{def.}}{=} \frac{1}{2\pi} \int_{\mathbb{T}} f(x)\bar{g}(x)dx$, one can use the Fourier basis

$$\varphi_n(x) \stackrel{\text{def.}}{=} e^{inx} \quad \text{for } n \in \mathbb{Z}. \tag{2.1}$$

One thus has

$$f = \sum_n \hat{f}_n e^{inx} \quad \text{where } \hat{f}_n \stackrel{\text{def.}}{=} \frac{1}{2\pi} \int_0^{2\pi} f(x)e^{-inx} dx, \tag{2.2}$$

in $L^2(\mathbb{T})$ sense. Pointwise convergence is delicate, see Section 1.2.

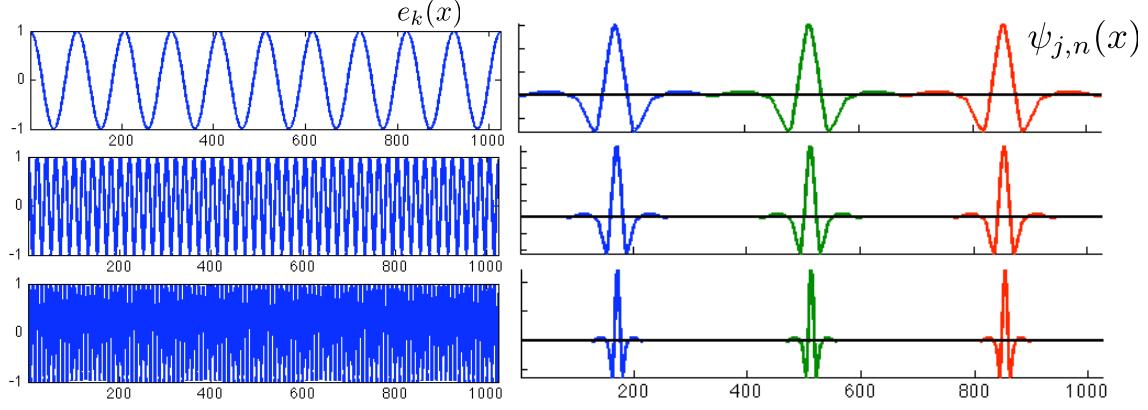


Figure 2.1: Left: 1D Fourier (real part), right: wavelet bases.

We recall that for $f \in L^1(\mathbb{R})$, its Fourier transform is defined as

$$\forall \omega \in \mathbb{R}, \quad \hat{f}(\omega) \stackrel{\text{def.}}{=} \int_{\mathbb{R}} f(x) e^{-ix\omega} dx.$$

and this is extended to $L^2(\mathbb{R})$ by density.

The connexion between the Fourier transform on \mathbb{R} and the Fourier coefficients on \mathbb{T} is given by the following diagram

$$\begin{array}{ccccc} f(x) & \xrightarrow{\mathcal{F}} & \hat{f}(\omega) & & \\ \downarrow \text{sampling} & & \downarrow & & \text{periodization .} \\ (f(n))_n & \xrightarrow{\text{Fourier serie}} & \sum_n f(n) e^{-i\omega n} & & \end{array}$$

Its commutativity sates

$$\sum_n f(n) e^{-i\omega n} = \sum_n \hat{f}(\omega - 2\pi n) \quad (2.3)$$

and this is in fact the celebrated Poisson formula (Proposition 1).

2.2 Convolution on \mathbb{R} and \mathbb{T}

2.2.1 Convolution

On $\mathbb{X} = \mathbb{R}$ or \mathbb{T} , one defines

$$f \star g(x) = \int_{\mathbb{X}} f(t) g(x-t) dt. \quad (2.4)$$

Young's inequality shows that this quantity is well defined if $(f, g) \in L^p(\mathbb{X}) \times L^q(\mathbb{X})$

$$\frac{1}{p} + \frac{1}{q} = 1 + \frac{1}{r} \implies f \star g \in L^r(\mathbb{X}) \quad \text{and} \quad \|f \star g\|_{L^r(\mathbb{X})} \leq \|f\|_{L^p(\mathbb{X})} \|g\|_{L^q(\mathbb{X})}. \quad (2.5)$$

This shows that if $f \in L^1(\mathbb{X})$, then one has the map $g \in L^p(\mathbb{X}) \mapsto f \star g \in L^p(\mathbb{X})$ is a continuous map on $L^p(\mathbb{X})$. Furthermore, when $r = \infty$, $f \star g \in \mathcal{C}^0(\mathbb{X})$ is a

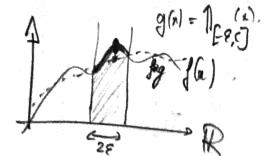


Figure 2.3: Convolution on \mathbb{R} .

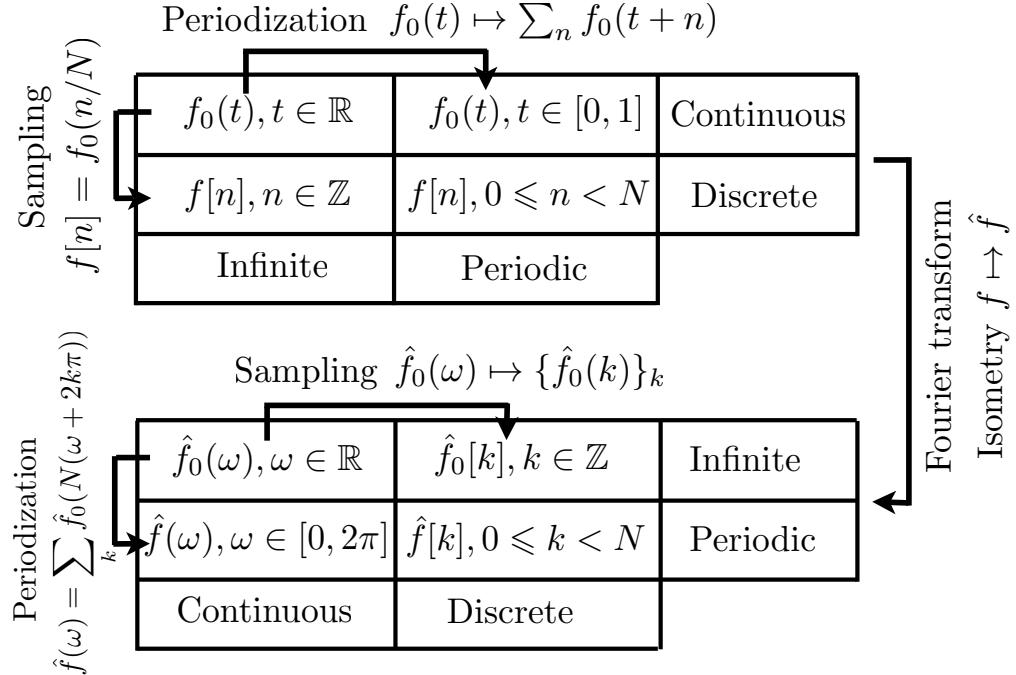


Figure 2.2: The four different settings for Fourier analysis, and the sampling-periodization relationship.

continuous function (which shows some regularizing effect). Note that for $\mathbb{X} = \mathbb{T}$, $p < q \implies L^q(\mathbb{T}) \subset L^p(\mathbb{T})$, so that $L^\infty(\mathbb{X})$ is the smallest space.

Convolution is mostly used in order to regularize functions. For instance, if $f \in L^1(\mathbb{X})$ and $g \in C^1(\mathbb{X})$ is bounded, then $f \star g$ is differentiable and $(f \star g)' = f \star g'$. This is used to produce smooth approximate identity $(\rho_\varepsilon = \frac{1}{\varepsilon} \rho(\cdot/\varepsilon))_\varepsilon$ with convergence $f \star \rho_\varepsilon \rightarrow f$ in $L^p(\mathbb{X})$ for $1 \leq p < +\infty$ of smooth approximations (and convergence in $L^\infty(\mathbb{X})$ if f is uniformly continuous). This is also used for denoising applications in signal and image processing.

For $(f, g) \in L^1(\mathbb{X})^2$ (so on $\mathbb{X} = \mathbb{T}$, also in any $L^p(\mathbb{X})$), one has

$$\mathcal{F}(f \star g) = \hat{f} \odot \hat{g} \quad (2.6)$$

which means that \mathcal{F} is a morphism of algebra. For instance if $\mathbb{X} = \mathbb{R}$, its range is included in the algebra of continuous functions with vanishing limits in $\pm\infty$.

$$\begin{array}{ccc} (f, g) & \xrightarrow{*} & f \star g \\ \downarrow \mathcal{F} & & \uparrow \mathcal{F}^{-1} \\ (\hat{f}, \hat{g}) & \xrightarrow{\odot} & \hat{f} \odot \hat{g} \end{array}$$

2.2.2 Translation Invariant Operators

Translation invariant operators (which commutes with translation) are fundamental because in most situations, input (signal, image, etc) should be processed without spatial preference. The following propositions shows that any translation invariant¹ (i.e. which commutes with translations) operator is actually a “convolution” against a distribution with bounded Fourier transform. The proof and conclusion (regularity of the convolution kernel) vary depending on the topology on the input and output spaces. We first study the case of convolution mapping to continuous functions.

Proposition 2. *We define $T_\tau f = f(\cdot - \tau)$. A bounded linear operator $H : (L^2(\mathbb{X}), \|\cdot\|_2) \rightarrow (C^0(\mathbb{X}), \|\cdot\|_\infty)$*

¹One should rather actually say “translation equivariant”.

Figure 2.6: Commutative diagram of convolution-Fourier.

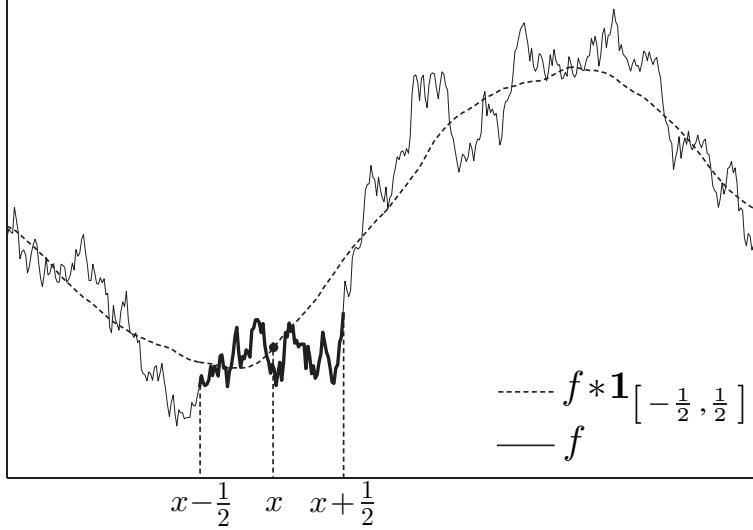


Figure 2.4: Signal filtering with a box filter (running average).

is such that for all τ , $H \circ T_\tau = T_\tau \circ H$ if and only if

$$\forall f \in L^2(\mathbb{T}), \quad H(f) = f \star g$$

with $g \in L^2(\mathbb{X})$.

Proof. Thanks to (2.5) (and the remark in the case $r = \infty$), $T : f \mapsto f \star g$ with $g \in L^2(\mathbb{X})$ is indeed a continuous operator from $L^2(\mathbb{X})$ to $C^0(\mathbb{X})$. Furthermore

$$(H \circ T_\tau)(f) = \int_{\mathbb{X}} f((\cdot - \tau) - y)g(y)d\tau = (f \star g)(\cdot - \tau) = T_\tau(Hf),$$

so that such an H is translation-invariant.

Conversely, we define $\ell : f \mapsto H(f)(0) \in \mathbb{R}$, which is legit since $H(f) \in C^0(\mathbb{X})$. Since H is continuous, there exists C such that $\|Hf\|_\infty \leq C\|f\|_2$, and hence $|\ell(f)| \leq C\|f\|_2$, so that f is a continuous linear form on the Hilbert space $L^2(\mathbb{X})$. Hence, according to Fréchet-Riesz theorem, there exists $h \in L^2(\mathbb{X})$ such that $\ell(f) = \langle f, h \rangle$. Hence, $\forall x \in \mathbb{X}$,

$$H(f)(x) = T_{-x}(Hf)(0) = H(T_{-x}f)(0) = \ell(T_{-x}f) = \langle T_{-x}f, h \rangle = \int_{\mathbb{X}} f(y + x)h(y)dy = f \star \bar{h}(x).$$

where $g \stackrel{\text{def}}{=} \bar{h} = h(-\cdot) \in L^2(\mathbb{X})$. □

We now study, on \mathbb{T} , the case of convolution which can output non-continuous functions. In this case, the kernel can be a “distribution”, so the convolution is defined over the Fourier domain.

Proposition 3. A bounded linear operator $H : L^2(\mathbb{T}) \rightarrow L^2(\mathbb{T})$ is such that for all τ , $H \circ T_\tau = T_\tau \circ H$ if and only if

$$\forall f \in L^2(\mathbb{T}), \quad \mathcal{F}(H(f)) = \hat{f} \odot c$$

where $c \in \ell^\infty(\mathbb{Z})$ (a bounded sequence).

$$\begin{array}{ccc} \int & \xrightarrow{H} & H(f) \\ T_\tau \downarrow & & \uparrow \overline{1_{-\tau}} \\ \int f(\cdot - \tau) & \xrightarrow{H} & H(f(\cdot - \tau)) \end{array}$$

Figure 2.7:
Commutative
diagram for trans-
lation invariance.

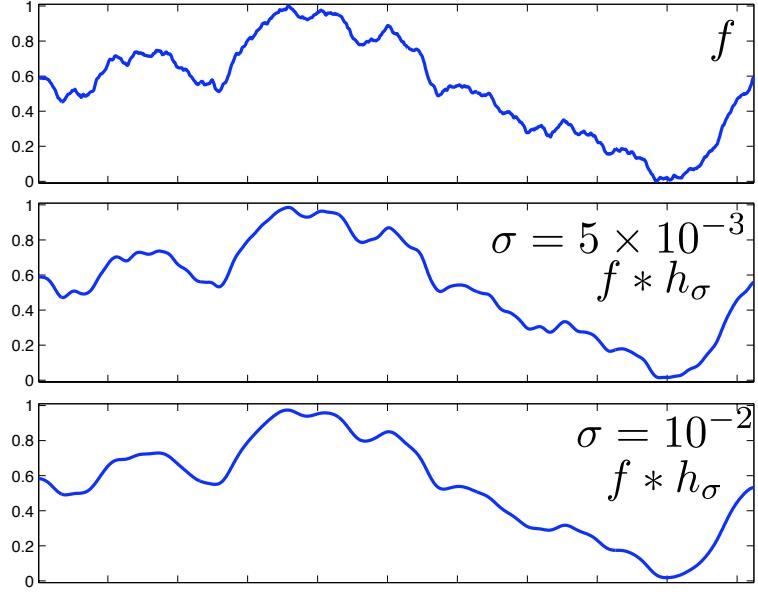


Figure 2.5: Filtering an irregular signal with a Gaussian filter of increasing filter size σ .

Proof. We denote $\varphi_n \stackrel{\text{def.}}{=} e^{in\cdot}$. One has

$$H(\varphi_n) = e^{in\tau} H(T_\tau(\varphi_n)) = e^{in\tau} T_\tau(H(\varphi_n)).$$

Thus, for all n ,

$$\langle H(\varphi_n), \varphi_m \rangle = e^{in\tau} \langle T_\tau \circ H(\varphi_n), \varphi_m \rangle = e^{in\tau} \langle H(\varphi_n), T_{-\tau}(\varphi_m) \rangle = e^{i(n-m)\tau} \langle H(\varphi_n), \varphi_m \rangle$$

So for $n \neq m$, $\langle H(\varphi_n), \varphi_m \rangle = 0$, and we define $c_n \stackrel{\text{def.}}{=} \langle H(\varphi_n), \varphi_n \rangle$. Since H is continuous, $\|Hf\|_{L^2(\mathbb{T})} \leq C\|f\|_{L^2(\mathbb{T})}$ for some constant C , and thus by Cauchy-Schwartz

$$|c_n| = |\langle H(\varphi_n), \varphi_n \rangle| \leq \|H(\varphi_n)\| \|\varphi_n\| \leq C$$

because $\|\varphi_n\| = 1$, so that $c \in \ell^\infty(\mathbb{Z})$. By continuity, recalling that by definition $\hat{f}_n \stackrel{\text{def.}}{=} \langle f, \varphi_n \rangle$,

$$H(f) = \lim_N H\left(\sum_{n=-N}^N \hat{f}_n \varphi_n\right) = \lim_N \sum_{n=-N}^N \hat{f}_n H(\varphi_n) = \lim_N \sum_{n=-N}^N c_n \hat{f}_n \varphi_n = \sum_{n \in \mathbb{Z}} c_n \hat{f}_n \varphi_n$$

so that in particular one has the desired result. \square

This theorem thus states that translation invariant operators are those which are “diagonal” in the Fourier ortho-basis.

2.2.3 Revisiting Poisson formula using distributions.

Informally, the Fourier series

$$\sum_n f(n) e^{-i\omega n}$$

can be thought as the Fourier transform $\mathcal{F}(\Pi_1 \odot f)$ of the discrete distribution

$$\Pi_1 \odot f = \sum_n f(n) \delta_n \quad \text{where} \quad \Pi_s = \sum_n \delta_{sn}$$

for $s \in \mathbb{R}$, where δ_a is the Dirac mass at location $a \in \mathbb{R}$, i.e. the distribution such that $\int f d(\delta_a) = f(a)$ for any continuous f . Indeed, one can multiply a distribution by a continuous function, and the definition of the Fourier transform of a distribution μ is a distributions $\mathcal{F}(\mu)$ such that that

$$\forall g \in \mathcal{S}(\mathbb{R}), \quad \int_{\mathbb{R}} g(x) d\mathcal{F}(\mu) = \int_{\mathbb{R}} \mathcal{F}^*(g) d\mu, \quad \text{where } \mathcal{F}^*(g) \stackrel{\text{def.}}{=} \int_{\mathbb{R}} g(x) e^{ix} dx,$$

where $\mathcal{S}(\mathbb{R})$ are smooth and rapidly decaying (Schwartz class) functions.

The Poisson formula (2.3) can thus be interpreted as

$$\mathcal{F}(\Pi_1 \odot f) = \sum_n \hat{f}(\cdot - 2\pi n) = \int_{\mathbb{R}} \hat{f}(\cdot - \omega) d\Pi_{2\pi}(\omega) = \hat{f} \star \Pi_{2\pi}$$

Since $\mathcal{F}^{-1} = \frac{1}{2\pi} \mathcal{S} \circ \mathcal{F}$ where $\mathcal{S}(f) = f(-\cdot)$, one has, applying this operator on both sides

$$\Pi_1 \odot f = \frac{1}{2\pi} \mathcal{S} \circ \mathcal{F}(\hat{f} \star \Pi_{2\pi}) = \mathcal{S}\left(\frac{1}{2\pi} \mathcal{F}(\hat{f}) \odot \hat{\Pi}_{2\pi}\right) = \mathcal{S}\left(\frac{1}{2\pi} \mathcal{F}(\hat{f})\right) \odot \mathcal{S}(\hat{\Pi}_{2\pi}) = \hat{\Pi}_{2\pi} \odot f.$$

This can be interpreted as the relation

$$\hat{\Pi}_{2\pi} = \Pi_1 \implies \hat{\Pi}_1 = 2\pi \Pi_{2\pi}.$$

To intuitively understand this relation, one can compute a finite Fourier series

$$\sum_{n=-N}^N e^{-inx} = \frac{\sin((N+1/2)x)}{\sin(x/2)}$$

which is a smooth function which grows unbounded with $N \rightarrow +\infty$ as $N \rightarrow +\infty$.

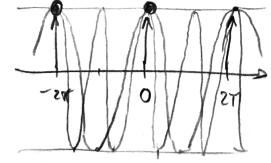


Figure 2.8: Sine wave being summed in the Poisson formula.

2.3 Finite Fourier Transform and Convolution

2.3.1 Discrete Ortho-bases

Discrete signals are finite dimensional vector $f \in \mathbb{C}^N$ where N is the number of samples and where each f_n is the value of the signal at a 1D or 2D location. For a 2-D images $f \in \mathbb{C}^N \simeq \mathbb{C}^{N_0 \times N_0}$, $N = N_0 \times N_0$, where N_0 is the number of pixels along each direction.

Discrete signals and images are processed using a discrete inner product that mimics the continuous L^2 inner product

$$\langle f, g \rangle = \sum_{n=0}^{N-1} f_n \bar{g}_n.$$

One thus defines a distance between discretized vectors as

$$\|f - g\|^2 = \sum_{n=0}^{N-1} |f_n - g_n|^2.$$

Exactly as in the continuous case, a discrete orthogonal basis $\{\psi_m\}_{0 \leq m < N}$ of \mathbb{C}^N , satisfies

$$\langle \psi_m, \psi_{m'} \rangle = \delta_{m-m'}. \tag{2.7}$$

The decomposition of a signal in such an ortho-basis is written

$$f = \sum_{m=0}^{N-1} \langle f, \psi_m \rangle \psi_m.$$

It satisfies a conservation of energy

$$\|f\|^2 = \sum_{n=0}^{N-1} |f_n|^2 = \sum_{m=0}^{N-1} |\langle f, \psi_m \rangle|^2$$

Computing the set of all inner product $\{\langle f, \psi_m \rangle\}_{0 \leq m < N}$ is done in a brute force way in $O(N^2)$ operations. This is not feasible for large datasets where N is of the order of millions. When designing an ortho-basis, one should keep this limitation in mind and enforce some structure in the basis elements so that the decomposition can be computed with fast algorithm. This is the case for the Fourier and wavelet bases, that enjoy respectively $O(N \log(N))$ and $O(N)$ algorithms.

2.3.2 Discrete Fourier transform

We denote $f = (f_n)_{n=0}^{N-1} \in \mathbb{R}^N$, but we insist that such vector should really be understood as being indexed by $n \in \mathbb{Z}/N\mathbb{Z}$, which is a finite commutative group for the addition. This corresponds to using periodic boundary conditions.

The discrete Fourier transform is defined as

$$\forall k = 0, \dots, N-1, \quad \hat{f}_k \stackrel{\text{def.}}{=} \sum_{n=0}^{N-1} f_n e^{-\frac{2i\pi}{N} kn} = \langle f, \varphi_k \rangle \quad \text{where} \quad \varphi_k \stackrel{\text{def.}}{=} (e^{\frac{2i\pi}{N} kn})_{n=0}^{N-1} \in \mathbb{C}^N \quad (2.8)$$

where the canonical inner product on \mathbb{C}^N is $\langle u, v \rangle = \sum_{n=1}^N u_n \bar{v}_n$ for $(u, v) \in (\mathbb{C}^N)^2$. This definition can intuitively be motivated by sampling the Fourier basis $x \mapsto e^{ikx}$ on $\mathbb{R}/2\pi\mathbb{Z}$ at equi-spaced points $(\frac{2\pi}{N}n)_{n=0}^{N-1}$. The following proposition shows that this corresponds to a decomposition in an ortho-basis.

Proposition 4. *One has*

$$\langle \varphi_k, \varphi_\ell \rangle = \begin{cases} N & \text{if } k = \ell, \\ 0 & \text{otherwise.} \end{cases}$$

In particular, this implies

$$\forall n = 0, \dots, N-1, \quad f_n = \frac{1}{N} \sum_k \hat{f}_k e^{\frac{2i\pi}{N} kn}. \quad (2.9)$$

Proof. One has, if $k \neq \ell$

$$\langle \varphi_k, \varphi_\ell \rangle = \sum_n e^{\frac{2i\pi}{N}(k-\ell)n} = \frac{1 - e^{\frac{2i\pi}{N}(k-\ell)N}}{1 - e^{\frac{2i\pi}{N}(k-\ell)}} = 0$$

which is the sum of a geometric serie (equivalently, sum of equi-spaced points on a circle). The inversion formula is simply $f = \sum_k \langle f, \varphi_k \rangle \frac{\varphi_k}{\|\varphi_k\|^2}$. \square

2.3.3 Fast Fourier transform

Assuming $N = 2N'$, one has

$$\begin{aligned} \hat{f}_{2k} &= \sum_{n=0}^{N'-1} (f_n + f_{n+N/2}) e^{-\frac{2i\pi}{N'} kn} \\ \hat{f}_{2k+1} &= \sum_{n=0}^{N'-1} e^{-\frac{2i\pi}{N'} n} (f_n - f_{n+N/2}) e^{-\frac{2i\pi}{N'} kn}. \end{aligned}$$

For the second line, we used the computation

$$e^{-\frac{2i\pi}{N}(2k+1)(n+N/2)} = e^{-\frac{2i\pi}{N}(2kn+kN+n+N/2)} = -e^{-\frac{2i\pi}{N'}n} e^{-\frac{2i\pi}{N'}kn}.$$

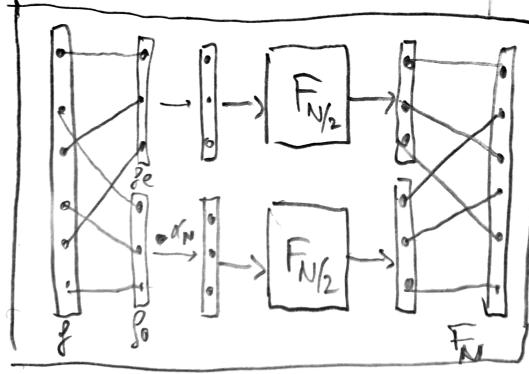


Figure 2.9: Diagram of one step of the FFT.

Denoting $\mathcal{F}_N(f) = \hat{f}$ the discrete Fourier transform on \mathbb{R}^N , and introducing the notation $f_e = (f_n + f_{n+N/2})_n \in \mathbb{R}^{N'}$, $f_o = (f_n - f_{n+N/2})_n \in \mathbb{R}^{N'}$ and $\alpha_N = (e^{-\frac{2i\pi}{N'} n})_n \in \mathbb{R}^{N'}$, one has the following recursion formula

$$\mathcal{F}_N(f) = \mathcal{I}_N(\mathcal{F}_{N/2}(f_e), \mathcal{F}_{N/2}(f_o \odot \alpha_N))$$

where \mathcal{I}_N is the “interleaving” operator, defined by $\mathcal{I}_N(a, b) \stackrel{\text{def.}}{=} (a_1, b_1, a_2, b_2, \dots, a_{N'}, b_{N'})$. These iterations define the so-called Fast Fourier Transform algorithm, which works here when N is a power of 2. These iterations can be extended to arbitrary number N , but a workaround is to simply pad with 0 (or use more complicated extensions) to have vector with size that are power of 2.

Denoting $C(N)$ the numerical complexity (number of elementary operations) associated to the computation of \hat{f} , one thus has

$$C(N) = 2C(N/2) + NK \quad (2.10)$$

where K is the complexity of N complex additions and $N/2$ multiplications. Making the change of variable

$$\ell \stackrel{\text{def.}}{=} \log_2(N) \quad \text{and} \quad T(\ell) \stackrel{\text{def.}}{=} \frac{C(N)}{N}$$

i.e. $C(N) = 2^\ell T(\ell)$, the relation (2.10) becomes

$$2^\ell T(\ell) = 2 \times 2^{\ell-1} T(\ell-1) + 2^\ell K \implies T(\ell) = T(\ell-1) + K \implies T(\ell) = T(0) + K\ell$$

and using the fact that $T(0) = C(1)/1 = 0$, one obtains

$$C(N) = KN \log_2(N).$$

This complexity should be contrasted with the complexity $O(N^2)$ of directly computing the N coefficients (2.8), each involving a sum of size N .

2.3.4 Finite convolution

For $(f, g) \in (\mathbb{R}^N)^2$, one defines $f \star g \in \mathbb{R}^N$ as

$$\forall n = 0, \dots, N-1, \quad (f \star g)_n \stackrel{\text{def.}}{=} \sum_{k=0}^{N-1} f_k g_{n-k} = \sum_{k+\ell=n} f_k g_\ell \quad (2.11)$$

where one should be careful that here $+$ and $-$ should be understood modulo N (vectors should be seen as being defined on the group $\mathbb{Z}/N\mathbb{Z}$, or equivalently, one uses periodic boundary conditions). This defines an

commutative algebra structure $(\mathbb{R}^N, +, \star)$, with neutral element the ‘‘Dirac’’ $\delta_0 \stackrel{\text{def.}}{=} (1, 0, \dots, 0)^\top \in \mathbb{R}^N$. The following proposition shows that $\mathcal{F} : f \mapsto \hat{f}$ is an algebra bijective isometry (up to a scaling by \sqrt{N} of the norm) mapping to $(\mathbb{R}^N, +, \odot)$ with neutral element $\mathbb{1}_N = (1, \dots, 1) \in \mathbb{R}^N$.

Proposition 5. *One has $\mathcal{F}(f \star g) = \hat{f} \odot \hat{g}$.*

Proof. We denote $T : g \mapsto f \star g$. One has

$$(T\varphi_\ell)_n = \sum_k f_k e^{\frac{2i\pi}{N}\ell(n-k)} = e^{\frac{2i\pi}{N}\ell n} \hat{f}_\ell.$$

This shows that $(\varphi_\ell)_\ell$ are the N eigenvectors of T with associated eigenvalues \hat{f}_ℓ . So T is diagonalizable in this basis. Denoting $F = (e^{-\frac{2i\pi}{N}kn})_{k,n}$ the matrix of the Fourier transform, the Fourier inversion formula (2.9) reads $F^{-1} = \frac{1}{N} F^*$ where $F^* = \bar{F}^\top$ is the adjoint matrix (trans-conjugate). The diagonalization of T now reads

$$T = F^{-1} \operatorname{diag}(\hat{f}) F = \implies \mathcal{F}(Tg) = \operatorname{diag}(\hat{f}) F g \implies \mathcal{F}(f \star g) = \operatorname{diag}(\hat{f}) \hat{g}.$$

□

This proposition shows that one can compute in $O(N \log(N))$ operation via the formula

$$f \star g = \mathbb{F}^{-1}(\hat{f} \odot \hat{g}).$$

This is very advantageous with respect to the naive implementation of formula (2.11), in the case where f and g have large support. In case where $|\operatorname{Supp}(g)| = P$ is small, then direct implementation is $O(PN)$ which might be advantageous. An example is $g = [1, 1, 0, \dots, 0, 1]/3$, the moving average, where

$$(f \star g)_n = \frac{f_{n-1} + f_n + f_{n+1}}{3}$$

needs $3N$ operations.

An example of application of the FFT is the multiplication of large polynomial, and thus of large integers (viewing the expansion in a certain basis as a polynomial). Indeed

$$\left(\sum_{i=0}^A a_i X_i \right) \left(\sum_{j=0}^B b_j X^j \right) = \sum_{k=0}^{A+B} \left(\sum_{i+j=k} a_i b_j \right) X^k$$

One can write $\sum_{i+j=k} a_i b_j = (\bar{a} \star \bar{b})_k$ when one defines $\bar{a}, \bar{b} \in \mathbb{R}^{A+B}$ by zero padding.

2.4 Discretisation Issues

Beside computing convolutions, another major application of the FFT is to approximate the Fourier transform and its inverse, thus leading to a computationally efficient spectral interpolation method.

2.4.1 Fourier approximation via spatial zero padding.

It is possible to view the discrete finite Fourier transform (2.8) as a first order approximation to compute Fourier coefficients, or rather actually samples from the Fourier transform (1.2). Supposing that f is a smooth enough function supported on $[0, 1]$, we consider the discrete Fourier transform of the vector $f^Q \stackrel{\text{def.}}{=} (f(n/N))_{n=0^Q-1} \in \mathbb{R}^Q$ where $Q \geq N$ induced a padding by 0 (since $f(n/N) = 0$ for $n > N$)

$$\forall k \in [-\frac{Q}{2}, \frac{Q}{2}], \quad \frac{1}{N} \hat{f}_k^Q = \frac{1}{N} \sum_{n=0}^{N-1} f\left(\frac{n}{N}\right) e^{-\frac{2i\pi}{Q} nk} \approx \int_0^1 f(x) e^{-\frac{2ki\pi}{T} x} dx = \hat{f}\left(\frac{2k\pi}{T}\right) \quad \text{where } T \stackrel{\text{def.}}{=} \frac{Q}{N}.$$

The approximation is first order accurate, i.e. $O(1/N)$ for a C^1 function f . Increasing the amount Q of zero padding is a way to compute larger frequencies. Increasing the discretization precision N is on contrary a way to increase the precision of the Fourier sampling (using smaller step size $2\pi/T$).

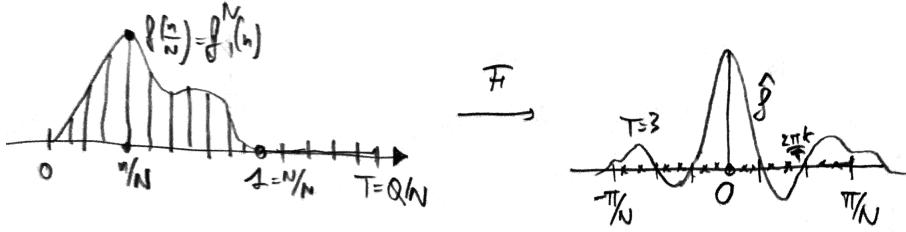


Figure 2.10: Fourier transform approximation by zero-padding in the spatial domain.

2.4.2 Fourier approximation via spatial zero padding.

If one has at its disposal N uniform discrete samples $f^N = (f_n^N)_{n=0}^{N-1}$, one can compute its discrete Fourier transform $\mathcal{F}(f^N) = \hat{f}^N$ (in $O(N \log(N))$ operations with the FFT),

$$\hat{f}_k^N \stackrel{\text{def.}}{=} \sum_{n=0}^{N-1} f_n^N e^{-\frac{2i\pi}{N} nk},$$

and then zero-pad it to obtain a vector of length Q . For simplicity, we assume $N = 2N' + 1$ is odd, and this computation can be also done (but is more involved) with even size. Indexing the frequencies as $-N' \leq k \leq N'$ The padding vector is of the form,

$$\tilde{f}^Q \stackrel{\text{def.}}{=} (0, \dots, 0, \hat{f}^N, 0, \dots, 0) \in \mathbb{R}^Q$$

One can then compute the (with a normalization constant Q/N) inverse discrete Fourier transform of size Q (in $O(Q \log(Q))$ operations with the FFT) to obtain

$$\begin{aligned} \frac{Q}{N} \mathcal{F}^{-1}(\tilde{f}^Q)_\ell &= \frac{Q}{N} \times \frac{1}{N} \sum_{k=-N'}^{N'} \hat{f}_k^N e^{\frac{2i\pi}{Q} \ell k} = \frac{1}{N} \sum_{k=-N'}^{N'} \sum_{n=0}^{N-1} f_n^N e^{\frac{2i\pi}{N} nk} e^{\frac{2i\pi}{Q} \ell k} \\ &= \sum_{n=0}^{N-1} f_n^N \frac{1}{N} \sum_{k=-N'}^{N'} e^{2i\pi(-\frac{n}{N} + \frac{\ell}{Q})k} = \sum_{n=0}^{N-1} f_n^N \frac{\sin\left[\pi N \left(\frac{\ell}{Q} - \frac{n}{N}\right)\right]}{N \sin\left[\pi \left(\frac{\ell}{Q} - \frac{n}{N}\right)\right]} \\ &= \sum_{n=0}^{N-1} f_n^N \operatorname{sinc}_N\left(\frac{\ell}{T} - n\right) \quad \text{where } T \stackrel{\text{def.}}{=} \frac{Q}{N} \quad \text{and} \quad \operatorname{sinc}_N(u) \stackrel{\text{def.}}{=} \frac{\sin(\pi u)}{N \sin(\pi u/N)}. \end{aligned}$$

Here we use the following summation rule for geometric series for $\rho = e^{i\omega}$, $a = -b$, $\omega = 2\pi\left(-\frac{n}{N} + \frac{\ell}{Q}\right)$,

$$\sum_{i=a}^b \rho^i = \frac{\rho^{a-\frac{1}{2}} - \rho^{b+\frac{1}{2}}}{\rho^{-\frac{1}{2}} - \rho^{\frac{1}{2}}} = \frac{\sin((b+\frac{1}{2})\omega)}{\sin(\omega/2)}.$$

This zero-padding method leads to a discrete version of the Shannon interpolation formula (1.9), which allows to comput the interpolation on a grid of size Q are cost $O(Q \log(Q))$. Increasing N increases the accuracy of the formula, since $\operatorname{sinc}_N \rightarrow \operatorname{sinc}$ as $N \rightarrow +\infty$.

2.5 Fourier in Multiple Dimensions

The Fourier transform is extended from 1-D to arbitrary finite dimension $d > 1$ by tensor product.

2.5.1 On Continuous Domains

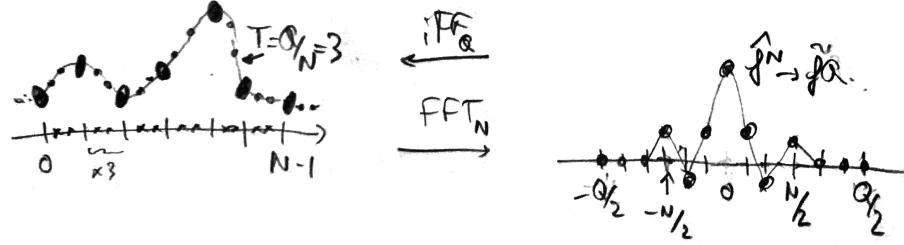


Figure 2.11: Interpolation by zero-padding in the frequency domain.

On \mathbb{R}^d . The crux of the power of Fourier transform in arbitrary dimension is that a product of elementary 1-D sine waves is still a sine wave

$$\prod_{\ell=1}^d e^{ix_\ell \omega_\ell} = e^{i\langle x, \omega \rangle}$$

moving orthogonally to the wave vector $\omega = (\omega_\ell)_{\ell=1}^d \in \mathbb{R}^d$. Here $\langle x, \omega \rangle = \sum_\ell x_\ell \omega_\ell$ is the canonical inner product on \mathbb{R}^d .

The definition of the Fourier transform and its inverse are

$$\begin{aligned} \forall \omega \in \mathbb{R}^d, \quad \hat{f}(\omega) &\stackrel{\text{def.}}{=} \int_{\mathbb{R}^d} f(x) e^{-i\langle x, \omega \rangle} dx, \\ \forall x \in \mathbb{R}^d, \quad f(x) &= \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \hat{f}(\omega) e^{i\langle x, \omega \rangle} d\omega, \end{aligned}$$

under hypotheses of integrability matching exactly those in 1-D.

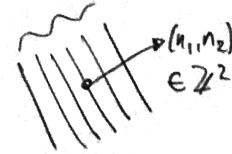


Figure 2.12: 2-D sine wave.

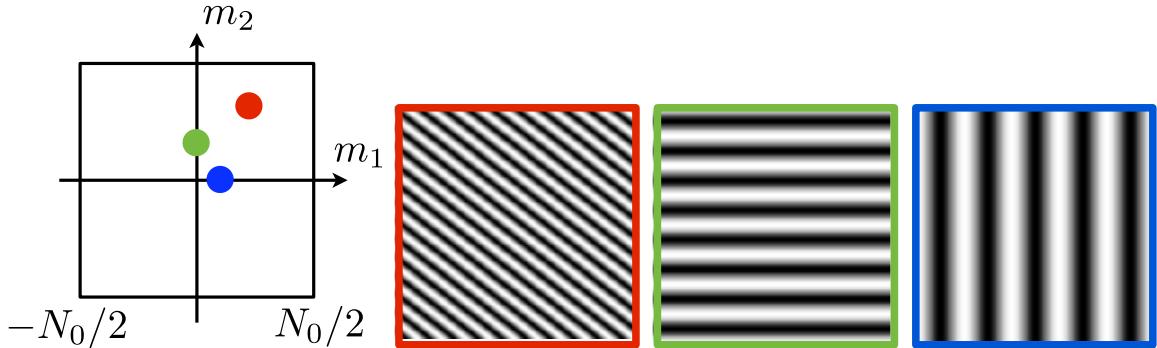


Figure 2.13: 2D Fourier orthogonal bases.

On $(\mathbb{R}/2\pi\mathbb{Z})^d$. Given an Hilbertian basis $(\varphi_{n_1})_{n_1 \in \mathbb{N}}$ of $L^2(\mathbb{X})$, one construct an Hilbertian basis of $L^2(\mathbb{X}^d)$ by tensorization

$$\forall n = (n_1, \dots, n_d) \in \mathbb{N}^d, \quad \forall x \in \mathbb{X}^d, \quad \varphi_n(x) = \varphi_{n_1}(x_1) \dots \varphi_{n_d}(x_d). \quad (2.12)$$

Orthogonality is simple to check, and one can also prove convergence for sum of the form $\sum_{\|n\|_\infty \leq N} \langle f, \varphi_n \rangle \varphi_n \rightarrow f$ in $L^2(\mathbb{X}^d)$.

For the multi-dimensional torus $(\mathbb{R}/2\pi\mathbb{Z})^d$, using the Fourier basis (2.1), this leads to consider the basis

$$\forall n \in \mathbb{R}^d, \quad \varphi_n(x) = e^{i\langle x, n \rangle}$$

which is indeed an Hilbertian orthonormal basis for the inner product $\langle f, g \rangle \stackrel{\text{def.}}{=} \frac{1}{(2\pi)^d} \int_{\mathbb{T}^d} f(x)\bar{g}(x)dx$. This defines the Fourier transform and the reconstruction formula on $L^2(\mathbb{T}^d)$

$$\hat{f}_n \stackrel{\text{def.}}{=} \frac{1}{(2\pi)^d} \int_{\mathbb{T}^d} f(x)e^{-i\langle x, n \rangle} dx \quad \text{and} \quad f = \sum_{n \in \mathbb{Z}^d} \hat{f}_n e^{i\langle x, n \rangle}.$$

2.5.2 On Discrete Domains

Discrete Fourier Transform. On d -dimensional discrete domain of the form

$$n = (n_1, \dots, n_d) \in \mathbb{Y}_d \stackrel{\text{def.}}{=} [\![1, N_1]\!] \times \dots \times [\![1, N_d]\!]$$

(we denote $[\![a, b]\!] \stackrel{\text{def.}}{=} \{i \in \mathbb{Z} ; a \leq i \leq b\}$) of $N = N_1 \dots N_d$ points, with periodic boundary conditions, one defines an orthogonal basis $(\varphi_k)_k$ by the same tensor product formula as (2.12) but using the 1-D discrete Fourier basis (2.8)

$$\forall (k, n) \in \mathbb{Y}_d^2, \quad \varphi_k(n) = \varphi_{k_1}(n_1) \dots \varphi_{k_d}(n_d) = \prod_{\ell=1}^d e^{\frac{2i\pi}{N_\ell} k_\ell n_\ell} = e^{2i\pi \langle k, n \rangle_{\mathbb{Y}_d}} \quad (2.13)$$

where we used the (rescaled) inner product

$$\langle k, n \rangle_{\mathbb{Y}_d} \stackrel{\text{def.}}{=} \sum_{\ell=1}^d \frac{k_\ell n_\ell}{N_\ell}. \quad (2.14)$$

The basis $(\varphi_k)_k$ is indeed orthonormal for this inner product. The Fourier transform gathers inner products in this basis, and (similarly to the 1-D case) the convention is to not normalize them with $(N_\ell)_\ell$, so that

$$\begin{aligned} \forall k \in \mathbb{Y}_d, \quad \hat{f}_k &\stackrel{\text{def.}}{=} \sum_{n \in \mathbb{Y}_d} f_n e^{-i\langle k, n \rangle_{\mathbb{Y}_d}}, \\ \forall n \in \mathbb{Y}_d, \quad f_n &= \frac{1}{N} \sum_{k \in \mathbb{Y}_d} \hat{f}_k e^{i\langle k, n \rangle_{\mathbb{Y}_d}}. \end{aligned}$$

Fast Fourier Transform. We detail the algorithm in dimension $d = 2$ for notation simplicity, but this extends similarly in arbitrary dimension. The general idea is that if a fast algorithm is available to compute ortho-decompositions on two 1-D bases $(\varphi_{k_1}^1)_{k_1=1}^{N_1}, (\varphi_{k_2}^2)_{k_2=1}^{N_2}$, is extended to compute decomposition on the tensor product basis $(\varphi_{k_1}^1 \otimes \varphi_{k_2}^2)_{k_1, k_2}$ by apply successively the algorithm on the “rows” and then “columns” (the order does not matters) of the matrix $(f_n)_{n=(n_1, n_2)} \in \mathbb{R}^{N_1 \times N_2}$. Indeed

$$\forall k = (k_1, k_2), \quad \langle f, \varphi_{k_1}^1 \otimes \varphi_{k_2}^2 \rangle = \sum_{n=(n_1, n_2)} f_n \varphi_{k_1}^1(n_1) \varphi_{k_2}^2(n_2) = \sum_{n_1} \left(\sum_{n_2} f_{n_1, n_2} \varphi_{k_1}^1(n_2) \right) \varphi_{k_1}^1(n_1).$$

Denoting $C(N_1)$ the complexity of the 1-D algorithm on \mathbb{R}^{N_1} , the complexity of the resulting 2-D decomposition is $N_2 C(N_1) + N_1 C(N_2)$, and hence for the FFT, it is $O(N_1 N_2 \log(N_1 N_2)) = O(N \log(N))$ for $N = N_1 N_2$.

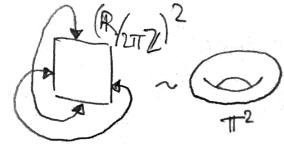


Figure 2.14: The 2-dimensional torus $\mathbb{T}^2 = (\mathbb{R}/2\pi\mathbb{Z})^2$

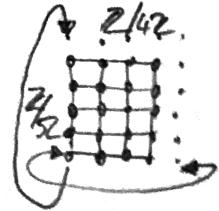


Figure 2.15: Discrete 2-D torus.

If we represent $f \in \mathbb{R}^{N_1 \times N_2}$ as a matrix, and denote $F_N = (e^{-\frac{2i\pi}{N} kn})_{k,n}$ the Fourier transform matrix (or the matrix where rows are the φ_k^*), then one can compute the 2-D Fourier transform as matrix-matrix products

$$\hat{f} = F_{N_1} \times f \times F_{N_2}^* \in \mathbb{R}^{N_1 \times N_2}.$$

But of course these multiplications are not computed explicitly (one uses the FFT).

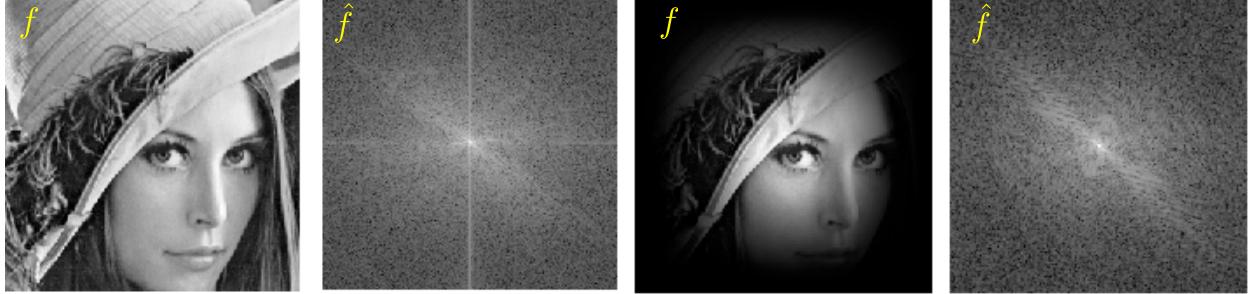


Figure 2.16: 2D Fourier analysis of a image (left), and attenuation of the periodicity artifact using masking (right).

2.5.3 Shannon sampling theorem.

The sampling Theorem 1 extends easily to \mathbb{R}^d by tensorization, assuming that the sampling is on a uniform Cartesian grid. In 2-D for instance, if $\text{supp}(\hat{f}) \subset [-\pi/s_1, \pi/s_1] \times [-\pi/s_2, \pi/s_2]$ and f is decaying fast enough,

$$\forall x \in \mathbb{R}^2, \quad f(x) = \sum_{n \in \mathbb{Z}^2} f(n_1 s_1, n_2 s_2) \text{sinc}(x_1/s_1 - n_1) \text{sinc}(x_2/s_2 - n_2) \quad \text{where} \quad \text{sinc}(u) = \frac{\sin(\pi u)}{\pi u}.$$

2.5.4 Convolution in higher dimension.

Convolution on \mathbb{X}^d with $\mathbb{X} = \mathbb{R}$ or $\mathbb{X} = \mathbb{R}/2\pi\mathbb{Z}$ are defined in the very same way as in 1-D (2.4) as

$$f \star g(x) = \int_{\mathbb{X}^d} f(t)g(x-t)dt.$$

Similarly, finite discrete convolution of vectors $f \in \mathbb{R}^{N_1 \times N_2}$ extend formula (2.11) as

$$\forall n \in \llbracket 0, N_1 - 1 \rrbracket \times \llbracket 0, N_2 - 1 \rrbracket, \quad (f \star g)_n \stackrel{\text{def.}}{=} \sum_{k_1=0}^{N_1-1} \sum_{k_2=0}^{N_2-1} f_{k_1} g_{n-k_1}$$

where additions and subtractions of vectors are performed modulo (N_1, N_2) .

The Fourier-convolution theorem is still valid in all this cases, namely $\mathcal{F}(f \star g) = \hat{f} \odot \hat{g}$. In the finite case, this offers a fast $O(N \log(N))$ method to compute convolutions even if f and g do not have small support.

2.6 Application to ODEs and PDEs

2.6.1 On Continuous Domains

We here give only the intuition without formal proofs.

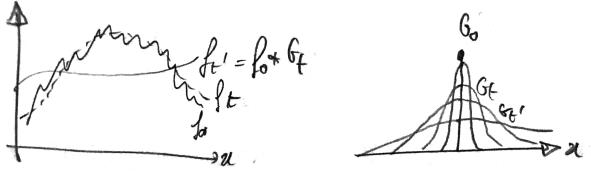


Figure 2.17: Heat diffusion as a convolution.

One $\mathbb{X} = \mathbb{R}$ or \mathbb{T} , one has

$$\mathcal{F}(f^{(k)})(\omega) = (\mathrm{i}\omega)^k \hat{f}(\omega).$$

Intuitively, $f^{(k)} = f \star \delta^{(k)}$ where $\delta^{(k)}$ is a distribution with Fourier transform $\hat{\delta}^{(k)}(\omega) = (\mathrm{i}\omega)^k$. Similarly on $\mathbb{X} = \mathbb{R}^d$ (see Section 2.5 for the definition of the Fourier transform in dimension d), one has

$$\mathcal{F}(\Delta f)(\omega) = -\|\omega\|^2 \hat{f}(\omega) \quad (2.15)$$

(and similarly on \mathbb{T} replacing ω by $n \in \mathbb{Z}^d$). The Fourier transform (or Fourier coefficients) are thus powerful to study linear differential equations with constant coefficients, because they are turned into algebraic equations.

As a typical example, we consider the heat equation

$$\frac{\partial f_t}{\partial t} = \Delta f_t \implies \forall \omega, \quad \frac{\partial \hat{f}_t(\omega)}{\partial t} = -\|\omega\|^2 \hat{f}(\omega).$$

This shows that $\hat{f}_t(\omega) = \hat{f}_0(\omega) e^{-\|\omega\|^2 t}$ and by inverse Fourier transform and the convolution theorem

$$f_t = G_t \star f_0 \quad \text{where} \quad G_t = \frac{1}{(4\pi t)^{d/2}} e^{-\frac{\|x\|^2}{4t}}$$

which is a Gaussian of standard deviation $\sqrt{2t}$.

2.6.2 Finite Domain and Discretization

On $\mathbb{R}/N\mathbb{Z}$ (i.e. discrete domains with periodic boundary conditions), one typically considers forward finite differences (first and second order)

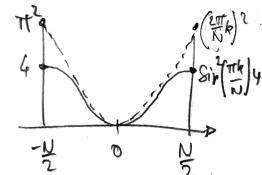
$$D_1 f \stackrel{\text{def.}}{=} N(f_{n+1} - f_n)_n = f \star d_1 \quad \text{where} \quad d_1 = [-1, 0, \dots, 0, 1]^\top \in \mathbb{R}^N, \quad (2.16)$$

$$D_2 f = D_1^\top D_1 f \stackrel{\text{def.}}{=} N^2(f_{n+1} + f_{n-1} - 2f_n)_n = f \star d_2 \quad \text{where} \quad d_2 = d_1 \star \bar{d}_1 = [2, -1, 0, \dots, 0, -1]^\top \in \mathbb{R}^N. \quad (2.17)$$

Thanks to Proposition 5, one can alternatively computes

$$\mathcal{F}(D_2 f) = \hat{d}_2 \odot \hat{f} \quad \text{where} \quad (\hat{d}_2)_k = N^2(e^{\frac{2\mathrm{i}\pi k}{N}} + e^{-\frac{2\mathrm{i}\pi k}{N}} - 2) = -4N^2 \sin\left(\frac{\pi k}{N}\right)^2. \quad (2.18)$$

For $N \gg k$, one thus has $(\hat{d}_2)_k \sim -(2\pi k)^2$ which matches the scaling of (2.15).



2.7 A Bit of Group Theory

The reference for this section is [31].

Figure 2.18: Comparison of the spectrum of Δ and D_2 .

2.7.1 Characters

For $(G, +)$ a commutative group, a character is a group morphism $\chi : (G, +) \rightarrow (\mathbb{C}^*, \cdot)$, i.e. it satisfies

$$\forall (n, m) \in G, \quad \chi(n + m) = \chi(n)\chi(m).$$

The set of characters is the so-called dual (\hat{G}, \odot) and is a group for the pointwise multiplication $(\chi_1 \odot \chi_2)(n) \stackrel{\text{def.}}{=} \chi_1(n)\chi_2(n)$. Indeed, the inverse of a character χ is $\chi^{-1}(n) = \chi(-n)$.

Note that for a finite group G with $|G| = N$, then since $N \times n = 0$ for any $n \in G$, then $\chi(n)^N = \chi(Nn) = \chi(0) = 1$, so that characters assume values in the unit circle, and more precisely

$$\chi(n) \in \left\{ e^{\frac{2i\pi}{N}k} ; 0 \leq k \leq N - 1 \right\}. \quad (2.19)$$

So in particular \hat{G} is a finite group (since there is a finite number of applications between two finite sets) and $\chi^{-1} = \bar{\chi}$. In the case of a cyclic group, the dual is actually simple to describe.

Proposition 6. *For $G = \mathbb{Z}/N\mathbb{Z}$, then $\hat{G} = (\varphi_k)_{k=0}^{N-1}$ where $\varphi_k = (e^{\frac{2i\pi}{N}nk})_n$ and $k \mapsto \varphi_k$ defines a (non-canonical) isomorphism $G \sim \hat{G}$.*

Proof. The φ_k are indeed characters.

Conversely, for any $\chi \in \hat{G}$, according to (2.19), $\chi(1) = e^{\frac{2i\pi}{N}k}$ for some k . Then

$$\chi(n) = \chi(1)^n = e^{\frac{2i\pi}{N}kn} = \varphi_k(n).$$

Note that all these applications are different (because $\varphi_k(1)$ are all distincts) which shows that $|G| = |\hat{G}|$ so that they are isomorphic. \square

This proposition thus shows that characters of cyclic groups are exactly the discrete Fourier orthonormal basis defined in (2.8).

Commutative groups. For more general commutative groups with a finite number of generators, according to the celebrated structure theorem, one can “decompose” them as a product of cyclic groups (which are in some sense the basic building blocks), i.e. there is the following isomorphism of groups

$$G \sim (\mathbb{Z}/N_1\mathbb{Z}) \times \dots \times (\mathbb{Z}/N_d\mathbb{Z}) \times \mathbb{Z}^Q. \quad (2.20)$$

If G is finite, then $Q = 0$ and $N = N_1 \times N_d$. In this case, G is simply a discrete d -dimensional “rectangle” with periodic boundary conditions.

For two finite groups (G_1, G_2) one has

$$\widehat{G_1 \times G_2} = \hat{G}_1 \otimes \hat{G}_2 = \left\{ \chi_1 \otimes \chi_2 ; (\chi_1, \chi_2) \in \hat{G}_1 \times \hat{G}_2 \right\}. \quad (2.21)$$

Here \otimes is the tensor product of two functions

$$\forall (n_1, n_2) \in G_1 \times G_2, \quad (\chi_1 \otimes \chi_2)(n_1, n_2) \stackrel{\text{def.}}{=} \chi_1(n_1)\chi_2(n_2).$$

Indeed, one verifies that $\chi_1 \otimes \chi_2$ is a morphism, and in fact one has the factorization $\chi = \chi(\cdot, 0) \otimes \chi(0, \cdot)$ because one decomposes $(n_1, n_2) = (n_1, 0) + (0, n_2)$.

This construction, thanks to the structure theorem, leads to a constructive proof of the isomorphism theorem.

Proposition 7. *If G is commutative and finite then $\hat{G} \sim G$.*

Proof. The structure theorem (2.20) for $Q = 0$ and the dual of a product (2.21) shows that

$$\hat{G} \sim \hat{G}_1 \otimes \dots \otimes \hat{G}_d$$

where we denoted $G_\ell \stackrel{\text{def.}}{=} \mathbb{Z}/N_\ell\mathbb{Z}$. One then remark that $\hat{G}_1 \otimes \hat{G}_2 \sim G_1 \times \hat{G}_2$. One conclude thanks to Proposition 6, since one has $\hat{G}_k \sim G_k$. \square

Note that the isomorphism $\hat{G} \sim G$ is not “canonical” since it depends on the indexing of the roots of unity on the circle. Similarly to the case of duality of vector space, the isomorphism $\hat{G} \sim G$ can be made canonical by considering the evaluation map

$$g \in G \mapsto e_g \in \hat{G} \quad \text{where} \quad (e_g : \chi \in \hat{G} \mapsto \chi(g) \in \mathbb{C}^*).$$

Discrete Fourier transform from character's point of view. One can be even more constructive by remarking that characters in \hat{G}_ℓ are the discrete Fourier atoms (2.8), i.e. are of the form

$$(e^{\frac{2i\pi}{N_\ell} k_\ell n_\ell})_{n_\ell=0}^{N_\ell-1} \quad \text{for some } 0 \leq k_\ell < N_\ell.$$

Identifying G and $G_1 \times \dots \times G_d$, by tensorizing these functions together, one thus obtains that the characters composing \hat{G} are exactly the orthogonal multi-dimensional discrete Fourier basis (2.13).

2.7.2 More General cases

Infinite groups. For an infinite group with a finite number of generator, one has $Q > 0$, and the definition of \hat{G} should impose the continuity of the characters (and also use an invariant measure on G to define inner products). In the case $G = \mathbb{Z}$, the dual are indexed by a continuous parameter,

$$\hat{\mathbb{Z}} = \{\varphi_\omega : n \mapsto e^{in\omega} \in L^2(\mathbb{R}/2\pi\mathbb{Z}) ; \omega \in \mathbb{R}/2\pi\mathbb{Z}\}$$

so that $\hat{\mathbb{Z}} \sim \mathbb{R}/2\pi\mathbb{Z}$. The case $G = \mathbb{Z}^Q$ follows by tensorization. The $(\varphi_\omega)_\omega$ are “orthogonal” in the sense that $\langle \varphi_\omega, \varphi_{\omega'} \rangle_{\mathbb{Z}} = \delta(\omega - \omega')$ can be understood as a Dirac kernel (this is similar to the Poisson formula), where $\langle u, v \rangle_{\mathbb{Z}} \stackrel{\text{def.}}{=} \sum_n u_n \bar{v}_n$. The “decomposition” of a sequence $(c_n)_{n \in \mathbb{Z}}$ on the set of characters is equivalent to forming a Fourier series $\sum_n c_n e^{-in\omega}$.

Similarly, for $G = \mathbb{R}/2\pi\mathbb{Z}$, one has $\hat{G} = \mathbb{Z}$, with orthonormal characters $\varphi_n = e^{in\omega}$, so that the decomposition of functions in $L^2(G)$ is the computation of Fourier coefficients.

Non-commutative groups. For non-commutative group, one also observe that G is not isometric to \hat{G} . A typical example is the symmetric group Σ_N of N elements, where one can show that $\hat{G} = \{\text{Id}, \varepsilon\}$ where $\varepsilon(\sigma) = (-1)^q$ is the signature, where q is the number of permutations involved in a decomposition of $\sigma \in \Sigma_N$.

In order to study non-commutative groups, one has to replace morphisms $\chi : G \rightarrow \mathbb{C}^*$ by morphisms $\rho : G \rightarrow \text{GL}(\mathbb{C}^d)$ for some d , which are called “representations” of the group G . For $(g, g') \in G$ (denoting now multiplicatively the operation on G), one should thus have $\rho(gg') = \rho(g) \circ \rho(g')$. When $d = 1$, identifying $\text{GL}(\mathbb{C}) \sim \mathbb{C}^*$, one retrieve the definition of characters. Note that if ρ is a representation, then $\chi(g) \stackrel{\text{def.}}{=} \text{tr}(\rho(g))$, where tr is the trace, defines a character. In order to limit the set of such representations, one is only interested in “elementary” ones, which does not have invariant sub-spaces, and are called “irreducible” (otherwise one could create arbitrary large representation by stacking others in a block diagonal way). One can show that the dimensions of these irreducible representations sum to N and that the entries of the matrices involved in these representation define an orthogonal basis of the space of functions $f : G \rightarrow \mathbb{C}$, thus leading to a theory extending the one on commutative groups. In particular, this Fourier transform in some sense also diagonalizes convolutions over G

$$f \star h(a) \stackrel{\text{def.}}{=} \sum_{bc=a} f(b)h(c).$$

For the symmetric group, there is an explicit description of the set of irreducible representations.

2.8 A Bit of Spectral Theory

In order to define Fourier methods on general domains \mathbb{X} , one can use the aforementioned group-theoretic approach if $\mathbb{X} = G$ is a group, or also if a group acts transitively on \mathbb{X} . An alternative way is to describe the equivalent of Fourier basis functions as diagonalizing a specific differential operator (as we have seen in Section 2.6 that it is in some sense a way to characterise the Fourier basis). Of particular interest is the Laplacian, since it is the lowest order rotation-invariant differential operator, and that there exists natural generalization on domains such as surfaces or graphs.

2.8.1 On a Surface or a Manifold

The presentation here is very informal. One can define the Laplacian of a smooth function $f : \mathbb{X} \rightarrow \mathbb{C}$ defined on a “surface” \mathbb{X} as

$$\forall x \in \mathbb{X}, \quad (\Delta f)(x) \stackrel{\text{def.}}{=} \lim_{\varepsilon \rightarrow 0} \frac{1}{\text{Vol}(B_\varepsilon(x))} \int_{B_\varepsilon(x)} f(x) d\mu(x) - f(x).$$

Here $\mu(x)$ is the area measure on \mathbb{X} , $\text{Vol}(B) \stackrel{\text{def.}}{=} \int_B d\mu(x)$, and $B_\varepsilon(x) = \{y ; d_{\mathbb{X}}(x, y) \leq \varepsilon\}$ is the geodesic ball of radius ε at x , where $d_{\mathbb{X}}$ is the geodesic distance on \mathbb{X} (length of the shortest path).

If the surface \mathbb{X} is smooth, compact and connected, then it is possible to show that Δ is itself a compact operator with a negative spectrum $0 > \lambda_1 > \lambda_2 > \dots$ and an orthogonal set of eigenvectors $(\varphi_n)_{n \geq 0}$ where $\varphi_1 = 1$. Here the inner product is $\langle f, g \rangle_{\mathbb{X}} \stackrel{\text{def.}}{=} \int_{\mathbb{X}} f(x) g(x) d\mu(x)$ on $L^2(\mathbb{X})$. In the case of a flat torus $\mathbb{X} = (\mathbb{R}/\mathbb{Z})^d$, then writing $x = (x_1, \dots, x_d)$,

$$\Delta f = \sum_{s=1}^d \frac{\partial^2 f}{\partial^2 x_s}.$$

Similarly to (2.15) (which was for an unbounded domain), then one can chose for this eigen-functions φ_n the Fourier basis (2.1) and $\lambda_n = -\|n\|^2$

2.8.2 Spherical Harmonics

Of particular interest is the special case of the previous construction on the $(d-1)$ -dimensional sphere $\mathbf{S}^{d-1} = \{x \in \mathbb{R}^d ; \|x\|_{\mathbb{R}^d} = 1\}$. In this case, there exists a closed form expression for the eigenvectors of the Laplacian. In the 3-D case $d = 3$, they are indexed by $n = (\ell, m)$

$$\forall \ell \in \mathbb{N}, \quad \forall m = -\ell, \dots, \ell, \quad \varphi_{\ell,m}(\theta, \varphi) = e^{im\varphi} P_{\ell}^m(\cos(\theta))$$

and then the eigenvalue of the Laplacian is $\lambda_{\ell,m} = -\ell(\ell + 1)$. Here P_{ℓ}^m are associated Legendre polynomials, and we used spherical coordinates $x = (\cos(\varphi), \sin(\varphi) \sin(\theta), \sin(\varphi) \cos(\theta)) \in \mathbf{S}^3$ for $(\theta, \varphi) \in [0, \pi] \times [0, 2\pi]$. The index ℓ is analogous to the amplitude of Fourier frequencies in 2-D. For a fixed ℓ , the space $V_{\ell} = \text{span}(\varphi_{\ell,m})$ is an eigenspace of Δ , and is also invariant under rotation.

2.8.3 On a Graph

We assume \mathbb{X} is a graph of N vertices, simply indexed $\{1, \dots, N\}$. Its “geometry” is described by a connectivity matrix of weights $W = (w_{i,j})_{i \sim j}$ where we denote $i \sim j$ to indicate that (i, j) is an edge of the graph for $(i, j) \in \mathbb{X}^2$. We assume that this weight matrix and the connectity is symmetric, $w_{i,j} = w_{j,i}$.



Figure 2.19: Computing Laplacian on a surface



Figure 2.20: Spherical coordinates.

The graph Laplacian $\Delta : \mathbb{R}^N \rightarrow \mathbb{R}^N$ is computing the difference between the average of values around a point and the value at this point

$$\forall f \in \mathbb{R}^N, \quad (\Delta f)_i \stackrel{\text{def.}}{=} \sum_{j \sim i} w_{i,j} f_j - \left(\sum_{j \sim i} w_{i,j} \right) f_i \quad \Rightarrow \quad \Delta = W - D$$

where $D \stackrel{\text{def.}}{=} \text{diag}_i(\sum_{j \sim i} w_{i,j})$. In particular, note $\Delta \mathbf{1} = 0$

For instance, if $\mathbb{X} = \mathbb{Z}/N\mathbb{Z}$ with the graph $i \sim i-1$ and $i \sim i+1$ (modulo N), then Δ is the finite difference Laplacian operator $\Delta = D_2$ defined in (2.17). This extends to any dimension by tensorization.

Proposition 8. Denoting $G : f \in \mathbb{R}^N \mapsto (\sqrt{w_{i,j}}(f_i - f_j))_{i < j}$ the graph-gradient operator, one verifies that

$$-\Delta = G^\top G \quad \Rightarrow \quad \forall f \in \mathbb{R}^N, \quad \langle \Delta f, f \rangle_{\mathbb{R}^N} = -\langle Gf, Gf \rangle_{\mathbb{R}^P}.$$

where P is the number of (ordered) edges $E = \{(i, j) ; i \sim j, i < j\}$.

Proof. One has

$$\begin{aligned} \|Gf\|^2 &= \sum_{(i,j) \in E} w_{i,j} |f_i - f_j|^2 = \sum_{i < j} w_{i,j} f_i^2 + \sum_{i < j} w_{i,j} f_j^2 - 2 \sum_{i < j} w_{i,j} f_i f_j \\ &= \sum_{i < j} w_{i,j} f_i^2 + \sum_{i > j} w_{i,j} f_i^2 - \sum_{i,j} w_{i,j} f_i f_j = \sum_j f_i^2 \sum_{i,j} w_{i,j} - \sum_i f_i \sum_j w_{i,j} f_j \\ &= \langle Df, f \rangle - \langle Lf, f \rangle = -\langle Lf, f \rangle. \end{aligned}$$

□

This proposition shows that Δ is a negative semi-definite operator, which thus diagonalizes in an ortho-basis $(\varphi_n)_{n=1}^N$, with $\varphi_1 = 1$, with eigenvalues $0 \geq \lambda_1 \geq \lambda_N$. If \mathbb{X} is connected, one can show that $\lambda_1 < 0$. In the case of a regular graph associated to a uniform grid, one retrieves the discrete Fourier basis (2.8).

More details and application of Laplacians on graphs can be found in Chapter 3, see in particular Section 3.2.5.

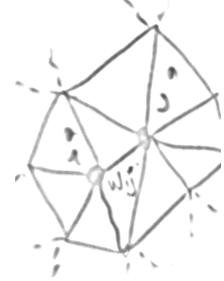


Figure 2.21:
Weighted graph.

Chapter 3

Linear Mesh Processing

This chapter exposes the basics of surface approximation with 3D meshes and the way to process such meshes with linear operators. In particular, it studies filtering on 3D meshes and explains how a Fourier theory can be built to analyze these filters.

3.1 Surface Discretization with Triangulated Mesh

3.1.1 Continuous Geometry of Surfaces

In this course, in order to simplify the mathematical description of surfaces, we consider only globally parameterized surfaces. We begin by considering surfaces embedded in euclidean space $\mathcal{M} \subset \mathbb{R}^k$.

Definition 1 (Parameterized surface). *A parameterized surface is a mapping*

$$u \in \mathcal{D} \subset \mathbb{R}^2 \mapsto \varphi(u) \in \mathcal{M}.$$

Of course, most surfaces do not benefit from such a simple parameterization. For instance, a sphere should be split into two parts in order to be mapped on two disks $\mathcal{D}_1, \mathcal{D}_2$. These topological difficulties require the machinery of manifolds in order to incorporate a set of charts $\mathcal{D} = \{\mathcal{D}_i\}_i$ that overlap in a smooth manner. All the explanations of this course extend seamlessly to this multi-charts setting.

A curve is defined in parameter domain as a 1D mapping $t \in [0, 1] \mapsto \gamma(t) \in \mathcal{D}$. This curve can be traced over the surface and its geometric realization is $\bar{\gamma}(t) \stackrel{\text{def.}}{=} \varphi(\gamma(t)) \in \mathcal{M}$. The computation of the length of γ in ambient k -dimensional space \mathbb{R}^k follows the usual definition, but to do the computation over the parametric domain, one needs to use a local metric defined as follow.

Definition 2 (First fundamental form). *For an embedded manifold $\mathcal{M} \subset \mathbb{R}^k$, the first fundamental form is*

$$I_\varphi = \left(\left\langle \frac{\partial \varphi}{\partial u_i}, \frac{\partial \varphi}{\partial u_j} \right\rangle \right)_{i,j=1,2}.$$

This local metric I_φ defines at each point the infinitesimal length of a curve as

$$L(\gamma) \stackrel{\text{def.}}{=} \int_0^1 \|\bar{\gamma}'(t)\| dt = \int_0^1 \sqrt{\gamma'(t)^T I_\varphi(\gamma(t)) \gamma'(t)} dt.$$

This fundamental form is an intrinsic invariant that does not depends on how the surfaces is isometrically embedded in space (since the length depends only on this tensor field I_φ). In contrast, higher order differential quantities such as curvature might depend on the bending of the surface and are thus usually not intrinsic (with the notable exception of invariants such as the gaussian curvature). In this course, we restrict ourselves to first order quantities since we are mostly interested in lengths and the intrinsic study of surfaces.

Example 1 (Isometry and conformality). A surface \mathcal{M} is locally isometric to the plane if $I_\varphi = \text{Id}_2$. This is for instance the case for a cylinder. The mapping φ is said to be conformal if $I_\varphi(u) = \lambda(u)\text{Id}_2$. It means that the length of a curve over the plane is only locally scaled when mapped to the surface. In particular, the angle of two intersecting curves is the same over the parametric domain and over the surface. This is for instance the case for the stereographic mapping between the plane and a sphere.

3.1.2 Discretization of Surfaces with Triangulations

Mesh Data Structure A triangulated mesh is a discrete structure that can be used to approximate a surface embedded in Euclidean space \mathbb{R}^k . It is composed of a topological part $M = (V, E, F)$ and a geometrical realization $\mathcal{M} = (\mathcal{V}, \mathcal{E}, \mathcal{F})$. It is important to make the distinction between these two parts since many algorithms rely only on geometry (point clouds processing such as dimension reduction) or on topology (such as compression).

The topology M of the mesh is composed of

- *Vertices* (0D): this is an abstract set of indices $V \simeq \{1, \dots, n\}$.
- *Edges* (1D): this is a set of pair of vertices $E \subset V \times V$. This set is assumed to be symmetric

$$(i, j) \in E \iff i \sim j \Leftrightarrow (j, i) \in E.$$

- *Faces* (2D): this is a collection of 3-tuples of vertices $F \subset V \times V \times V$, with the additional compatibility condition

$$(i, j, k) \in F \implies (i, j), (j, k), (k, i) \in E.$$

We further assume that there is no isolated edges

$$\forall (i, j) \in E, \exists k, (i, j, k) \in F.$$

The set of edges can be stored in a symmetric matrix $A \in \mathbb{R}^{n \times n}$ such that $A_{ij} = 1$ if $(i, j) \in E$ and $A_{ij} = 0$ otherwise. This matrix is often stored as a sparse matrix since the number of edges is usually much smaller than n^2 . The set of vertices and edges form a non-oriented graph $\mathcal{G} = (V, E)$. Faces are often stored as a matrix $A_F \in \{1, \dots, n\}^{3 \times m}$ where m is the number of faces and a column $((A_F)_{i,1}, (A_F)_{i,2}, (A_F)_{i,3})$ stores the indices of a face. In a triangulation, the face matrix A_F allows to recover the edge incidence matrix A . The face data structure allows to really capture the 2D geometry of surfaces, which is not possible with graphs alone.

The geometric realization \mathcal{M} is defined through a spatial localization of the vertices (for instance in 3D space)

$$\mathcal{V} \stackrel{\text{def.}}{=} \{x_i ; i \in V\} \subset \mathbb{R}^3.$$

This allows to define a piecewise linear mesh

$$\mathcal{F} \stackrel{\text{def.}}{=} \bigcup_{(i,j,k) \in F} \text{Conv}(x_i, x_j, x_k) \subset \mathbb{R}^3,$$

where the convex envelop $\text{Conv}(x, y, z)$ of three points is the Euclidean triangle generated by (x, y, z) .

This piecewise linear realization \mathcal{M} can be displayed as a 3D surface on a computer screen. This is performed through a perspective projection of the points and a linear interpolation of color and light inside the triangle. Figure 3.1 shows an example of 3D display, with a zoom on the faces of the mesh.

Adjacency Relationships From the basic topological information given by $M = (V, E, F)$, one can deduce several adjacency data-structures that are important to navigate over the triangulation.

Definition 3 (Vertex 1-ring). *The vertex 1-ring of a vertex $i \in V$ is*

$$V_i \stackrel{\text{def.}}{=} \{j \in V ; (i, j) \in E\} \subset V. \quad (3.1)$$

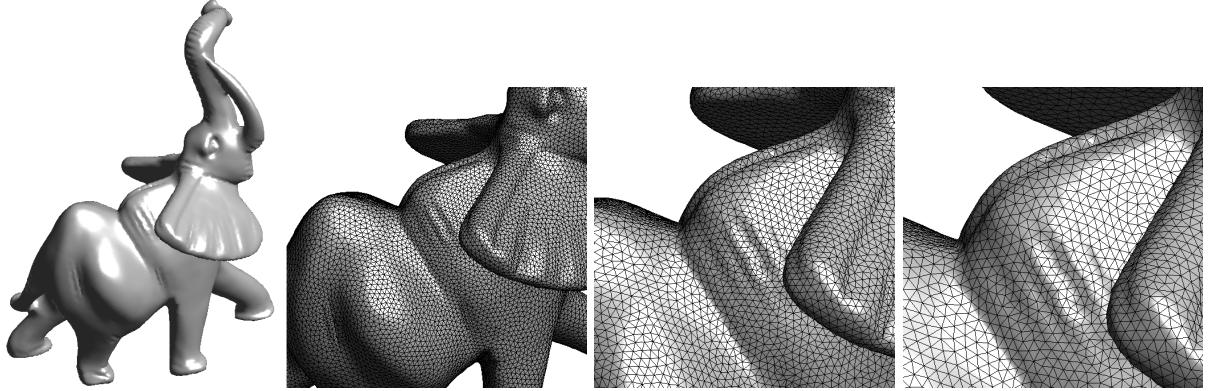


Figure 3.1: Example of display of a 3D mesh.

The s -ring is defined by induction as

$$\forall s > 1, \quad V_i^{(s)} = \left\{ j \in V ; (k, j) \in E \text{ and } k \in V_i^{(s-1)} \right\}. \quad (3.2)$$

Definition 4 (Face 1-ring). The face 1-ring of a vertex $i \in V$ is

$$F_i \stackrel{\text{def.}}{=} \{(i, j, k) \in F ; i, j \in V\} \subset F.$$

The geometrical realization of a vertex 1-ring is

$$\mathcal{V}_i = \bigcup_{(i, j, k) \in V_i} \text{Conv}(x_i, x_j, x_k).$$

A triangulated mesh is a manifold mesh if all the rings \mathcal{V}_i for $i \in V$ are homeomorphic to either a disk (for interior vertices) or to a half disk (for boundary vertices). This ensures that the geometrical mesh really has the topology of a 2D surface embedded in \mathbb{R}^3 (possibly with boundaries). In particular, it implies that there is at most two faces connected to each edge

$$\forall (i, j) \in E, \quad \#\{k ; (i, j, k) \in F\} \leq 2.$$

As an application of these local rings, one can compute a normal at each point using a simple rule

$$\forall f = (i, j, k) \in F, \quad \vec{n}_f \stackrel{\text{def.}}{=} \frac{(x_j - x_i) \wedge (x_k - x_i)}{\|(x_j - x_i) \wedge (x_k - x_i)\|}.$$

and where

$$\forall i \in V, \quad \vec{n}_i \stackrel{\text{def.}}{=} \frac{\sum_{f \in F_i} \vec{n}_f}{\|\sum_{f \in F_i} \vec{n}_f\|}.$$

These normals are used to define for instance a light intensity $I(i) = \max(\langle n_i, \ell(i), \rangle 0)$, where $\ell(i)$ is the incident light. In practice one uses a infinite light source $\ell(i) = \ell = \text{constant}$ or a local spot located at position $s \in \mathbb{R}^3$ through $\ell(i) = (v_i - s)/\|v_i - s\|$. This light intensity is interpolated on the whole mesh during display.

3.2 Linear Mesh Processing

The light intensity I is a particular example of a function defined at each vertex of the mesh. Mesh processing is intended to process such functions and we thus define carefully vector spaces and operators on meshes.

3.2.1 Functions on a Mesh

In this course, a function is a discrete set of values defined at each vertex location.

Definition 5 (Linear space on a mesh). *A function on a mesh is a mapping $f \in \ell^2(\mathcal{V}) \simeq \ell^2(V) \simeq \mathbb{R}^n$ and can be viewed equivalently as*

$$f : \begin{cases} \mathcal{V} & \rightarrow \mathbb{R} \\ x_i & \mapsto f(x_i) \end{cases} \iff f : \begin{cases} V & \rightarrow \mathbb{R} \\ i & \mapsto f_i \end{cases} \iff f = (f_i)_{i \in V} \in \mathbb{R}^n.$$

The linear space of the functions on a mesh is equipped with an Hilbert space structure that allows to quantify approximation error and compute projections of functions.

Definition 6 (Inner product and norm). *One defines the following inner product and norm for vector $f, g \in \mathbb{R}^n$*

$$\langle f, g \rangle \stackrel{\text{def.}}{=} \sum_{i \in V} f_i g_i \quad \text{and} \quad \|f\|^2 = \langle f, f \rangle.$$

In order to modify (process) functions on a mesh (such as a light intensity I), this course considers only linear operations that are defined through a large matrix.

Definition 7 (Linear operator A). *A linear operator A is defined as*

$$A : \ell^2(V) \rightarrow \ell^2(V) \iff A = (a_{ij})_{i,j \in V} \in \mathbb{R}^{n \times n} \text{ (matrix).}$$

and operate on a function f as follow

$$(Af)(x_i) = \sum_{j \in V} a_{ij} f(x_j) \iff (Af)_i = \sum_{j \in V} a_{ij} f_j.$$

Example 2. If the coordinates of the point of a mesh are written $x_i = (x_i^1, x_i^2, x_i^3) \in \mathbb{R}^3$, then the X -coordinate defines a function $f : i \in V \mapsto x_i^1 \in \mathbb{R}$. A geometric mesh \mathcal{M} is thus 3 functions defined on M .

Mesh processing is the task of modifying functions $f \in \ell^2(V)$. For instance, one can denoise a mesh \mathcal{M} as 3 functions on M . The usual strategy applies a linear operator $f \mapsto Af$. Sometimes, A can be computed from M only (for instance for compression) but most of the times it requires both M and \mathcal{M} .

3.2.2 Local Operators

In most applications, one can not store and manipulate a full matrix $A \in \mathbb{R}^{n \times n}$. Furthermore, one is usually interested in exploiting the local redundancies that exist in most usual functions $f \in \mathbb{R}^n$ defined on a mesh. This is why we restrict our attention to local operators that can be conveniently stored as sparse matrices (the zeros are not kept in memory).

Definition 8 (Local operator). *A local operator $W \in \mathbb{R}^{n \times n}$ satisfies $w_{ij} = 0$ if $(i, j) \notin E$.*

$$(Wf)_i = \sum_{(i,j) \in E} w_{ij} f_j.$$

A particularly important class of local operators are local smoothings (also called filterings) that perform a local weighted sum around each vertex of the mesh. For this averaging to be consistent, we define a normalized operator \tilde{W} whose set of weights sum to one.

Definition 9 (Local averaging operator). *A local normalized averaging is $\tilde{W} = (\tilde{w}_{ij})_{i,j \in V} \geq 0$ where*

$$\forall (i, j) \in E, \quad \tilde{w}_{ij} = \frac{w_{ij}}{\sum_{(i,j) \in E} w_{ij}}.$$

It can be equivalently expressed in matrix form as

$$\tilde{W} = D^{-1}W \quad \text{with} \quad D = \text{diag}_i(d_i) \quad \text{where} \quad d_i = \sum_{(i,j) \in E} w_{ij}.$$

The smoothing property corresponds to $\tilde{W}1 = 1$ which means that the unit vector is an eigenvector of W with eigenvalue 1.

Example 3. In practice, we use three popular kinds of averaging operators.

- *Combinatorial weights:* they depends only on the topology (V, E) of the vertex graph

$$\forall (i, j) \in E, \quad w_{ij} = 1.$$

- *Distance weights:* they depends both on the geometry and the topology of the mesh, but do not require faces information,

$$\forall (i, j) \in E, \quad w_{ij} = \frac{1}{\|x_j - x_i\|^2}.$$

- *Conformal weights:* they depends on the full geometrical realization of the 3D mesh since they require the face information

$$\forall (i, j) \in E, \quad w_{ij} = \cot(\alpha_{ij}) + \cot(\beta_{ij}). \quad (3.3)$$

Figure 3.2 shows the geometrical meaning of the angles α_{ij} and β_{ij}

$$\alpha_{ij} = \angle(x_i, x_j, x_{k_1}) \quad \text{and} \quad \beta_{ij} = \angle(x_i, x_j, x_{k_2}),$$

where $(i, j, k_1) \in F$ and $(i, j, k_2) \in F$ are the two faces adjacent to edge $(i, j) \in E$. We will see in the next section the explanation of these celebrated cotangent weights.

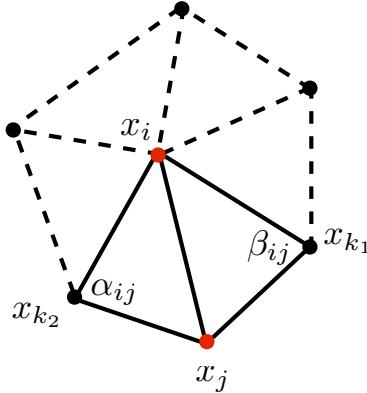


Figure 3.2: One ring around a vertex i , together with the geometrical angles α_{ij} and β_{ij} used to compute the conformal weights.

One can use iteratively a smoothing in order to further filter a function on a mesh. The resulting vectors $\tilde{W}f, \tilde{W}^2, \dots, \tilde{W}^k f$ are increasingly smoothed version of f . Figure 3.3 shows an example of such iterations applied to the three coordinates of mesh. The sharp features of the mesh tend to disappear during iterations. We will make this statement more precise in the following, by studying the convergence of these iterations.

3.2.3 Approximating Integrals on a Mesh

Before investigating algebraically the properties of smoothing operators, one should be careful about what are these discrete operators really approximating. In order for the derivation to be simple, we make computation for a planar triangulation M of a mesh $\mathcal{M} \subset \mathbb{R}^2$.

In the continuous domain, filtering is defined through integration of functions over the mesh. In order to discretize integrals, one needs to define a partition of the plane into small cells centered around a vertex or an edge.

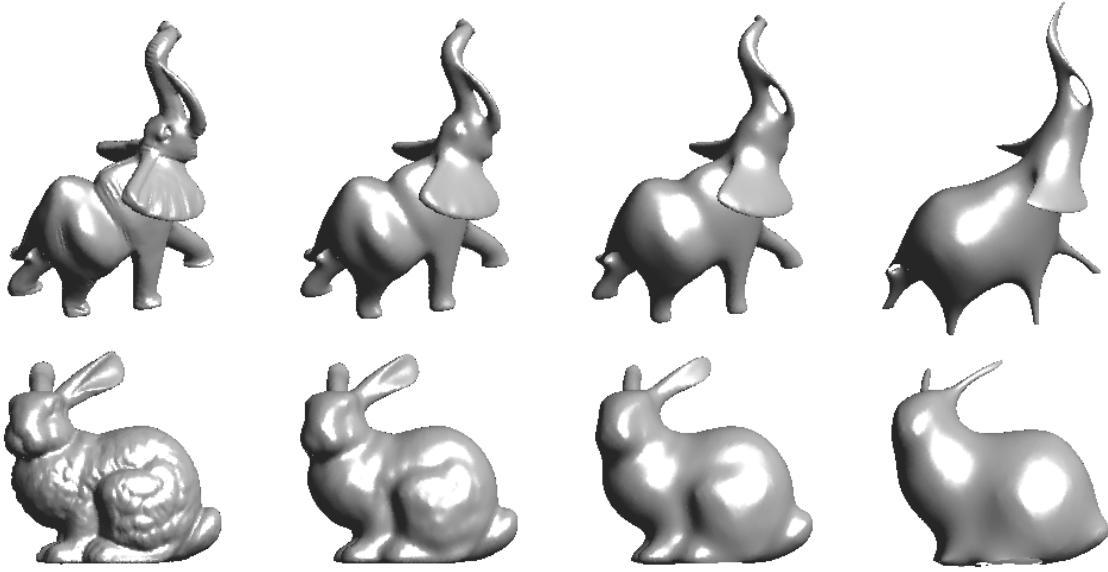


Figure 3.3: Examples of iterative smoothing of a 3D mesh.

Definition 10 (Vertices Voronoi). *The Voronoi diagram associated to the vertices is*

$$\forall i \in V, \quad E_i = \{x \in \mathcal{M} ; \forall j \neq i, \|x - x_i\| \leq \|x - x_j\|\}$$

Definition 11 (Edges Voronoi). *The Voronoi diagram associated to the edges is*

$$\forall e = (i, j) \in E, \quad E_e = \{x \in \mathcal{M} ; \forall e' \neq e, d(x, e) \leq d(x, e')\}$$

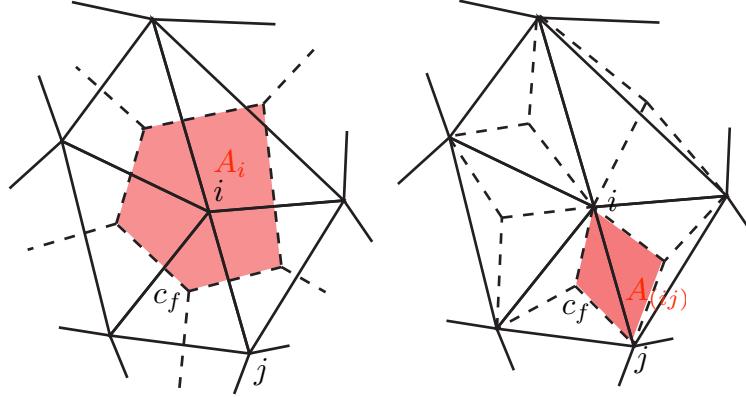


Figure 3.4: Left: vertex Voronoi cell, right: delaunay Voronoi cell. The point c_f is the orthocenter of a face $f = (i, j, k)$.

These Voronoi cells indeed form a partition of the mesh

$$\mathcal{M} = \bigcup_{i \in V} E_i = \bigcup_{e \in E} E_e.$$

The following theorem gives the formula for the area of these cells.

Theorem 3 (Voronoi area formulas). *For all $e = (i, j) \in E$, $\forall i \in V$, one has*

$$A_e = \text{Area}(E_e) = \frac{1}{2} \|x_i - x_j\|^2 (\cot(\alpha_{ij}) + \cot(\beta_{ij}))$$

$$A_i = \text{Area}(E_i) = \frac{1}{2} \sum_{j \in N_i} A_{(ij)}.$$

With these areas, one can approximate integrals on vertices and edges using

$$\int_M f(x) dx \approx \sum_{i \in V} A_i f(x_i) \approx \sum_{e=(i,j) \in E} A_e f([x_i, x_j]).$$

Of particular interest is the approximation of the so-called Dirichlet energy $\int_M \|\nabla_x f\|^2 dx$. In order to compute it on a triangular mesh, one can use a finite difference approximation of the gradient of a function at the point $x_{ij} = (x_i + x_j)/2$ along an edge (i, j)

$$\langle \nabla_{x_{ij}} f, \frac{x_i - x_j}{\|x_i - x_j\|} \rangle \approx \frac{f(x_i) - f(x_j)}{\|x_i - x_j\|}.$$

This leads to the following approximation of the Dirichlet energy

$$\int_M \|\nabla_x f\|^2 dx \approx \sum_{(i,j) \in E} A_{(i,j)} \langle \nabla_{x_{ij}} f, \frac{x_i - x_j}{\|x_i - x_j\|} \rangle^2 \approx \sum_{(i,j) \in E} A_{(i,j)} \frac{|f(x_j) - f(x_i)|^2}{\|x_j - x_i\|^2} \quad (3.4)$$

$$= \sum_{(i,j) \in E} w_{ij} |f(x_j) - f(x_i)|^2 \quad \text{where} \quad w_{ij} = \cot(\alpha_{ij}) + \cot(\beta_{ij}). \quad (3.5)$$

This discrete formulation shows that the correct weights to approximate the Dirichlet energy are the cotangent one, already introduced in equation (3.3).

3.2.4 Example on a Regular Grid

A regular grid is an uniform discretization with n points of $[0, 1]$ (in 1D) or $[0, 1]^2$ (in 2D). One usually assumes periodic boundary conditions, which means that each side of the square is associated with its opposite.

Since the geometry of a regular grid is invariant under translation, local averaging operators can be computed as convolution on $D = (\mathbb{Z}/p\mathbb{Z})^d$ where $n = p^d$ for d the dimension of the domain ($d = 1$ or $d = 2$)

$$\forall i \in D, \quad \tilde{W}f(i) = \sum_{k \in D} f(k) \tilde{w}(i - k),$$

where the operation $+$ and $-$ should be computed modulo p and $\tilde{w}(k) = \tilde{W}(0, k)$ is the convolution kernel.

Example 4 (Averaging). The uniform averaging filter is defined as

$$\tilde{W}f(i) = \frac{1}{|N|} \sum_{k \in N} f(i + k),$$

where N is the set of neighbors of the point 0 and $|N| = 2^d$. In this case, in dimension 1, $\tilde{w} = (1, 0, 1)/2$, where this notation assumes that \tilde{w} is centered at the point 0 .

In order to study translation invariant operators like local filtering, one needs to use the discrete Fourier transform that diagonalizes these operators.

Definition 12 (Discrete Fourier transform). *The 1D discrete Fourier transform $\Phi(f) \in \mathbb{C}^n$ of the vector $f \in \mathbb{C}^n$*

$$\Phi(f)(\omega) = \hat{f}(\omega) \stackrel{\text{def.}}{=} \frac{1}{n} \sum_k f_k e^{\frac{2i\pi}{n} k\omega}.$$

A similar definition can be given for the 2D discrete Fourier transform. The main property of the Fourier transform is the following diagonalization result.

Theorem 4 (Convolution and Fourier). *For any vector f , one has*

$$\Phi(\tilde{W}^k f) = \Phi(\tilde{w} * \dots * \tilde{w} * f) \implies \Phi(\tilde{W}^k f)(\omega) = \widehat{\tilde{w}}(\omega)^k \hat{f}(\omega).$$

The main interest of this tools is that $\Phi(f)$ can be computed in $O(n \log(n))$ operations with the FFT algorithm. Using the following theorem, it gives an alternative expression of a local filtering. This expression in the Fourier domain can be used to speed up the computation of $\tilde{w} * f$ if \tilde{w} has a lot of non zero entries (which is not the case in our setting of local operators). It is also useful to analyze theoretically the behavior of iterated filterings.

Theorem 5 (Convergence). *For any function f defined on a regular grid in 1D or 2D, one has*

$$\tilde{W}^k f \xrightarrow{k \rightarrow +\infty} \frac{1}{|V|} \sum_{i \in V} f_i$$

This Fourier theory can only be developed for domains that have a group structure that enables translation invariant filtering. In particular, it does not carry over easily to an arbitrary surface. In the remaining, we define a corresponding theory for graphs and triangulated surfaces using the eigenvector of Laplacian operators. This Fourier transform on meshes enables the analysis of the convergence of many filtering schemes.

3.2.5 Gradients and Laplacians on Meshes

Gradient operator A gradient operator defines directional derivatives on a triangulation. It maps functions defined on vertices to functions defined on the set of oriented edges

$$\bar{E} \stackrel{\text{def.}}{=} \{(i, j) \in E ; i > j\}.$$

Definition 13 (Gradient). *Given a local averaging W , the gradient operator G is defined as*

$$\forall (i, j) \in \bar{E}, i < j, \quad (Gf)_{(i,j)} \stackrel{\text{def.}}{=} \sqrt{w_{ij}}(f_j - f_i) \in \mathbb{R}.$$

This mapping can be viewed equivalently as

$$G : \ell^2(V) \longrightarrow \ell^2(E), \quad \text{or} \quad G : \mathbb{R}^n \longrightarrow \mathbb{R}^p \quad \text{where } p = |E|, \\ \text{or} \quad G \in \mathbb{R}^{n \times p} \quad (\text{a matrix}).$$

The value of $(Gf)_e$ for an edge $e = (i, j)$ can be thought as a derivative along direction $\overrightarrow{x_i x_j}$.

Example 5. For the local averaging based on square distances, one has

$$w_{ij} = \|x_i - x_j\|^{-2}, \quad (Gf)_{(i,j)} = \frac{f(x_j) - f(x_i)}{\|x_i - x_j\|}.$$

which is exactly the finite difference discretization of a directional derivative.

One a regular grid, one can note that

- Gf discretizes $\nabla f = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y} \right)^T$.
- $G^T v$ discretizes $\div(v) = \frac{\partial v_1}{\partial x} + \frac{\partial v_2}{\partial y}$.

Laplacian Operator A Laplacian operator is a discrete version of a second order derivative operator.

Definition 14 (Laplacian). *Given a local averaging W , the Laplacian operator D is defined as*

$$L \stackrel{\text{def}}{=} D - W, \quad \text{where} \quad D = \text{diag}_i(d_i), \quad \text{with} \quad d_i = \sum_j w_{ij}.$$

In the remaining, we also make use of normalized operators, which have an unit diagonal.

Definition 15 (Normalized Laplacian). *The normalized Laplacian is defined as*

$$\tilde{L} \stackrel{\text{def}}{=} D^{-1/2} L D^{-1/2} = \text{Id}_n - D^{-1/2} W D^{1/2} = \text{Id}_n - D^{1/2} \tilde{W} D^{-1/2}.$$

This normalized Laplacian correspond to the weighted graph Laplacian used in graph theory, see for instance [12].

Remark 1. One can note that

- Laplacians are symmetric operators $L, \tilde{L} \in \mathbb{R}^{n \times n}$.
- L acts like a (second order) derivative since $L\mathbf{1} = 0$.
- in contrast, the normalized Laplacian is not a real derivative since $\tilde{L}\mathbf{1} \neq 0$ in general.

The main interest of the gradient operator is that it factorizes the Laplacian as follow.

Theorem 6 (Laplacian factorization). *One has*

$$L = G^T G \quad \text{and} \quad \tilde{L} = (GD^{-1/2})^T (GD^{-1/2}).$$

This theorem proves in particular that L and \tilde{L} are symmetric positive definite operators. The inner product defined by the Laplacian can be expressed as an energy summed over all the edges of the mesh

$$\langle Lf, f \rangle = \|Gf\|^2 = \sum_{(i,j) \in E} w_{ij} \|f_i - f_j\|^2.$$

In the particular case of the cotangent weights introduced in equation (3.3), one can see that the Laplacian norm $\langle Lf, f \rangle$ is exactly the finite differences approximation of the continuous Dirichlet energy $\int_M |\nabla_x f|^2 dx$ derived in equation (3.5). This is why these cotangent weights are the best choice to compute a Laplacian that truly approximates the continuous Laplace Beltrami operator (see definition 16).

A similar expression is derived for the normalized laplacian

$$\langle \tilde{L}f, f \rangle = \|GD^{-1/2}f\|^2 = \sum_{(i,j) \in E} w_{ij} \left\| \frac{f_i}{\sqrt{d_i}} - \frac{f_j}{\sqrt{d_j}} \right\|^2.$$

Of particular interest for the study of filtering on meshes is the behavior of the spectrum of the Laplacian. We can first study its kernel.

Theorem 7 (Kernel of the Laplacian). *If M is connected, then*

$$\ker(L) = \text{span}(1) \quad \text{and} \quad \ker(\tilde{L}) = \text{span}(D^{1/2}).$$

3.2.6 Examples in 1D and 2D

In 1D, all local weights are equivalent since the points are equi-spaced. The corresponding Laplacian is a convolution that can be written as

$$(Lf)_i = \frac{1}{h^2} (2f_i - f_{i+1} - f_{i-1}) = \frac{1}{h^2} f * (-1, 2, 1),$$

where it is important to remember that the notation $(-1, 2, 1)$ means that the vector is centered around 0.

This discrete 1D Laplacian is the finite difference approximation of the continuous Laplacian on the torus \mathcal{T} of the segment $[0, 1]$ modulo 1. Up to a minus sign, this Laplacian is just the second order derivative

$$L \xrightarrow{h \rightarrow 0} -\frac{d^2 f}{dx^2}(x_i)$$

One should be careful with our notation that consider positive semi-definite Laplacian, that have the opposite sign with respect to second order derivative operators (which are definite negative).

The gradient operator corresponds to a discretization of the first order derivative $f \mapsto f'$ (which is anti-symmetric). The continuous counterpart of the factorization $L = G^T G$ is the integration by part formula on the torus

$$\int_{\mathcal{T}} f''(x)g(x)dx = - \int_{\mathcal{T}} -f(x)g'(x)dx \implies \int_{\mathcal{T}} f''(x)f(x)dx = - \int_{\mathcal{T}} |f'(x)|^2 \leq 0.$$

The discrete Laplacian on a 2D grid can also be written as a 2D convolution

$$(Lf)_i = \frac{1}{h^2} (4f_i - f_{j_1} - f_{j_2} - f_{j_3} - f_{j_4}) = \frac{1}{h^2} f * \begin{pmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{pmatrix}$$

where $\{j_k\}_k$ are the four neighbors of the point i . This operator is the finite difference approximation to the continuous 2D Laplacian

$$L \xrightarrow{h \rightarrow 0} -\frac{\partial^2 f}{\partial x^2}(x_i) - \frac{\partial^2 f}{\partial y^2}(x_i) = -\Delta f(x_i).$$

The factorization $Lf = G^T Gf$ corresponds to the decomposition $\Delta f = \div(\nabla f)$.

3.2.7 Example of a Parametric Surface

We recall that a parameterized surface is a mapping $u \in \mathcal{D} \subset \mathbb{R}^2 \mapsto \varphi(u) \in \mathcal{M}$. Whereas the continuous Laplacian is simple to define on the plane using partial derivatives, its definition on a surface requires the intervention of an arbitrary parameterization φ which makes its expression cumbersome.

Definition 16 (Laplace-Beltrami). *The Laplace-Beltrami operator on a parametric surface \mathcal{M} is defined as*

$$\sqrt{g}\Delta_{\mathcal{M}} \stackrel{\text{def.}}{=} \frac{\partial}{\partial u_1} \left(\frac{g_{22}}{\sqrt{g}} \frac{\partial}{\partial u_1} - \frac{g_{12}}{\sqrt{g}} \frac{\partial}{\partial u_2} \right) + \frac{\partial}{\partial u_2} \left(\frac{g_{11}}{\sqrt{g}} \frac{\partial}{\partial u_2} - \frac{g_{12}}{\sqrt{g}} \frac{\partial}{\partial u_1} \right)$$

where $g = \det(I_\varphi)$ and $I_\varphi = (g_{ij})_{i,j=1,2}$.

The Laplacian is however an intrinsic operator that does not depends on the chosen parameterization, as shown by the following approximation theorem.

Remark 2 (Laplacian using averaging).

$$\Delta_{\mathcal{M}} f(x) = \lim_{h \rightarrow 0} \frac{1}{|B_h(x)|} \int_{y \in \mathcal{M}} f(y)dy \quad \text{where } B_h(x) = \{y ; d_{\mathcal{M}}(x, y) \leq h\}$$

where $d_{\mathcal{M}}$ is the geodesic distance on \mathcal{M} and $h = \max_{(i,j) \in E} \|x_i - x_j\|$ is the discretization precision.

3.3 Diffusion and Regularization on Surfaces

3.3.1 Heat Diffusion

The main linear PDE for regularization of functions is the heat equation that governs the isotropic diffusion of the values of a function in time.

Definition 17 (Heat diffusion). $\forall t > 0$, one defines $F_t : M \rightarrow \mathbb{R}$ solving

$$\frac{\partial F_t}{\partial t} = -D^{-1}LF_t = -(\text{Id}_n - \tilde{W})F_t \quad \text{and} \quad \forall i \in V, F_0(i) = f(i)$$

In order to compute numerically the solution of this PDE, one can fix a time step $\delta > 0$ and use an explicit discretization in time \bar{F}_k as $F_0 = f$ and

$$\frac{1}{\delta} (\bar{F}_{k+1} - \bar{F}_k) = -D^{-1}L\bar{F}_k \implies \bar{F}_{k+1} = \bar{F}_k - \delta D^{-1}L\bar{F}_k = (\text{Id} - \delta)\bar{F}_k + \delta\tilde{W}\bar{F}_k. \quad (3.6)$$

If δ is small enough, one hopes that the discrete solution \bar{F}_k is close to the continuous time solution F_t for $t = \delta k$. This is indeed the case as proven later in these notes.

Remark 3. In order for this scheme to be stable, one needs $\delta < 1$. This is be proven later using the extension of Fourier theory to meshes.

Remark 4. If $\delta = 1$, then the discretization of the Heat equation corresponds to iterative smoothing since $\bar{F}_k = \tilde{W}^k f$. In this case stability is not guaranteed but only pathological meshes give unstable filtering (see theorem 15).

Instead of using the explicit discretization in time (3.6), one can use an implicit scheme which compute an approximate solution \tilde{F}_k at step k by solving

$$\frac{1}{\delta} (\tilde{F}_{k+1} - \tilde{F}_k) = -D^{-1}L\tilde{F}_{k+1} \implies ((\delta + 1)\text{Id}_n - \delta\tilde{W})\tilde{F}_{k+1} = \tilde{F}_k. \quad (3.7)$$

Computing \tilde{F}_k requires the solution of a sparse linear system at each step k . The implicit scheme (3.7) is thus computationally more involved than the explicit scheme (3.6). We will however see later that the implicit scheme is always stable for any value of $\delta \leq 1$.

Example 6 (Mesh smoothing). In order to smooth a mesh whose points are $x_i = (x_i^1, x_i^2, x_i^3)$, one can perform a heat diffusion for each component $f_i = (x_i^k)$, $k = 1, 2, 3$. Figure 3.5 shows an example of such a smoothing.

In practice, mesh smoothing is used to denoise a function $f = f_0 + \sigma g$ where $g \in \mathbb{R}^n$ is a realization of a gaussian white noise (each entry $g(i)$ are independent and follow a gaussian law with unit variance). The difficult task it to find an optimal stopping time t to minimize $\|F_t - f_0\|$, which is not available since one does not know f_0 . For uniformly smooth surfaces, the theory predicts that a linear filtering such as the heat equation requires a stopping time proportional to the noise level σ . This is however false for more complex surfaces such as the one used in computer graphics. In these case, alternate non linear diffusions such as non-linear PDE or wavelet thresholding usually perform better, see [27] for an overview of these methods in image processing.

Other differential equations. One can solve other partial differential equations involving the Laplacian over a 3D mesh $M = (V, E, F)$. For instance, one can consider the wave equation, which defines, for all $t > 0$, a vector $F_t \in \ell^2(V)$ as the solution of

$$\frac{\partial^2 F_t}{\partial t^2} = -D^{-1}LF_t \quad \text{and} \quad \begin{cases} F_0 = f \in \mathbb{R}^n, \\ \frac{d}{dt}F_0 = g \in \mathbb{R}^n, \end{cases} \quad (3.8)$$

In order to compute numerically the solution of this PDE, one can fix a time step $\delta > 0$ and use an explicit discretization in time \bar{F}_k as $F_0 = f$, $F_1 = F_0 + \delta g$ and for $k > 1$

$$\frac{1}{\delta^2} (\bar{F}_{k+1} + \bar{F}_{k-1} - 2\bar{F}_k) = -D^{-1}L\bar{F}_k \implies \bar{F}_{k+1} = 2\bar{F}_k - \bar{F}_{k-1} - \delta^2 D^{-1}L\bar{F}_k.$$

Figure 3.6 shows examples of the resolution of the wave equation on 3D meshes.

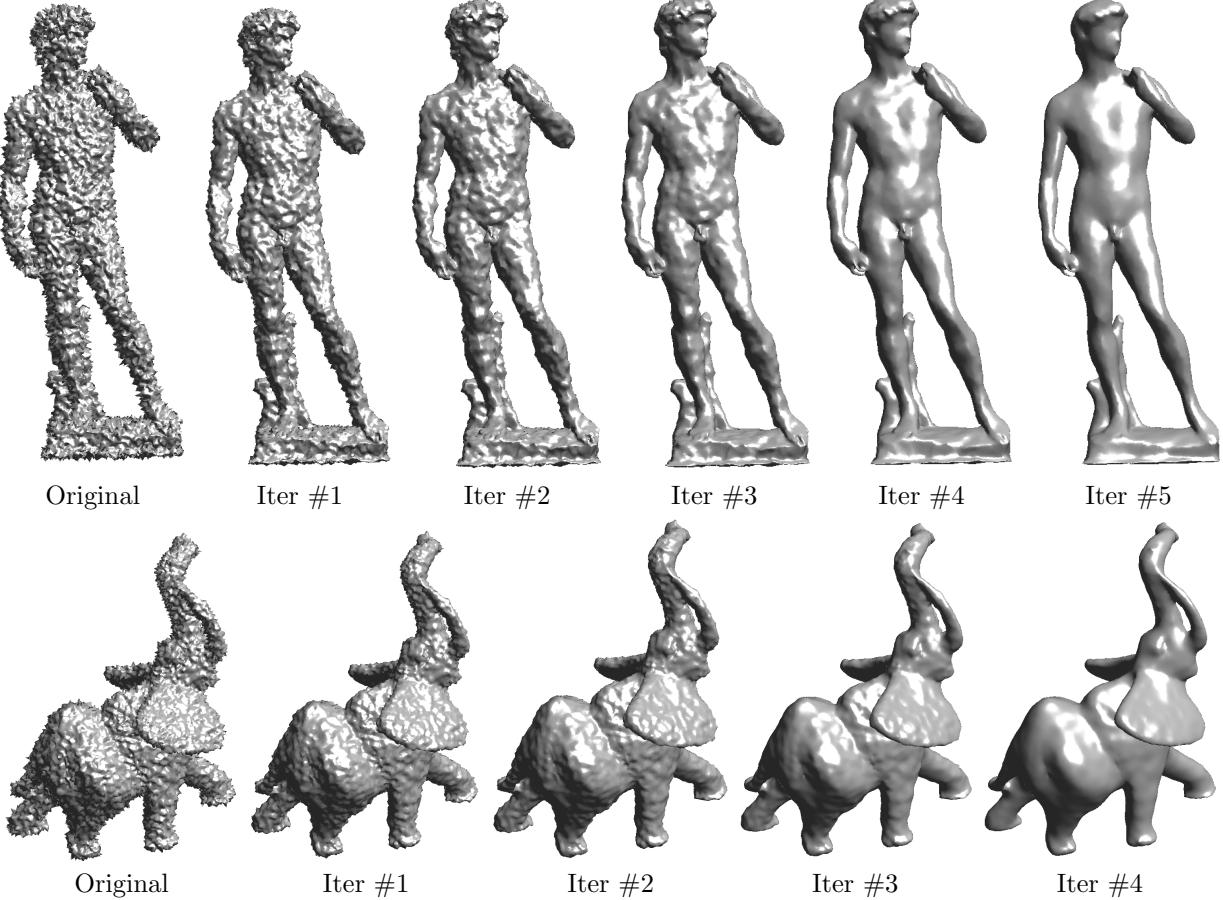


Figure 3.5: Examples of mesh denoising with the heat equation.

3.3.2 Spectral Decomposition

In order to better understand the behavior of linear smoothing on meshes, one needs to study the spectral content of Laplacian operators. This leads to the definition of a Fourier theory for meshes. The decomposition $\tilde{L} = (GD^{-1/2})^T (GD^{-1/2})$ of the Laplacian implies that it is a positive semi-definite operator. One can thus introduce the following orthogonal factorization.

Theorem 8 (Eigen-decomposition of the Laplacian). *It exists a matrix U , $U^T U = \text{Id}_n$ such that*

$$\tilde{L} = U \Lambda U^T \quad \text{where} \quad \Lambda = \text{diag}_{\omega}(\lambda_{\omega}), \quad \lambda_1 \leq \dots \leq \lambda_n.$$

The eigenvalues λ_{ω} correspond to a frequency index that ranks the eigenvectors u_{ω} of $U = (u_{\omega})_{\omega}$. One can first state some bounds on these eigenvalues.

Theorem 9 (Spectral bounds). $\forall i, \lambda_i \in [0, 2]$ and

- If M is connected then $0 = \lambda_1 < \lambda_2$.
- $\lambda_n = 2$ if and only if M is 2-colorable.

We recall the definition of a colorable graph next.

Definition 18 (Colorable graph). *A graph (V, E) is k -colorable if it exist a mapping $f : V \rightarrow \{1, \dots, k\}$ such that*

$$\forall (i, j) \in E, \quad f(i) \neq f(j).$$

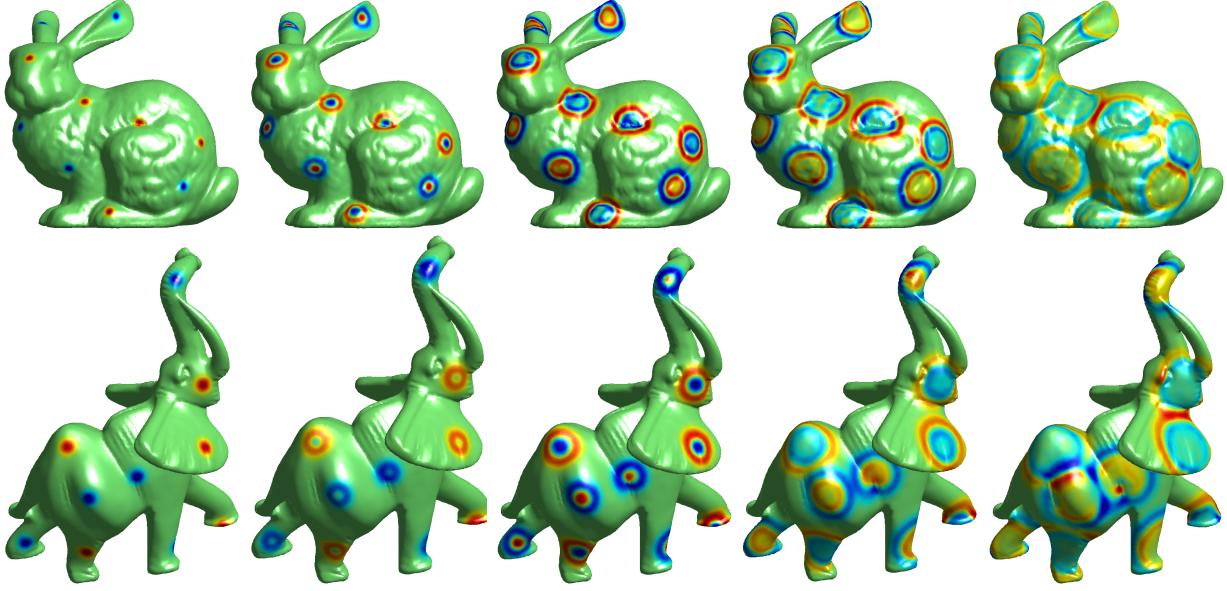


Figure 3.6: Example of evolution of the wave equation on 3D mesh. The initial condition f is a superposition of small positive and negative gaussians.

A 2-colorable graph is also called bi-partite. A 2-colorable mesh is pathological for filtering since one can split the set of vertices into two parts without inner connexions. The filtering process can oscillate by exchanging values between these sets, thus never converging.

The orthogonal eigen-basis $U = (u_\omega)_\omega$ is an orthogonal basis of the space $\mathbb{R}^n \simeq \ell^2(V)$, which can be written as

$$u_\omega : \begin{cases} V & \longrightarrow \mathbb{R} \\ i & \longmapsto u_\omega(x_i) \end{cases}$$

The orthogonality means that $\langle u_\omega, u_{\omega'} \rangle = \delta_{\omega}^{\omega'}$. This basis allows to compute an orthogonal decomposition of any functions f

$$\forall f \in \ell^2(V), \quad f = \sum_\omega \langle f, u_\omega \rangle u_\omega.$$

Having such a tool allows to split a function f in elementary contributions $\langle f, u_\omega \rangle$ with a control in the energy because of orthogonality

$$\|f\|^2 = \sum_\omega |\langle f, u_\omega \rangle|^2.$$

Figure 3.7 shows some examples of eigenfunctions depicted using color ranging from blue (negative values of the eigenfunction) to red (positive values). One can see that these functions are oscillating, in a way similar to the traditional Fourier basis. In some sense (made more precise latter), this basis is the extension of the Fourier basis to meshes. A function u_ω corresponding to a large spectral value λ_ω is highly oscillating and corresponds thus intuitively to a high frequency atom.

Extracting numerically eigenvectors from a large matrix is a difficult problem. If the matrix is sparse, a method of choice consists in using iterative powers of a shifted version of the laplacian. One starts from a random initial vector v_0 and iterates

$$v_{k+1} = \frac{w_{k+1}}{\|w_{k+1}\|} \quad \text{where} \quad w_{k+1} = (\tilde{L} - \lambda \text{Id}_n)^{-1} v_k. \quad (3.9)$$

These iterates converges to the eigenvectors corresponding to the eigenvalue the closest to λ , as staten in the following theorem.

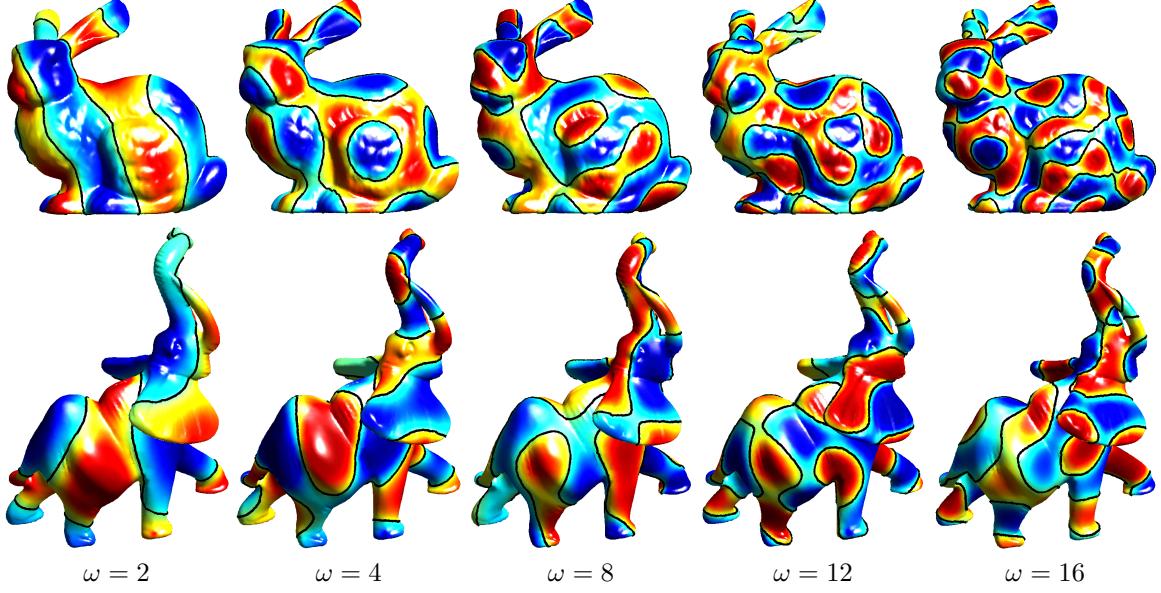


Figure 3.7: Examples of eigenvectors u_ω of the Laplacian \tilde{L} . The blue colors indicated negative values, red colors positive ones. The black curve is the 0 level set of the eigenvector.

Theorem 10 (Inverse iterations). *For a given shift λ , lets denote*

$$\omega^* = \operatorname{argmin}_\omega |\lambda - \lambda_\omega| \quad \text{and} \quad \omega^+ = \operatorname{argmin}_{\omega \neq \omega^*} |\lambda - \lambda_\omega|$$

If $|\lambda - \lambda_{\omega^*}| < |\lambda - \lambda_{\omega^+}|$, then

$$v_k \xrightarrow{k \rightarrow +\infty} u_{\omega^*} \quad \text{and} \quad \langle Lv_k, v_k \rangle \xrightarrow{k \rightarrow +\infty} \lambda_{\omega^*}.$$

The speed of convergence of these inverse iterations is governed by the conditioning of $(\tilde{L} - \lambda \operatorname{Id}_n)^{-1}$ since

$$\|v_k - u_{\omega^*}\| \leq C\rho(\lambda)^k \quad \text{where} \quad \rho(\lambda) \stackrel{\text{def.}}{=} \frac{|\lambda - \lambda_{\omega^*}|}{|\lambda - \lambda_{\omega^+}|} < 1.$$

The smallest $\rho(\lambda)$ is, the faster the method converges.

In order to compute an iteration (3.9) of the method, one needs to solve a sparse linear system $Aw_{k+1} = v_k$ whith $A = \tilde{L} - \lambda \operatorname{Id}_n$. In order to do so, one can use a direct method such as LU factorization. The advantage of such an approach is that the factorization is computed once for all and can be re-used to solve very quickly at each step k . These factorization are however quite slow to compute especially for large matrices. For large problems, one can solve this linear system using an iterative algorithm such as conjugate gradient. These iterative method are attractive for sparse matrices, but a fast convergence requires $1/\rho(\lambda)$, the conditioning of $\tilde{L} - \lambda \operatorname{Id}_n$ to be not large, with is contradictory with the constraint for iterations 3.9 to converge fast.

3.3.3 Spectral Theory on a Regular Grid

In the particular case of a 1D or 2D lattice, the eigenfunctions defined earlier correspond exactly to the Fourier basis used in the discrete Fourier transform.

Theorem 11 (Spectrum in 1D). *For a 1D regular lattice,*

$$u_\omega(k) = \frac{1}{\sqrt{n}} \exp\left(\frac{2i\pi}{n} k\omega\right) \quad \text{and} \quad \lambda_\omega = 4 \sin^2\left(\frac{2\pi}{n} \omega\right).$$

Theorem 12 (Spectrum in 2D). *For a 2D regular lattice, $n = n_1 n_2$, $\omega = (\omega_1, \omega_2)$*

$$u_\omega(k) = \frac{1}{\sqrt{n}} \exp\left(\frac{2i\pi}{n} \langle k, \omega \rangle\right) \quad \text{and} \quad \lambda_\omega = 4 \left(\sin^2\left(\frac{2\pi}{n_1} \omega_1\right) + \sin^2\left(\frac{2\pi}{n_2} \omega_2\right) \right).$$

As already mentioned, on a mesh, the eigenvectors of \tilde{L} correspond to a extension of the Fourier basis to meshes. The definition of the Fourier transform on meshes requires a little care since a diagonal normalization by D is used as defined next.

Definition 19 (Manifold-Fourier transform). *For $f \in \ell^2(V)$,*

$$\Phi(f)(\omega) = \hat{f}(\omega) \stackrel{\text{def.}}{=} \langle D^{1/2} f, u_\omega \rangle \iff \Phi(f) = \hat{f} = U^T D^{1/2}.$$

where $(u_\omega)_\omega$ are the eigenvectors of \tilde{L} .

One can note that there is still a degree of freedom in designing this Fourier transform since one can use any local weighting (for instance combinatorial, distance or conformal). Depending on the application, one might need to use weights depending only on the topology of the mesh (combinatorial for mesh compression).

A major theoretical interest of this Fourier transform is that it diagonalizes local averaging operators.

Theorem 13 (Spectral smoothing). *One has $\Phi \tilde{W} \Phi^{-1} = \text{Id}_n - \Lambda$ and thus for any function f*

$$\widehat{\tilde{W}f}(\omega) = (1 - \lambda_\omega) \hat{f}(\omega)$$

This diagonalization allows to prove the convergence of iterative smoothing.

Theorem 14 (Convergence of iterated smoothing). *If $\lambda_n < 2$ (i.e. M is not 2-colorable), then for any function f*

$$\tilde{W}^k f \xrightarrow{k \rightarrow +\infty} \frac{1}{n} \sum_{i \in V} f_i.$$

3.3.4 Spectral Resolution of the Heat Diffusion

Recall that the heat diffusion is defined as

$$\forall t > 0, \quad \frac{\partial F_t}{\partial t} = -D^{-1} L F_t = -(\text{Id}_n - \tilde{W}) F_t$$

Using the manifold Fourier expansion $\hat{F}_t \stackrel{\text{def.}}{=} U^T D^{1/2} F_t$, this differential equation can be re-written as

$$\frac{\partial \hat{F}_t(\omega)}{\partial t} = -\lambda_\omega \hat{F}_t(\omega) \implies \hat{F}_t(\omega) = \exp(-\lambda_\omega t) \hat{f}(\omega). \quad (3.10)$$

This allows to study the convergence of the continuous heat equation.

Theorem 15 (Convergence of heat equation). *If \mathcal{M} is connected,*

$$F_t \xrightarrow{t \rightarrow +\infty} \frac{1}{n} \sum_{i \in V} f_i.$$

Recall that the heat equation is discretized using the following explicit and implicit schemes, equations (3.6) and (3.7)

$$\begin{cases} \bar{F}_k = (1 - \delta) \bar{F}_k + \delta \tilde{W} \bar{F}_k, \\ ((1 + \delta) \text{Id}_n - \delta \tilde{W}) \tilde{F}_{k+1} = \tilde{F}_k. \end{cases}$$

These filtering iterations can be re-written over the Fourier domain as

$$\begin{cases} \widehat{\bar{F}_{k+1}}(\omega) = (1 - \delta \lambda_\omega) \widehat{\bar{F}_k}(\omega), \\ \widehat{\tilde{F}_{k+1}}(\omega) = \frac{1}{(1 + \delta \lambda_\omega)} \widehat{\tilde{F}_k}(\omega). \end{cases}$$

This allows to state the stability and convergence of the finite difference discretization.

Theorem 16 (Convergence of discretization). *The explicit scheme is stable if $\delta < 1$. The implicit scheme is always stable. One has*

$$\begin{cases} \bar{F}_{t/\delta} \xrightarrow{\delta \rightarrow 0} F_t, \\ \tilde{F}_{t/\delta} \xrightarrow{\delta \rightarrow 0} F_t. \end{cases}$$

with the restriction that for the explicit scheme, the mesh must not be 2-colorable.

Other Differential Equations. The manifold Fourier transform can also be used to solve the wave equation (3.8) since

$$\frac{\partial^2 \hat{F}_t(\omega)}{\partial t^2} = -\lambda_\omega \hat{F}_t(\omega) \implies \hat{F}_t(\omega) = \cos(\sqrt{\lambda_\omega} t) \hat{f}(\omega) + \frac{1}{\sqrt{\lambda_\omega}} \sin(\sqrt{\lambda_\omega} t) \hat{g}(\omega).$$

3.3.5 Quadratic Regularization

Instead of using a PDE for regularization, one can try to find a new function that is both close to the original one f and that is smooth in a certain sense. This leads to the notion of quadratic regularization, where one uses a Laplacian as a smoothness prior on the recovered function.

Definition 20 (Quadratic regularizer). *For $t > 0$, one defines*

$$F_t^q = \underset{g \in \mathbb{R}^n}{\operatorname{argmin}} \|f - g\|^2 + t \|\tilde{G}g\|^2 \quad \text{where} \quad \tilde{G} = GD^{-1/2}.$$

This optimization replaces $f \in \ell^2(V)$ by $F_t^q \in \ell^2(V)$ with small gradients. This optimization can be found in closed form by inverting a sparse linear system.

Theorem 17 (Solution of quadratic regularization). *F_t^q is unique and*

$$F_t^q = (\operatorname{Id}_n + t\tilde{L})^{-1} f.$$

Over the Fourier domain, this inversion reads

$$\hat{F}_t^q(\omega) = \frac{1}{1 + t\lambda_\omega} \hat{f}(\omega).$$

This corresponds to an attenuation of the high frequency content of f , in a way very similar to equation (3.10).

Once again, similarly to the heat equation, the spectral expression of the quadratic regularizer allows to study its convergence for large t .

Theorem 18 (Convergence of quadratic regularization). *If \mathcal{M} is connected,*

$$F_t^q \xrightarrow{t \rightarrow +\infty} \frac{1}{n} \sum_{i \in V} f_i.$$

3.3.6 Application to Mesh Compression

We have shown how the Fourier basis on meshes can be used to compute in a diagonal fashion filtering, heat diffusion and quadratic regularization. This Fourier transform is however of little interest in practice, since the original filterings (or finite difference approximation of the heat equation) are usually faster to compute directly than over the Fourier domain. The Fourier transform is thus mainly of theoretical interest in these cases since it allows to prove convergence results.

Another class of applications makes use of an orthogonal expansion such as the Fourier one to perform mesh compression. This section shows how to compute a linear M -term approximation in this Fourier basis and to do mesh compression. We refer to the survey [1] for more advanced non-linear mesh compression methods.

The orthogonal basis $U = (u_\omega)_\omega$ of $\ell^2(V) \simeq \mathbb{R}^n$, where $\tilde{L} = U\Lambda U^\top$ allows to define a linear approximation as followed.

Definition 21 (Linear M -term approximation). *For any $M > 0$, the linear M -term approximation of f is*

$$f = \sum_{\omega=1}^n \langle f, u_\omega \rangle u_\omega \stackrel{M\text{-term approx.}}{\implies} f_M \stackrel{\text{def.}}{=} \sum_{\omega=1}^M \langle f, u_\omega \rangle u_\omega.$$

The quality of the approximation is measured using the error decay, which can in turn be estimated using the removed coefficients

$$E(M) \stackrel{\text{def.}}{=} \|f - f_M\|^2 = \sum_{\omega>M} |\langle f, u_\omega \rangle|^2.$$

A good orthogonal basis U is a basis for which $E(M)$ decays fast on the signals of interest. Equivalently, a fast decay of E with M corresponds to a fast decay of $|\langle f, u_\omega \rangle|$ for large ω . Figure 3.8 shows the decay of the Fourier spectrum for two different functions defined on a 3D mesh. The smooth function (left in the figure) exhibits a fast decay of its spectrum, meaning that it can be well approximated with only a few Fourier coefficients.

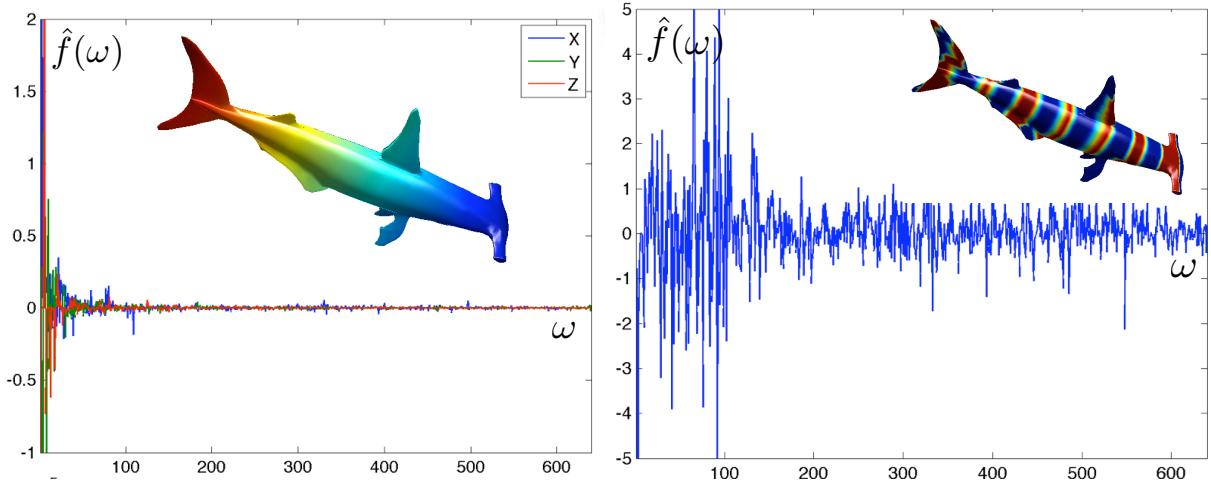


Figure 3.8: Examples of Fourier spectrum for a smooth and a non-smooth function.

We recall that the Fourier atoms

$$\forall \omega \in \mathbb{Z}, \quad u_\omega(x) = \frac{1}{\sqrt{2\pi}} e^{i\omega x}$$

are the eigenvectors of the compact, symmetric, semi-definite negative operator $f \mapsto f''$ (that should be defined on the Hilbert space of twice Sobolev derivable functions). This set of function is also an Hilbert basis of the space $L^2(\mathbb{R}/(2\pi\mathbb{Z}))$ of 2π -periodic square integrable functions and a Fourier coefficient is $\hat{f}(\omega) \stackrel{\text{def.}}{=} \langle f, u_\omega \rangle$.

Approximation theory studies this linear error decay for classical functional spaces. One can for instance study the Fourier expansion over euclidean spaces.

Theorem 19 (Fourier in 1D). *If f is C^α regular on $\mathbb{R}/(2\pi\mathbb{Z})$,*

$$|\hat{f}(\omega)| \leq \|f^{(\alpha)}\|_\infty |\omega|^{-\alpha}.$$

This result can be proven with a simple integration by parts. A slightly more difficult result shows that the linear approximation error decays like $M^{-\alpha}$.

Theorem 20 (Fourier approximation). *If f is C^α on $\mathbb{R}/(2\pi\mathbb{Z})$, then it exist $C > 0$ such that*

$$\sum_{\omega} |\omega|^{2\alpha} |\langle f, u_{\omega} \rangle|^2 < +\infty \implies E(M) \leq CM^{-\alpha}.$$

This kind of results can be extended to continuous surfaces thanks to the continuous Laplacian. We suppose that \mathcal{M} is a surface parameterized by φ , and a function $f = \varphi \circ \bar{f}$ is defined on it. By definition, this function f is C^α if \bar{f} is C^α in euclidean space. For a compact surface \mathcal{M} , the Laplace-Beltrami operator $\Delta_{\mathcal{M}}$ is symmetric (for the inner product on the surface), is negative semi-definite and has a discrete spectrum $\Delta_{\mathcal{M}} u_{\omega} = -\lambda_{\omega} u_{\omega}$ for $\omega \in \mathbb{N}$. The functions $\{u_{\omega}\}_{\omega}$ are an orthogonal basis for function of finite energy on the surface $L^2(\mathcal{M})$. The inner product of an arbitrary smooth function $f \in C^\alpha(\mathcal{M})$ can be bounded using integration by parts

$$\langle f, u_{\omega} \rangle = \frac{1}{\lambda_{\omega}^k} \langle \Delta_{\mathcal{M}}^k f, u_{\omega} \rangle \implies |\langle f, u_{\omega} \rangle| \leq \frac{\|f\|_{C^\alpha}}{\lambda_{\omega}^{\alpha/2}}.$$

This proves the efficiency of the Fourier basis on surfaces to approximate smooth functions.

When computing the M -term approximation f_M of f one removes the small amplitude Fourier coefficients of the orthogonal expansion of f . Figure 3.9 shows some examples of mesh approximation where one retains an increasing number of Fourier coefficients. Mesh compression is only a step further, since one also need to code the remaining coefficients. This requires first quantifying the coefficients up to some finite precision and then binary code these coefficients into a file.

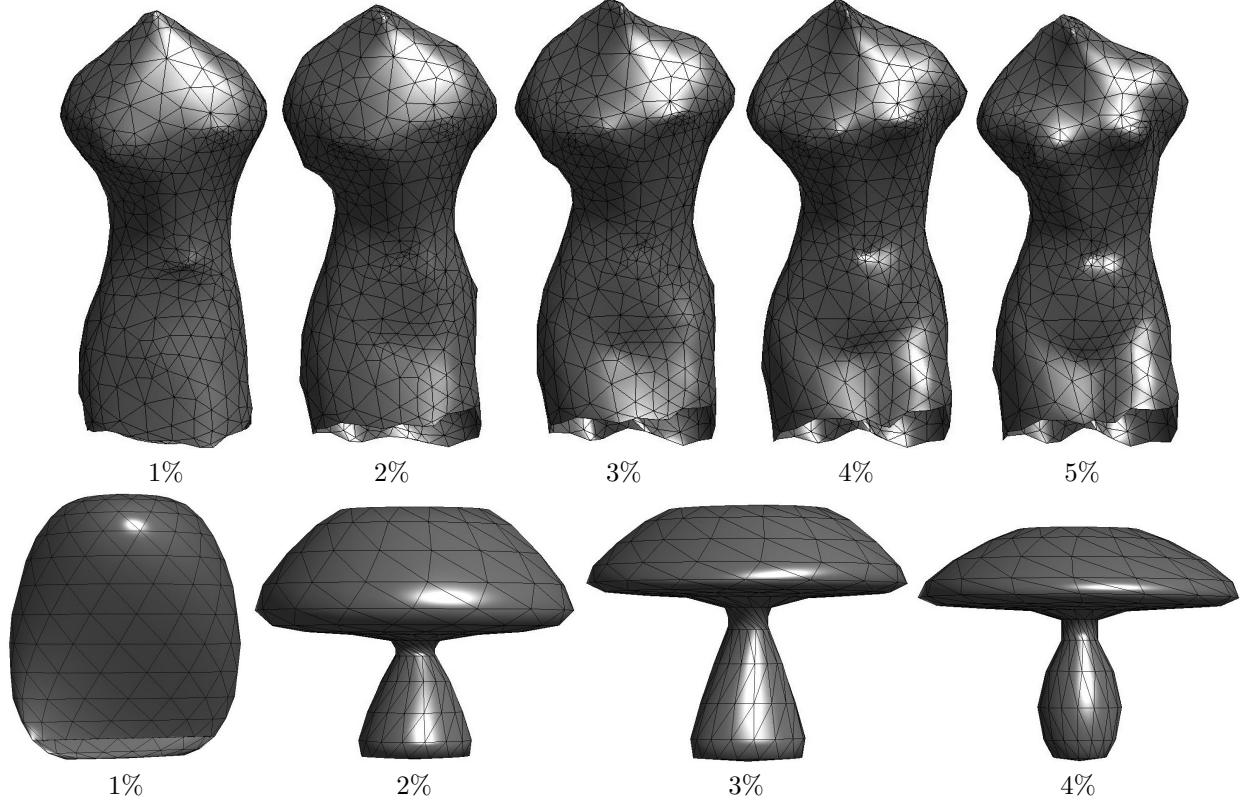


Figure 3.9: Examples of spectral mesh compression.

3.3.7 Application to Mesh Parameterization

This section is restricted to the study of meshes that can be globally parameterized on a plane. It means that they are topologically equivalent to a 2D disk. More complex meshes should be first segmented in cells that are equivalent to a disk.

A parameterization of a continuous surface \mathcal{M} is a bijection

$$\psi : \mathcal{M} \longrightarrow \mathcal{D} \subset \mathbb{R}^2.$$

A similar definition applies to a discrete mesh where one computes a 2D position $\psi(i)$ for all the vertices $i \in V$ and then interpolates linearly the mapping to the whole piecewise linear geometric mesh. This section explains the basics of linear methods for mesh parameterization. We refer to various surveys [21, 39] for more details on mesh parameterization.

Usually, a 2D mesh is computed from range scanning or artistic modeling, so it does not come with such a parameterization. In order to perform texture mapping or more general mesh deformations, it is however important to use such a parameterization. Since many bijections are possible to layout the mesh in 2D, the mapping ψ has to satisfy additional smoothness assumptions. Classically, one requires that each coordinate of ψ has a vanishing Laplacian (it is thus harmonic) outside a set of constrained vertices that enforce boundary conditions.

More precisely, $\psi = (\psi_1, \psi_2)$ is the solution of

$$\begin{cases} \forall i \notin \partial\mathcal{M}, & (L\psi_1)(i) = (L\psi_2)(i) = 0 \\ \forall i \in \partial\mathcal{M}, & \psi(i) = \psi^0(i) \in \partial\mathcal{D}, \end{cases}$$

where $\partial\mathcal{M}$ is the boundary of the mesh, which consists in vertices whose face ring is not homeomorphic to a disk but rather to a half disk. This formulation requires the solution of two sparse linear systems (one for each coordinate of ψ).

The boundary condition $\psi^0(i)$ for $i \in \partial\mathcal{M}$ describes a 1D piecewise linear curve in the plane, that is fixed by the user. In the following, we will see that this curve should be convex for the parameterization to be bijective.

Remark 5. For such an harmonic parameterization, each point is the average of its neighbors since

$$\forall i, \quad \psi(i) = \frac{1}{\sum_j w_{ij}} \sum_{(i,j) \in E} w_{i,j} \psi(j).$$

The powerful feature of this linear parameterization method is that it can be proven to produce a valid (bijective) parameterization as long as the constrained position (boundary values of ψ) are along a convex curve.

Theorem 21 (Tutte theorem). *If $\forall (i,j) \in E, w_{ij} > 0$, and if $\partial\mathcal{D}$ is a convex curve, then ψ is a bijection.*

Figure 3.10 shows several examples of parameterizations. One is free to use any laplacian (combinatorial, distance or conformal) as long as it produces positive weights. There is a issue with the conformal weights, which can be negative if the mesh contains obtuse triangles. In practice however it leads to the best results. The efficiency of a parameterization can be measured by some amount of distortion induced by the planar mapping. Linear methods cannot hope to cope with large isoperimetric distortions (for instance large extrusions in the mesh) since harmonicity leads to clustering of vertices.

3.3.8 Application to Mesh Flattening

One of the difficulty with linear parameterization methods is that they require to set up the positions of the vertices along the boundary of the mesh. In order to let the boundary free to evolve and find some optimal shape, one can replace the fixed point constraint by a global constraints of unit variance as follow

$$\min_{\psi_1, \psi_2 \in \mathbb{R}^n} \|\tilde{G}\psi_1\|^2 + \|\tilde{G}\psi_2\|^2 \quad \text{with} \quad \begin{cases} \|\psi_i\| = 1, \\ \langle \psi_1, \psi_2 \rangle = 0, \\ \langle \psi_i, 1 \rangle = 0. \end{cases}$$

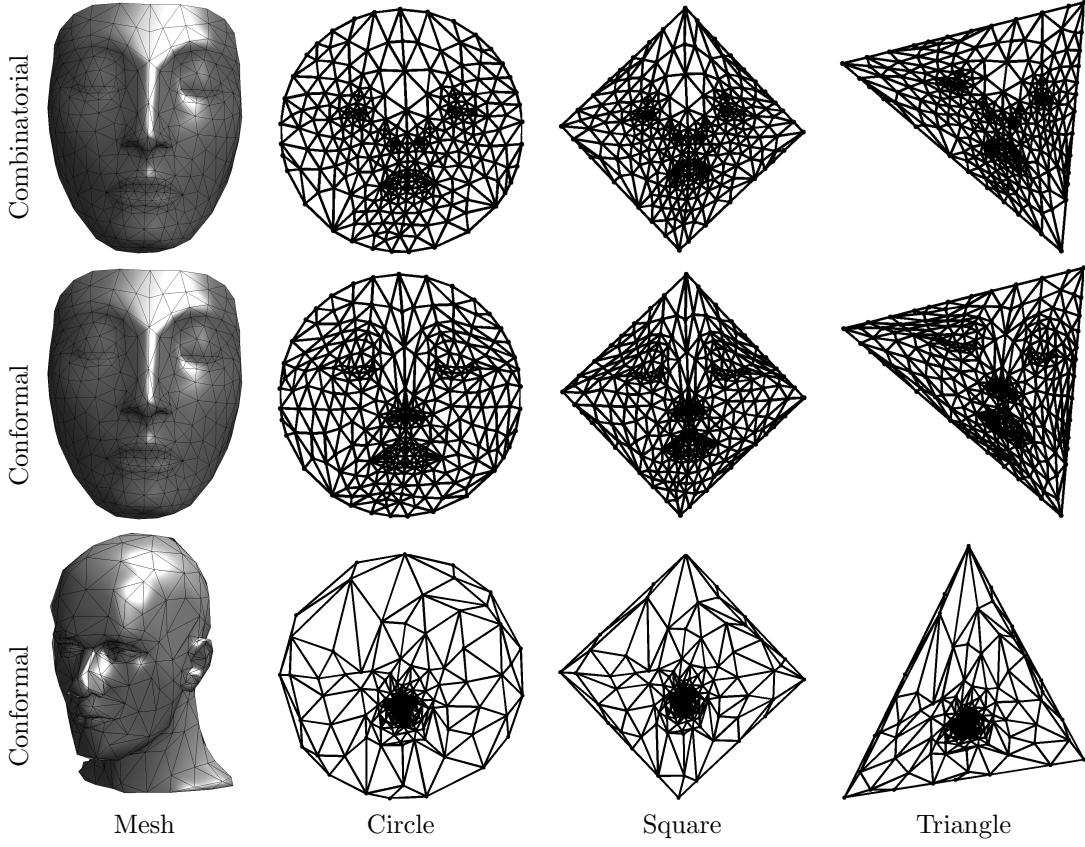


Figure 3.10: Examples of mesh parameterizations.

This optimization problem also has a simple global solution using eigenvectors of the Laplacian.

Theorem 22 (Mesh flattening solution). *The mesh flattening solution is given by*

$$\text{Span}(\psi_1, \psi_2) = \text{Span}(u_1, u_2) \quad \text{where} \quad \tilde{L} = U \Lambda U^T.$$

In order to compute this flattening, one thus needs to extract 2 eigenvectors from a sparse matrix. Note however that, in contrast to linear parameterization schemes, this flattening is not ensured to be bijective. Figure 3.11 shows that for meshes with large distortion, this flattening indeed leads to wrong parameterizations.

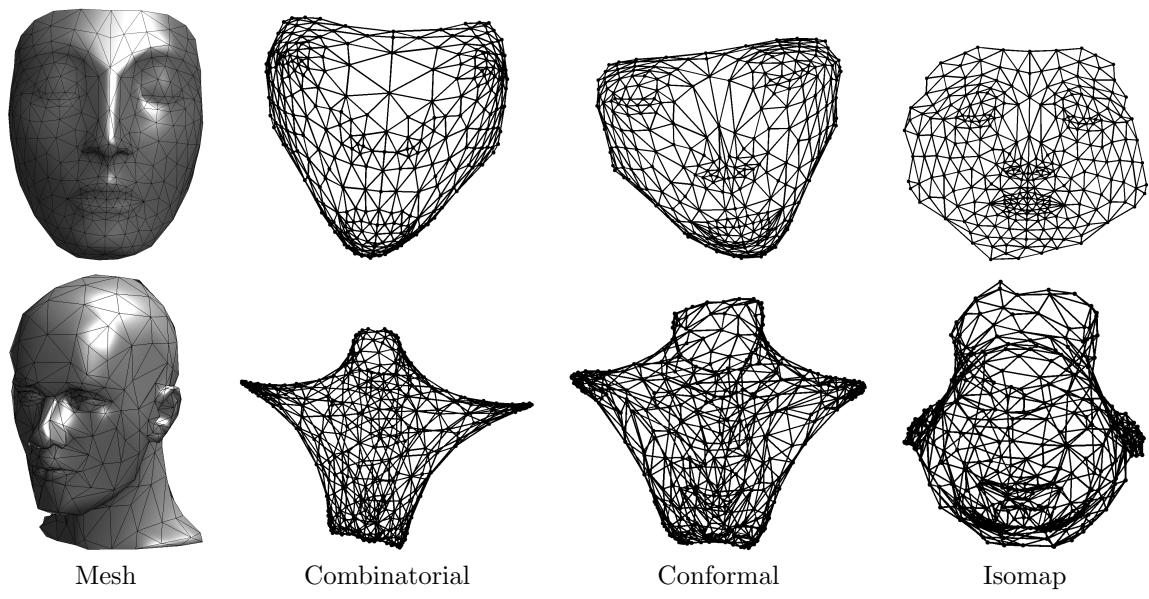


Figure 3.11: Examples of mesh flattening.

Chapter 4

Wavelets

The reference for this chapter is [28].

4.1 Multi-resolution Approximation Spaces

A multiresolution approximation of $L^2(\mathbb{R})$ is a set of nested closed subspaces $(V_j)_j$

$$L^2(\mathbb{R}) \supset \dots \supset V^{j-1} \supset V_j \supset V_{j+1} \supset \dots \supset \{0\} \quad (4.1)$$

which must be related one from each other by dyadic scaling and must also be stable by dyadic translation

$$f \in V_j \iff f(\cdot/2) \in V_{j+1} \quad \text{and} \quad f \in V_j \iff \forall n \in \mathbb{Z}, f(\cdot + t2^j) \in V_j$$

So large j corresponds to coarse approximation spaces, and 2^j is often call the “scale”.

The limit on the left of (4.1) means that $\cup_j V_j$ is dense in $L^2(\mathbb{R})$, or equivalently that $P_{V_j}(f) \rightarrow f$ as $j \rightarrow -\infty$ where P_V is the orthogonal projector on V

$$P_V(f) = \operatorname{argmin}_{f' \in V} \|f - f'\|.$$

The limit on the right of (4.1) means that $\cap_j V_j = \{0\}$, or equivalently that $P_{V_j}(f) \rightarrow 0$ as $j \rightarrow +\infty$.

Scaling functions. We also require that there exists a scaling function $\varphi \in L^2(\mathbb{R})$ so that

$$\{\varphi(\cdot - n)\}_n \text{ is an Hilertian orthonormal basis of } V_0.$$

By the dilation property, this implies that

$$\{\varphi_{j,n}\}_n \text{ is an Hilertian orthonormal basis of } V_j \quad \text{where} \quad \varphi_{j,n} \stackrel{\text{def.}}{=} \frac{1}{2^{j/2}} \varphi\left(\frac{\cdot - 2^j n}{2^j}\right).$$

Note that one then has

$$P_{V_j}(f) = \sum_n \langle f, \varphi_{j,n} \rangle \varphi_{j,n}.$$

Figure 4.1 illustrates the translation and scaling effect.

A typical example is approximation by piecewise constant signals

$$V_j \stackrel{\text{def.}}{=} \{f \in L^2(\mathbb{R}) ; \forall n, f \text{ is constant on } [2^j n, 2^j(n+1)[\} , \quad (4.2)$$

in which case one can use $\varphi = 1_{[0,1]}$ and $\varphi_{j,n} = 2^{-j/2} 1_{[2^j n, 2^j(n+1)[}$.

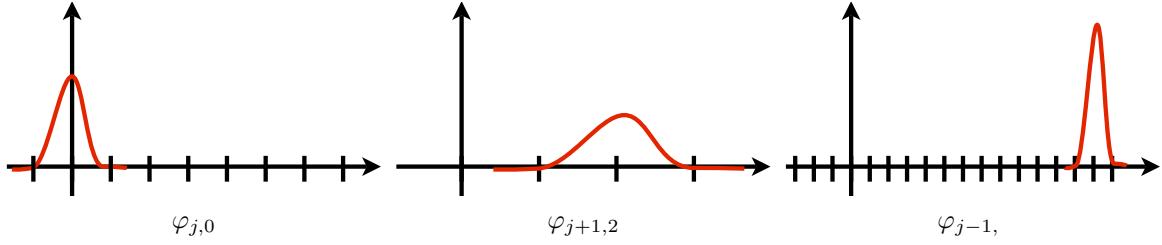


Figure 4.1: Translation and scaling to generate approximation spaces.

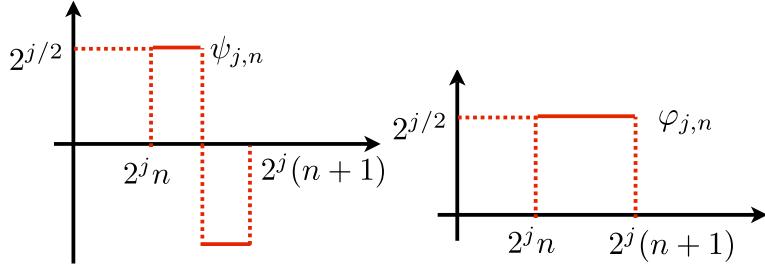


Figure 4.2: Haar scaling (left) and wavelet (right) functions.

Spectral orthogonalization. In many case of practical interest, the space V_j is described by a translation-invariant basis which is not-orthogonal, $V_j = \text{Span}(\theta(\cdot - n))_{n \in \mathbb{Z}}$. The following proposition shows how to orthogonalize it using the Fourier transform.

Proposition 9. For $\theta \in L^2(\mathbb{R})$ (assumed regular and with fast enough decay), $\{\theta(\cdot - n)\}_{n \in \mathbb{Z}}$ is orthonormal if and only if

$$\forall \omega, \quad A(\omega) \stackrel{\text{def.}}{=} \sum_k |\hat{\theta}(\omega - 2\pi k)|^2 = 1.$$

If there exists $0 < a \leq b < +\infty$ such that $a \leq A(\omega) \leq b$, then φ defined by

$$\hat{\varphi}(\omega) = \frac{\hat{\theta}(\omega)}{\sqrt{A(\omega)}}$$

is such that $\{\varphi(\cdot - n)\}_{n \in \mathbb{Z}}$ is an Hilbertian basis of $\text{Span}\{\theta(\cdot - n)\}_{n \in \mathbb{Z}}$.

Proof. One has that $\{\theta(\cdot - n)\}_{n \in \mathbb{Z}}$ is orthonormal if and only if

$$\langle \theta, \theta(\cdot - n) \rangle = (\theta \star \bar{\theta})(n) = \delta_{0,n} \stackrel{\text{def.}}{=} \begin{cases} 0 & \text{if } n = 0, \\ 1 & \text{otherwise.} \end{cases}$$

where $\bar{\theta} = \theta(\cdot)$ and δ_0 is the discrete Dirac vector. The Poisson summation formula (1.8) then reads

$$\sum_n \mathcal{F}(\theta \star \bar{\theta})(\omega - 2\pi n) = \sum_n \delta_{0,n} e^{-i\omega} = 1.$$

We conclude with the Fourier-convolution formula (2.6) shows that $\mathcal{F}(\theta \star \bar{\theta})(\omega) = \hat{\theta}(\omega) \hat{\theta}(\omega)^* = |\hat{\theta}(\omega)|^2$ which leads to the desired formula, and it is if and only if. Normalizing by $1/\sqrt{A(\omega)}$, which is a bounded function, shows that $\hat{\varphi}$ satisfies $\sum_k |\hat{\varphi}(\omega - 2\pi k)|^2 = 1$. \square

A typical example of application is spline (e.g. cubic ones) interpolations, which are generated by the box-spline function θ which is a piecewise polynomial with a compact support.

4.2 Multi-resolution Details Spaces

The details spaces are defined as orthogonal complement of $V_j \subset V_{j-1}$, which is legit because these are closed subspaces

$$\forall j, \quad W_j \text{ is such that } V_{j-1} = V_j \oplus^\perp W_j.$$

This leads to the following sequence of embedded spaces

$$\begin{array}{ccccccc} L^2(\mathbb{R}) & \longrightarrow & \cdots & \swarrow & V_{j-1} & \longrightarrow & V_j \\ & & & & \searrow & & \swarrow \\ & & & & W_{j-1} & & W_j \\ & & & & & & \searrow \\ & & & & & & W_{j+1} \end{array} \longrightarrow \cdots \longrightarrow \{0\}$$

Once again, we suppose that W_0 has an Hilbertian ortho-basis of the form $\{\psi(\cdot - n)\}_n$, so that

$$\{\psi_{j,n}\}_n \text{ is an Hilbertian orthonormal basis of } V_j \quad \text{where} \quad \psi_{j,n} \stackrel{\text{def.}}{=} \frac{1}{2^{j/2}} \psi\left(\frac{\cdot - 2^j n}{2^j}\right).$$

Due to the orthogonal complementarity property, one has

$$L^2(\mathbb{R}) = \bigoplus_{j=-\infty}^{j=+\infty} W_j = V_{j_0} \bigoplus_{j \leq j_0} V_j.$$

This means that for all $f \in L^2(\mathbb{R})$, one has the following convergence in $L^2(\mathbb{R})$,

$$f = \lim_{(j_-, j_+) \rightarrow (-\infty, +\infty)} \sum_{j=j_-}^{j_+} P_{V_j} f = \lim_{j_+ \rightarrow +\infty} P_{j_0} f + \sum_{j=j_0}^{j_+} P_{V_j} f.$$

This decomposition shows that

$$\{\psi_{j,n} ; (j, n) \in \mathbb{Z}^2\}$$

is an Hilbertian orthogonal basis of $L^2(\mathbb{R})$, which is called a wavelet basis. One also have a “truncated” ortho-basis

$$\{\psi_{j,n} ; j \leq j_0, n \in \mathbb{Z}\} \cup \{\varphi_{j_0,n} ; n \in \mathbb{Z}\}.$$

A (forward) Wavelet transform corresponds to the computation of all the inner products of some function f with the elements of these basis.

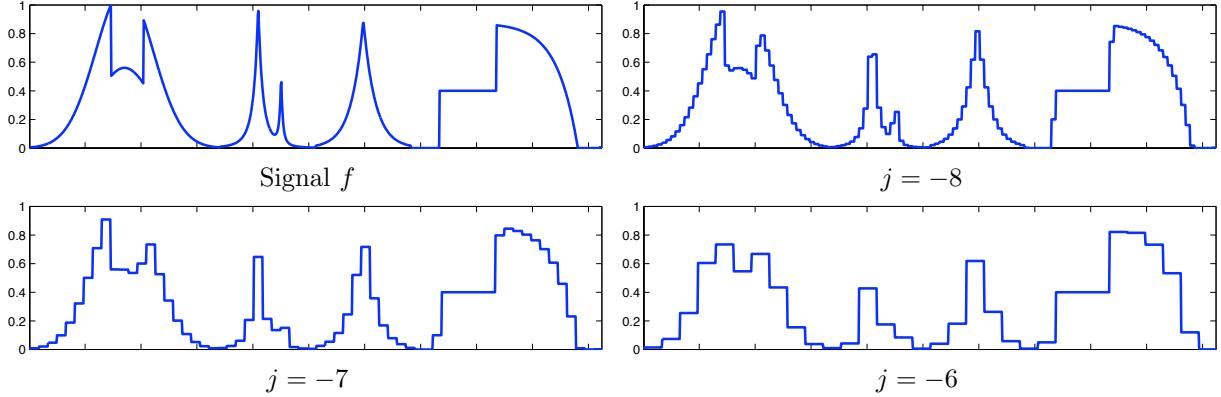


Figure 4.3: 1-D Haar multiresolution projection $P_{V_j} f$ of a function f .

Haar wavelets. For the Haar multiresolution (4.2), one has

$$W_j = \left\{ f ; \forall n \in \mathbb{Z}, f \text{ constant on } [2^{j+1}n, 2^{j+1}(n+1)) \text{ and } \int_{n2^j}^{(n+1)2^j} f = 0 \right\}. \quad (4.3)$$

A possible choice for a mother wavelet function is

$$\psi(t) = \frac{1}{\sqrt{2}} \begin{cases} 1 & \text{for } 0 \leq t < 1/2, \\ -1 & \text{for } 1/2 \leq t < 1, \\ 0 & \text{otherwise,} \end{cases}$$

as shown on Figure 4.2, right.

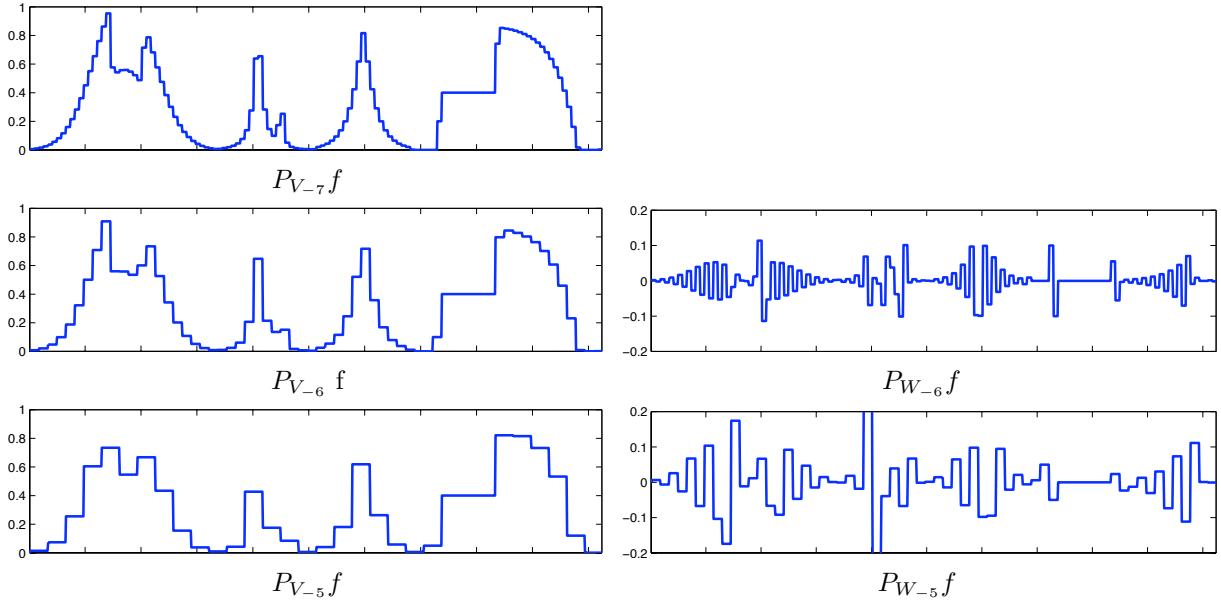


Figure 4.4: Projection on Haar approximation spaces (left) and detail spaces (right).

Figure 4.4 shows examples of projections on details spaces, and how they can be derived from projection on approximation spaces.

4.3 On Bounded Domains

On a periodic bounded domain $\mathbb{T} = \mathbb{R}/\mathbb{Z}$ (note that we use here 1-periodicity, in contrast to the convention we used for Fourier series of 2π -periodicity), one obtains an orthogonal wavelet basis of $L^2(\mathbb{T})$ by periodizing the original wavelets, and also restricting the translation points $2^j n$ to be in $[0, 1]$, i.e. $0 \leq n < 2^{-j}$. Similarly to (1.6), the periodization of a function $f \in L^1(\mathbb{R})$ is the function

$$f^P = \sum_{n \in \mathbb{Z}} f(\cdot - n) \in L^1(\mathbb{T}).$$

The wavelet basis is thus defined as

$$\{\psi_{j,n}^P ; j \leq j_0, 0 \leq n < 2^{-j}\} \cup \{\varphi_{j_0,n}^P ; 0 \leq n < 2^{-j_0}\}.$$

and one verifies that it defines an Hilbertian ortho-basis of $L^2(\mathbb{T})$. It is possible to define wavelet basis using Neumann (mirror) boundary conditions, but this is more involved.

4.4 Fast Wavelet Transform

4.4.1 Discretization

We now work over \mathbb{R}/\mathbb{Z} . The modelling hypothesis is that one has access to a discrete signal $a_J \in \mathbb{R}^N$ with $N = 2^{-J}$ at some fixed scale 2^J , and that this signal exactly matches the inner-product with the scaling functions, i.e.

$$\forall n \in \{0, \dots, N-1\}, \quad a_{J,n} = \langle f, \varphi_{J,n}^P \rangle \quad (4.4)$$

for some function of interest f we are sampling. This is equivalent to saying that the discretization process have exactly access to $P_{V_J}f$. This hypothesis is questionable, and similar to the Shannon bandlimit assumption. In practice, the scaling functions $\varphi_{j,n}$ are often quite close to the point-spread function of the acquisition device, so it is acceptable. One can however improves this by correcting the acquired values by the device to be closer to assumption (4.4).

The discrete wavelet transform then computes, from this input a_J , all the coefficients

$$\forall j \in \{J+1, J+2, \dots, 0\}, \quad \forall n \in \llbracket 0, 2^{-j}-1 \rrbracket, \quad a_{j,n} \stackrel{\text{def}}{=} \langle f, \varphi_{j,n}^P \rangle, \quad \text{and} \quad d_{j,n} \stackrel{\text{def}}{=} \langle f, \psi_{j,n}^P \rangle$$

in this order (increasing values of j). During the algorithm, the previously computed vector a_j can be discarded, and only the d_j are kept.

The forward discrete wavelet transform on a bounded domain is thus the orthogonal finite dimensional map

$$a_J \in \mathbb{R}^N \longmapsto \{d_{j,n} ; 0 \leq j < J, 0 \leq n < 2^{-j}\} \cup \{a_0 \in \mathbb{R}\}.$$

The inverse transform, which is thus the adjoint due to orthogonality, is the inverse map.

Figure 4.5 shows examples of wavelet coefficients. For each scale 2^j , there are 2^{-j} coefficients.

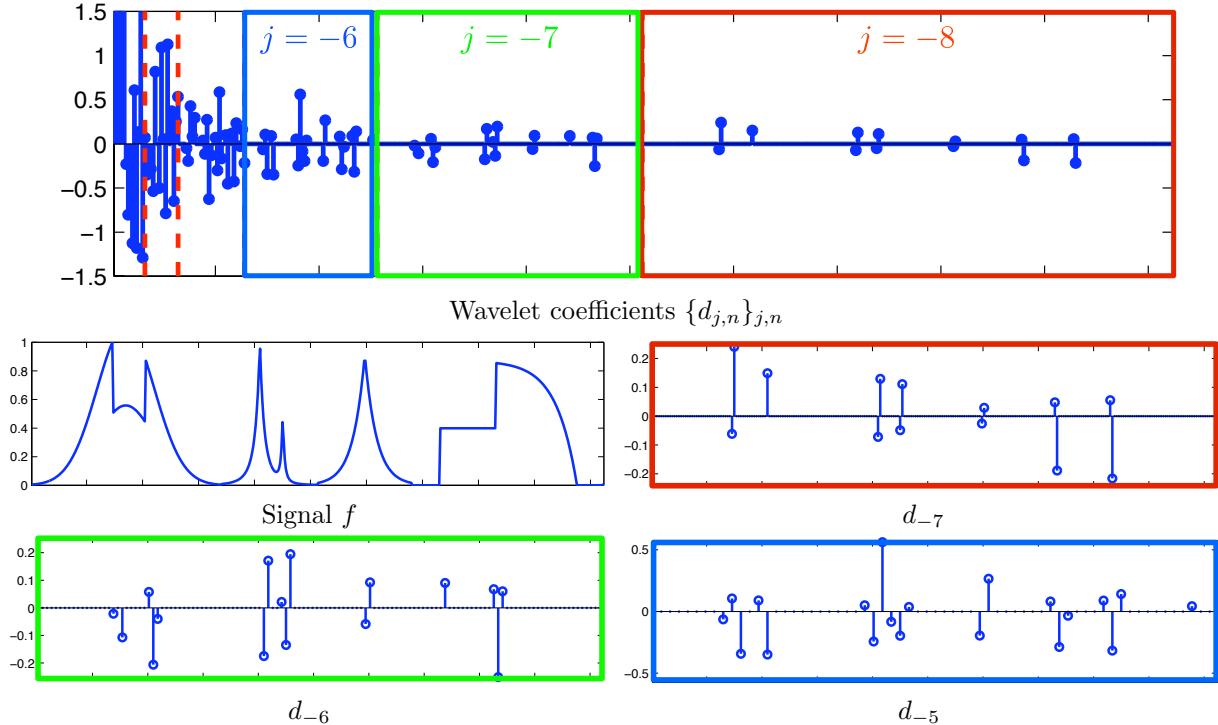


Figure 4.5: Wavelet coefficients. Top row: all the coefficients. Bottoms rows: zoom on the different scales

4.4.2 Forward Fast Wavelet Transform (FWT)

The algorithm proceeds by computing a series of simple operators

$$\forall j = J+1, \dots, 0, \quad (a_j, d_j) = \mathcal{W}_j(a_{j-1}) \quad \text{where} \quad \mathcal{W}_j : \mathbb{R}^{2^{-j+1}} \rightarrow \mathbb{R}^{2^{-j}} \times \mathbb{R}^{2^{-j}} \quad (4.5)$$

The number of such steps is thus $|J| = \log_2(N)$. Each \mathcal{W}_j is orthogonal since it corresponds to linear maps between coefficients in orthogonal bases.

In order to describe the algorithm that computes \mathcal{W}_j , we introduce the filters “filter” coefficients $f, g \in \mathbb{R}^{\mathbb{Z}}$

$$h_n \stackrel{\text{def.}}{=} \frac{1}{\sqrt{2}} \langle \varphi(\cdot/2), \varphi(\cdot - n) \rangle \quad \text{and} \quad g_n \stackrel{\text{def.}}{=} \frac{1}{\sqrt{2}} \langle \psi(\cdot/2), \varphi(\cdot - n) \rangle. \quad (4.6)$$

We denote as $\downarrow_2 : \mathbb{R}^K \rightarrow \mathbb{R}^{K/2}$ the subsampling operator by a factor of 2, i.e.

$$u \downarrow_2 \stackrel{\text{def.}}{=} (u_0, u_2, u_4, \dots, u_{K-2}, u_K).$$

In the following, we assume that these filters are decaying fast enough.

Proposition 10. *One has*

$$\mathcal{W}_j(a_j) = ((\bar{h} \star a_{j-1}) \downarrow_2, (\bar{g} \star a_{j-1}) \downarrow_2), \quad (4.7)$$

where \star denotes periodic convolutions on \mathbb{R}^{-j+1} .

Proof. We note that $\varphi(\cdot/2)$ and $\psi(\cdot/2)$ are in \mathcal{W}_j , so one has the decompositions

$$\frac{1}{\sqrt{2}} \varphi(t/2) = \sum_n h_n \varphi(t - n) \quad \text{and} \quad \frac{1}{\sqrt{2}} \psi(t/2) = \sum_n g_n \varphi(t - n) \quad (4.8)$$

Doing the change of variable $t \mapsto \frac{t-2^j p}{2^{j-1}}$ in (4.8), one obtains

$$\frac{1}{\sqrt{2}} \varphi\left(\frac{t-2^j p}{2^j}\right) = \sum_n h_n \varphi\left(\frac{t}{2^{j-1}} - (n + 2p)\right)$$

(similarly for ψ) and then doing the change $n \mapsto n - 2p$, one obtains

$$\varphi_{j,p} = \sum_{n \in \mathbb{Z}} h_{n-2p} \varphi_{j-1,n} \quad \text{and} \quad \psi_{j,p} = \sum_{n \in \mathbb{Z}} g_{n-2p} \psi_{j-1,n}.$$

When working with periodized function $(\varphi_{j,n}^P, \psi_{j,p}^P)$, this formula is still valid, but the summation over $n \in \mathbb{Z}$ should be done modulo 2^{-j+1} . Taking inner product of both size with respect to f (which is legit if h, g are decaying fast enough), one obtains the fundamental recursion fromula

$$a_{j,p} = \sum_{n \in \mathbb{Z}} h_{n-2p} a_{j-1,n} = (\bar{h} \star a_{j-1})_{2p} \quad \text{and} \quad d_{j,p} = \sum_{n \in \mathbb{Z}} g_{n-2p} a_{j-1,n} = (\bar{g} \star a_{j-1})_{2p} \quad (4.9)$$

where $\bar{u}_n \stackrel{\text{def.}}{=} u_{-n}$. One can show that this formula is still valid when working over a bounded interval $\mathbb{T} = \mathbb{R}/\mathbb{Z}$, but then \star denotes the periodic convolution over $\mathbb{Z}/2^{-j+1}\mathbb{Z}$. \square

Figure 4.6 shows two steps of application of these refinement relationships.

The FWT thus operates as follow:

- **Input:** signal $f \in \mathbb{C}^N$.
- **Initialization:** $a_J = f$.
- **For** $j = J, \dots, j_0 - 1$.

$$a_{j+1} = (a_j \star \tilde{h}) \downarrow 2 \quad \text{and} \quad d_{j+1} = (a_j \star \tilde{g}) \downarrow 2$$

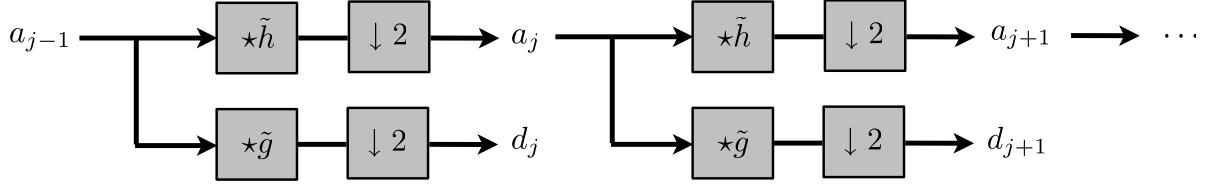


Figure 4.6: Forward filter bank decomposition.

– **Output:** the coefficients $\{d_j\}_{j_0 \leq j < J} \cup \{a_{j_0}\}$.

If $|h|, |g| \leq C$ so that both filters are compactly supported, then computing each \mathcal{W}_j is $(2C)2^{-j}$ operation, so that the complexity of the whole wavelet transform is

$$\sum_{j=J}^1 (2C)2^{-j} = (2C)2^{-J} = 2CN.$$

This shows that the fast wavelet transform is a linear time algorithm. Figure 4.7 shows the process of extracting iteratively the wavelet coefficients. Figure 4.8 shows an example of computation, where at each iteration, the coefficients of a_j and d_j are added to the left of the output vector.

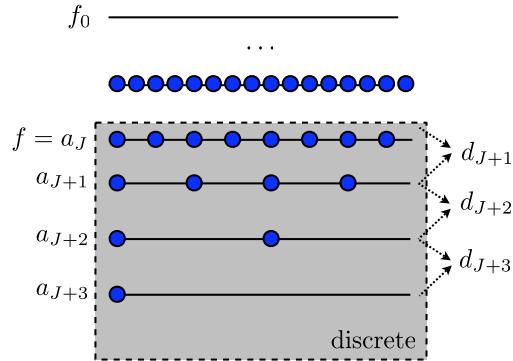


Figure 4.7: Pyramid computation of the coefficients.

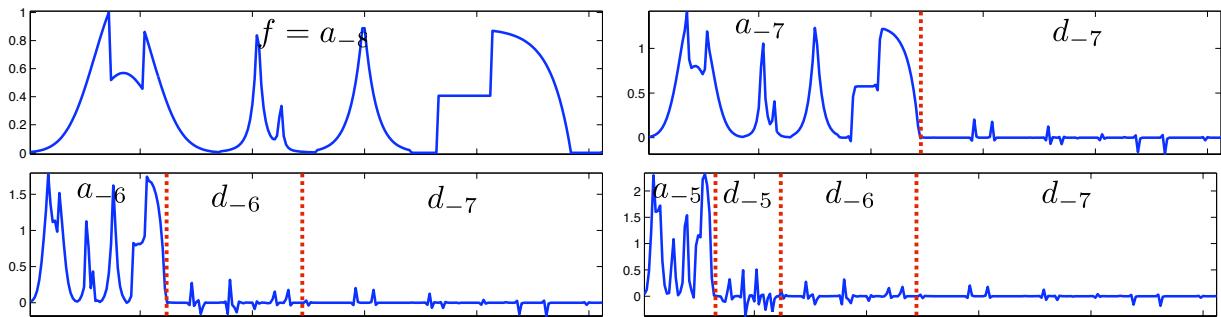


Figure 4.8: Wavelet decomposition algorithm.

Fast Haar transform. For the Haar wavelets, one has

$$\begin{aligned}\varphi_{j,n} &= \frac{1}{\sqrt{2}}(\varphi_{j-1,2n} + \varphi_{j-1,2n+1}), \\ \psi_{j,n} &= \frac{1}{\sqrt{2}}(\varphi_{j-1,2n} - \varphi_{j-1,2n+1}).\end{aligned}$$

This corresponds to the filters

$$\begin{aligned}h &= [\dots, 0, h[0] = \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0, \dots], \\ g &= [\dots, 0, g[0] = \frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}, 0, \dots].\end{aligned}$$

The Haar wavelet transform algorithm thus processes by iterating averaging and differences:

- **Input:** signal $f \in \mathbb{C}^N$.
- **Initialization:** $a_J = f$.
- **For** $j = J, \dots, j_0 - 1$.

$$a_{j+1,n} = \frac{1}{\sqrt{2}}(a_{j-1,2n} + a_{j-1,2n+1}) \quad \text{and} \quad d_{j+1,n} = \frac{1}{\sqrt{2}}(a_{j-1,2n} - a_{j-1,2n+1}).$$

- **Output:** the coefficients $\{d_j\}_{j_0 \leq j < J} \cup \{a_{j_0}\}$.

4.4.3 Inverse Fast Transform (iFWT)

The inverse algorithm proceeds by inverting each step (4.10)

$$\forall j = 0, \dots, J+1, \quad a_{j-1} = \mathcal{W}_j^{-1}(a_j, d_j), \quad (4.10)$$

where \mathcal{W}_j^* is the adjoint for the canonical inner product on $\mathbb{R}^{2^{-j+1}}$, i.e. when viewed as a matrix, the transpose.

We denote $\uparrow_2: \mathbb{R}^{K/2} \rightarrow \mathbb{R}^K$ the up-sampling operator

$$a \uparrow_2 = (a_0, 0, a_1, 0, \dots, 0, a_{K/2}, 0) \in \mathbb{R}^K.$$

Proposition 11. One has

$$\mathcal{W}_j^{-1}(a_j, d_j) = (a_j \uparrow_2) \star h + (d_j \uparrow_2) \star g.$$

Proof. Since \mathcal{W}_j is orthogonal, $\mathcal{W}_j^{-1} = \mathcal{W}_j^*$. We write the whole transform as

$$\mathcal{W}_j = S_2 \circ C_{\bar{h}, \bar{g}} \circ \mathcal{D} \quad \text{where} \quad \begin{cases} \mathcal{D}(a) = (a, a), \\ C_{\bar{h}, \bar{g}}(a, b) = (\bar{h} \star a, \bar{g} \star a), \\ S_2(a, b) = (a \downarrow_2, b \downarrow_2). \end{cases}$$

One has the following adjoint operator

$$\mathcal{D}^*(a, b) = a + b, \quad C_{\bar{h}, \bar{g}}^*(a, b) = (h \star a, g \star b), \quad \text{and} \quad S_2(a, b) = (a \uparrow_2, b \uparrow_2).$$

Indeed, let us check this for the convolution, assuming involved sequences are in ℓ_1 (they are actually finite sums when considering periodic signals),

$$\langle f \star \bar{h}, g \rangle = \sum_n (f \star \bar{h})_n g_n = \sum_n \sum_k f_k h_{k-n} g_n = \sum_k f_k \sum_n h_{k-n} g_n = \langle f, h \star g \rangle,$$

for the copying

$$\langle \mathcal{D}(a), (u, v) \rangle = \langle a, u \rangle + \langle b, v \rangle = \langle a, u + v \rangle = \langle a, \mathcal{D}^*(u, v) \rangle,$$

and for the down-sampling

$$\langle f \downarrow_2, g \rangle = \sum_n f_{2n}g_n = \sum_n (f_{2n}g_n + f_{2n+1}0) = \langle f, g \uparrow_2 \rangle.$$

This is shown using matrix notations in Figure 4.9. Putting everything together gives the desired formula. \square

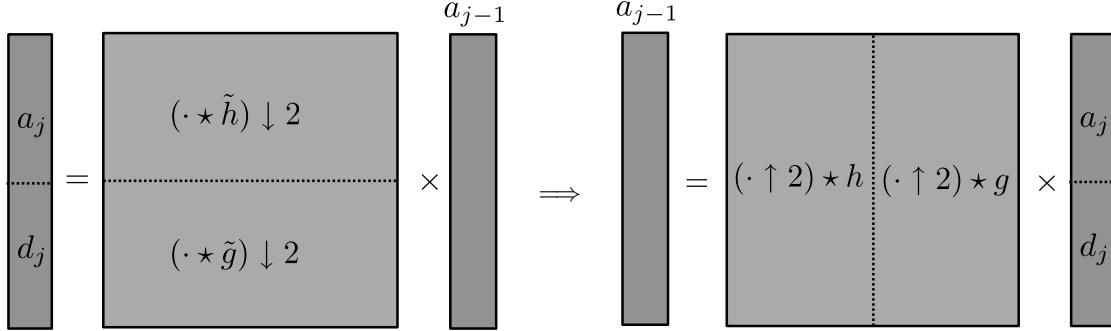


Figure 4.9: Wavelet inversion in matrix format.

The inverse Fast wavelet transform iteratively applies this elementary step

- **Input:** $\{d_j\}_{j_0 \leq j < J} \cup \{a_{j_0}\}$.
- **For** $j = j_0, \dots, J + 1$.

$$a_{j-1} = (a_j \uparrow 2) * h + (d_j \uparrow 2) * g.$$

- **Output:** $f = a_J$.

This process is shown using a block diagram in Figure 4.10, which is the inverse of the block diagram 4.6.

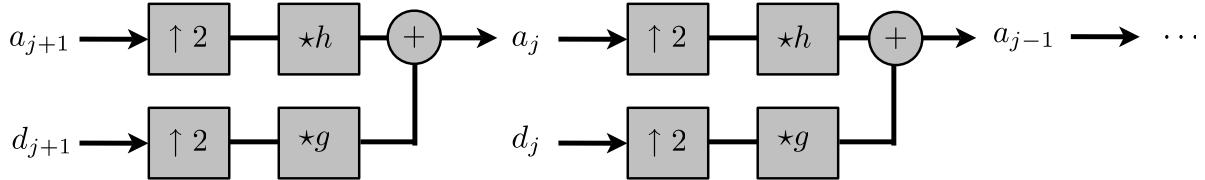


Figure 4.10: Backward filterbank recomposition algorithm.

4.5 2-D Wavelets

4.5.1 Anisotropic Wavelets

2-D anisotropic wavelets are defined using tensor product of Wavelet basis functions (2.12). The basis over \mathbb{T}^2 is thus of the form

$$\{\psi_{(j_1, j_2), (n_1, n_2)} ; (j_1, j_2) \in \mathbb{Z}^2, 0 \leq n_1 < 2^{-j_1}, 0 \leq n_2 < 2^{-j_2}\}$$

$$\text{where } \psi_{(j_1, j_2), (n_1, n_2)}(x_1, x_2) = \psi_{j_1, n_1}(x_1)\psi_{j_2, n_2}(x_2). \quad (4.11)$$

The computation of the fast anisotropic Wavelet transform in 2-D is similar to the 2-D FFT detailed in Section 2.5.2. Viewing the input image $a_J \in \mathbb{R}^{2^{-J} \times 2^{-J}}$ as a matrix, one first apply the 1-D FWT to each row, and then to each column, resulting in a linear time $O(N)$ algorithm, where $N = 2^{-2J}$.

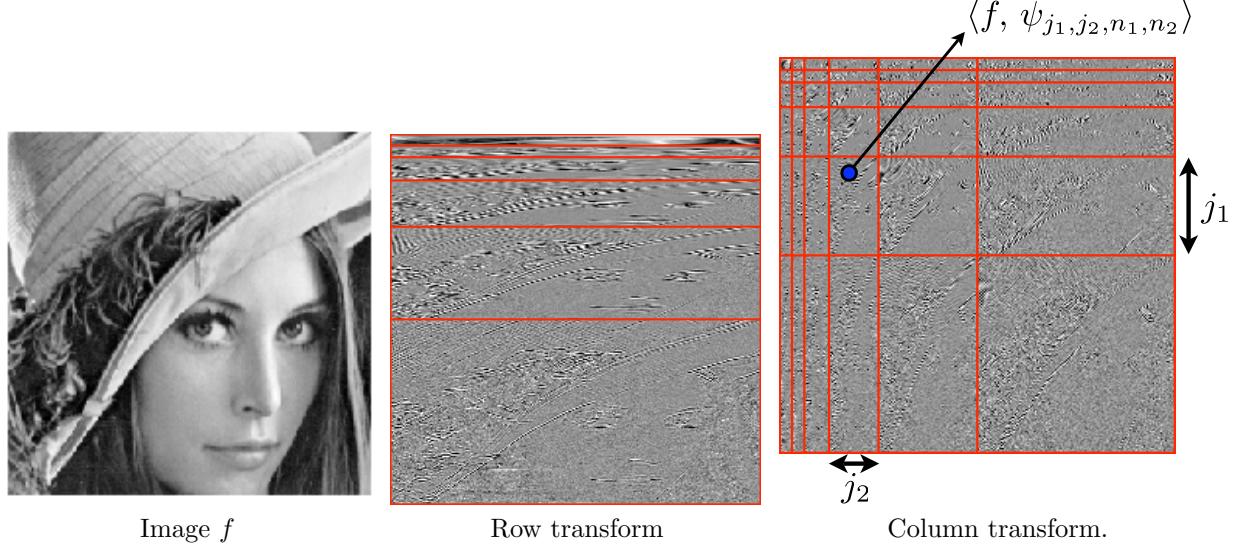


Figure 4.11: Steps of the anisotropic wavelet transform.

4.5.2 Isotropic Wavelets

A major issue with these anisotropic wavelet (4.11) is that a function $\psi_{(j_1, j_2), (n_1, n_2)}$ is scaled independently in each direction, leads to functions concentrated along an axis-oriented rectangle of size $2^{-j_1} \times 2^{j_2}$. This is not a very good features (since natural images usually do not exhibit such an anisotropy) and typically leads to visually unpleasant artifacts when used for processing (denoising or compression).

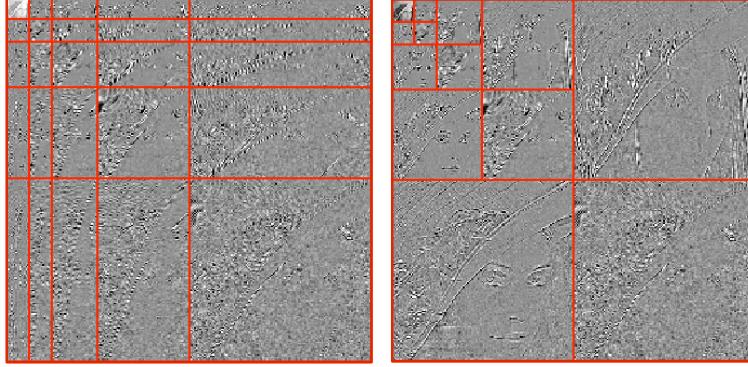


Figure 4.12: Anisotropic (left) versus isotropic (right) wavelet coefficients.

One rather use isotropic wavelet obtained by considering 2-D multi-resolution, obtained by tensor products of the 1-D approximation spaces

$$L^2(\mathbb{R}^2) \supset \dots \supset V^{j-1} \otimes V^{j-1} \supset V_j \otimes V_j \supset V_{j+1} \otimes V_{j+1} \supset \dots \supset \{0\}.$$

In the following, we denote

$$V_j^O \stackrel{\text{def.}}{=} V_j \otimes V_j$$

this isotropic 2-D multiresolution space.

Recall that the tensor product of two space $(V_1, V_2) \in L^2(\mathbb{R})^2$ is

$$V_1 \otimes V_2 = \text{Closure} (\text{Span} \{f_1(x_1)f_2(x_2) \in L^2(\mathbb{R}^2) ; f_1 \in V_1, f_2 \in V_2\}).$$

If $(\varphi_k^s)_k$ are Hilbertian bases for V_s , then one can show that $(\varphi_k^1(x_1)\varphi_k^2(x_2))_k$ is an Hilbertian basis for $V_1 \otimes V_2$.

One easily verify that one has the distributivity

$$(V_j \oplus^\perp W_J) \otimes (V_j \oplus^\perp W_J) = V_j^O \oplus^\perp W_j^V \oplus^\perp W_j^H \oplus^\perp W_j^D \quad \text{where} \quad \begin{cases} W_j^V \stackrel{\text{def.}}{=} (V_j \otimes W_j), \\ W_j^H \stackrel{\text{def.}}{=} (W_j \otimes V_j), \\ W_j^D \stackrel{\text{def.}}{=} (W_j \otimes W_j). \end{cases}$$

Here the letters $\{V, H, D\}$ stands for *Vertical*, *Horizontal*, *Diagonal* detail spaces. This leads to the following diagram of embedded spaces

$$\begin{array}{ccccccc} L^2(\mathbb{R}^2) & \xrightarrow{\quad} & \cdots & \xleftarrow{\quad} & V_{j-1} \otimes V_{j-1} & \xleftarrow{\quad} & V_j \otimes V_j \\ & & & & \searrow & & \searrow \\ & & & & W_{j-1}^{(2)} & & W_j^{(2)} \\ & & & & & & \searrow \\ & & & & & & W_{j+1}^{(2)} \end{array}$$

For $j \in \mathbb{Z}$, each of the three wavelet spaces is spanned with a wavelet, where basis elements are indexed by $n = (n_1, n_2) \in \mathbb{Z}$ (or in $\{0, \dots, 2^{-j} - 1\}^2$ on the interval \mathbb{T}),

$$\forall \omega \in \{V, H, D\}, \quad W_j^\omega = \text{Span}\{\psi_{j, n_1, n_2}^\omega\}_{n_1, n_2}$$

where

$$\forall \omega \in \{V, H, D\}, \quad \psi_{j, n_1, n_2}^\omega(x) = \frac{1}{2^j} \psi^\omega \left(\frac{x_1 - 2^j n_1}{2^j}, \frac{x_2 - 2^j n_2}{2^j} \right)$$

and where the three mother wavelets are

$$\psi^H(x) = \psi(x_1)\varphi(x_2), \quad \psi^V(x) = \varphi(x_1)\psi(x_2), \quad \text{and} \quad \psi^D(x) = \psi(x_1)\psi(x_2).$$

Figure 4.13 displays an examples of these wavelets.

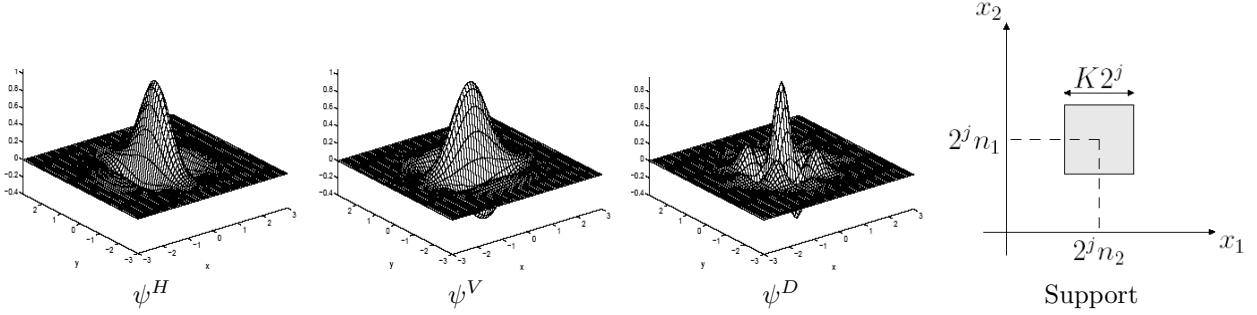


Figure 4.13: 2-D wavelets and their approximative support (right).

Haar 2-D multiresolution. For the Haar multiresolution, one obtains 2-D piecewise-constant Haar approximation. A function of $V_j \otimes V_j$ is constant on squares of size $2^j \times 2^j$. Figure 4.14 shows an example of projection of an image onto these 2-D Haar approximation spaces.



Figure 4.14: 2-D Haar approximation $P_{V_j^O} f$ for increasing j .

Discrete 2-D wavelet coefficients. Similarly to (4.4), we suppose that the sampling mechanism gives us access to inner product of the analog (continuous) signal f with the scaling function at scale $N = 2^{-J}$

$$\forall n \in \{0, \dots, N-1\}^2, \quad a_{J,n} = \langle f, \varphi_{J,n}^P \rangle$$

Discrete wavelet coefficients are defined as

$$\forall \omega \in \{V, H, D\}, \forall J < j \leq 0, \forall 0 \leq n_1, n_2 < 2^{-j}, \quad d_{j,n}^\omega = \langle f, \psi_{j,n}^\omega \rangle.$$

(we use here periodized wavelets). Approximation coefficients are defined as

$$a_{j,n} = \langle f_0, \varphi_{j,n}^O \rangle.$$

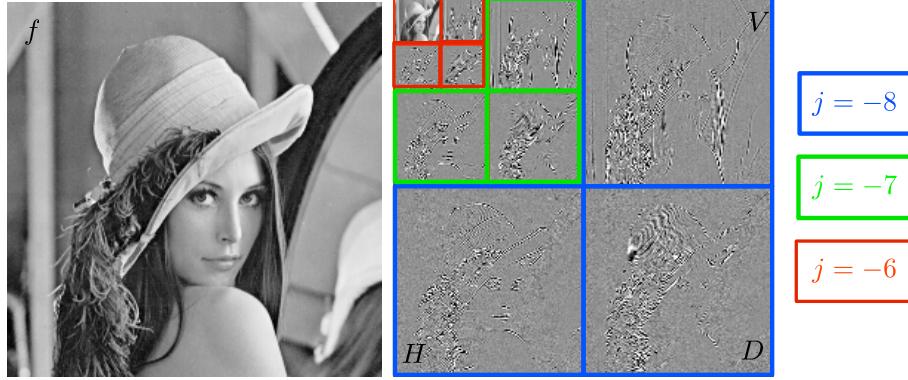


Figure 4.15: 2-D wavelet coefficients.

Figure 4.15 shows examples of wavelet coefficients, that are packed in an image of N pixels. Figure 4.16 shows other examples of wavelet decompositions.

Forward 2-D wavelet transform basic step. A basic step of the computation of the 2-D wavelet transform computes detail coefficients and a low pass residual from the fine scale coefficients

$$a_{j-1} \mapsto (a_j, d_j^H, d_j^V, d_j^D).$$

Similarly to the 1-D setting, this mapping is orthogonal, and is computed using the 1-D filtering and sub-sampling formula (4.7).

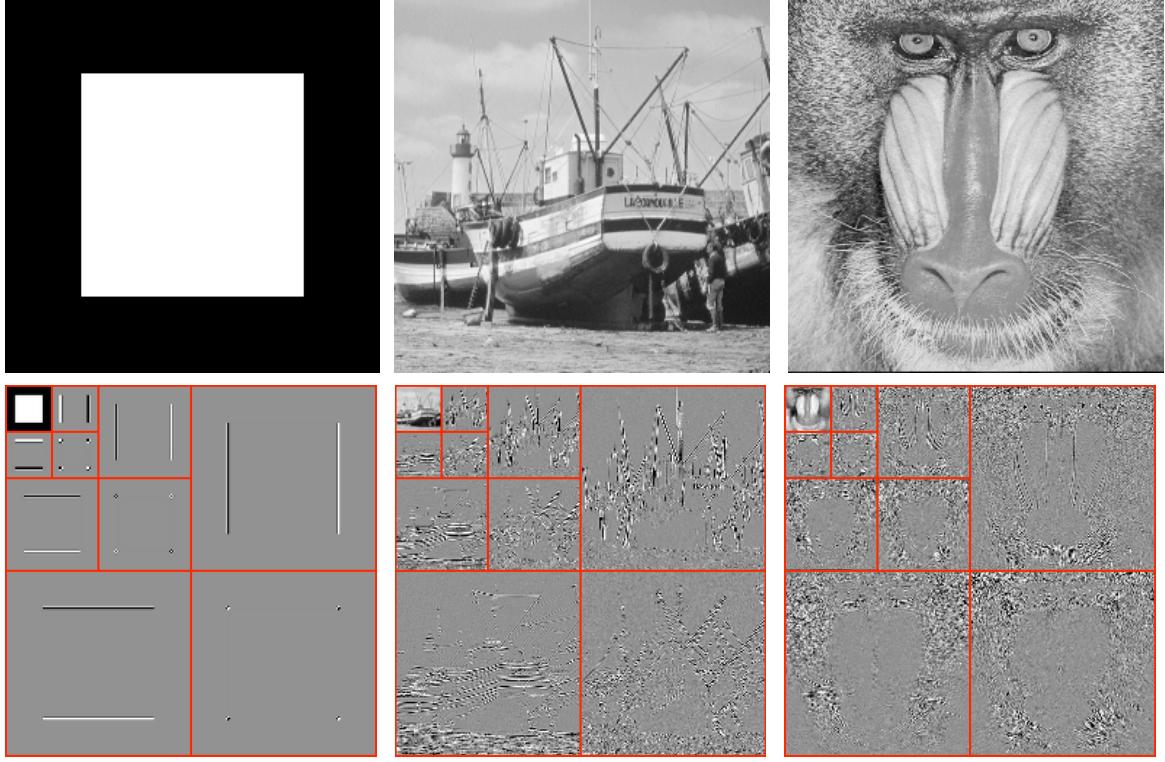


Figure 4.16: Examples of images (top row) and the corresponding wavelet coefficients (bottom row) .

One first applies 1-D horizontal filtering and sub-sampling

$$\begin{aligned}\tilde{a}_j &= (a_{j-1} \star^H \tilde{h}) \downarrow^H 2 \\ \tilde{d}_j &= (a_{j-1} \star^H \tilde{h}) \downarrow^H 2,\end{aligned}$$

where \star^H is the horizontal convolution, that applies the 1-D convolution to each column of a matrix

$$a \star^H b_{n_1, n_2} = \sum_{m_1=0}^{P-1} a_{n_1 - m_1, n_2} b_{m_1}$$

where $a \in \mathbb{C}^{P \times P}$ and $b \in \mathbb{C}^P$ are matrix and vectors. The notation $\downarrow^H 2$ accounts for sub-sampling in the horizontal direction

$$(a \downarrow^H 2)_{n_1, n_2} = a_{2n_1, n_2}.$$

One then applies 1-D vertical filtering and sub-sampling to \tilde{a}_j and \tilde{d}_j to obtain

$$\begin{aligned}a_j &= (\tilde{a}_j \star^V \tilde{h}) \downarrow^V 2, & d_j^H &= (\tilde{d}_j \star^V \tilde{h}) \downarrow^V 2, \\ d_j^V &= (\tilde{a}_j \star^V \tilde{g}) \downarrow^V 2, & d_j^D &= (\tilde{d}_j \star^V \tilde{g}) \downarrow^V 2,\end{aligned}$$

where the vertical operators are defined similarly to horizontal operators but operating on rows.

These two forward steps are shown in block diagram in Figure 4.17. These steps can be applied in place, so that the coefficients are stored in an image of N pixels, as shown in Figure 4.18. This gives the traditional display of wavelet coefficients used in Figure 4.16.

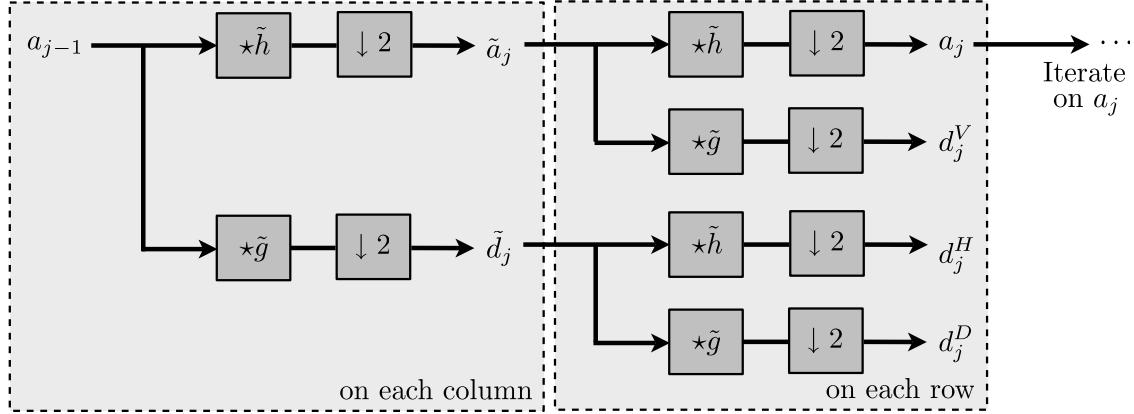


Figure 4.17: Forward 2-D filterbank step.

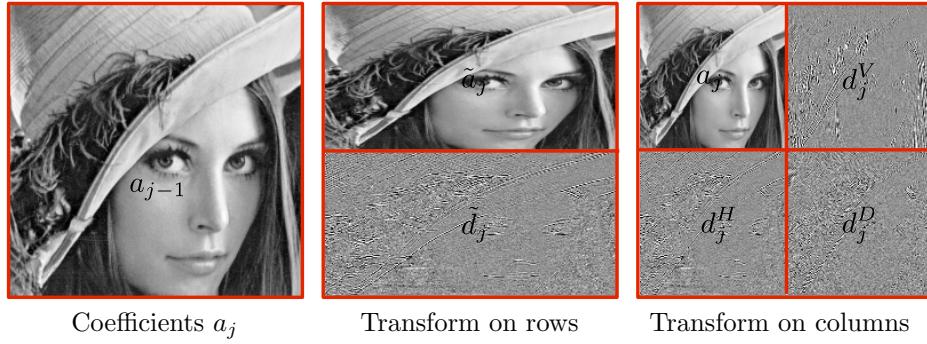


Figure 4.18: One step of the 2-D wavelet transform algorithm.

Fast 2-D wavelet transform. The 2-D FWT algorithm iterates these steps through the scales:

- **Input:** signal $f \in \mathbb{C}^N$.
- **Initialization:** $a_J = f$.
- **For** $j = J, \dots, j_0 - 1$.

$$\begin{aligned} \tilde{a}_j &= (a_{j-1} \star^H \tilde{h}) \downarrow^H 2, & d_j^V &= (\tilde{a}_j \star^V \tilde{g}) \downarrow^V 2, \\ \tilde{d}_j &= (a_{j-1} \star^H \tilde{h}) \downarrow^H 2, & d_j^H &= (\tilde{d}_j \star^V \tilde{h}) \downarrow^V 2, \\ a_j &= (\tilde{a}_j \star^V \tilde{h}) \downarrow^V 2, & d_j^D &= (\tilde{d}_j \star^V \tilde{g}) \downarrow^V 2. \end{aligned}$$

- **Output:** the coefficients $\{d_j^\omega\}_{j_0 \leq j < J, \omega} \cup \{a_{j_0}\}$.

Fast 2-D inverse wavelet transform. The inverse transform undo the horizontal and vertical filtering steps. The first step computes

$$\begin{aligned} \tilde{a}_j &= (a_j \star^V h) \uparrow^V 2 + (d_j^V \star^V g) \uparrow^V 2, \\ \tilde{d}_j &= (d_j^H \star^V h) \uparrow^V 2 + (d_j^D \star^V g) \uparrow^V 2, \end{aligned}$$

where the vertical up-sampling is

$$(a \uparrow^V 2)_{n_1, n_2} = \begin{cases} a_{k, n_2} & \text{if } n_1 = 2k, \\ 0 & \text{if } n_1 = 2k + 1. \end{cases}$$

The second inverse step computes

$$a_{j-1} = (\tilde{a}_j \star^H h) \uparrow^H 2 + (\tilde{d}_j \star^H g) \uparrow^H 2.$$

Figure 4.19 shows in block diagram this inverse filter banks, that is the inverse of the diagram 4.17.

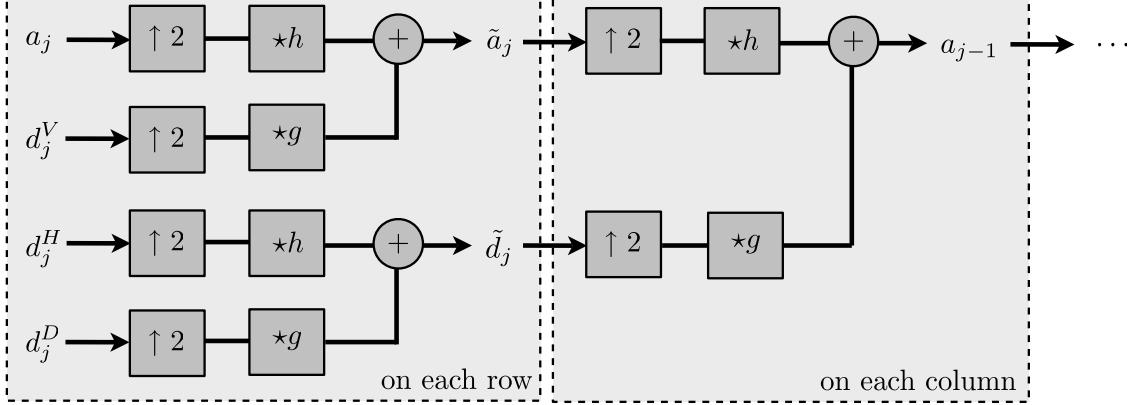


Figure 4.19: Backward 2-D filterbank step.

The inverse Fast wavelet transform iteratively applies these elementary steps

- **Input:** $\{d_j^\omega\}_{j_0 \leq j < J, \omega} \cup \{a_{j_0}\}$.
- **For** $j = j_0, \dots, J+1$.

$$\begin{aligned} \tilde{a}_j &= (a_j \star^V h) \uparrow^V 2 + (d_j^V \star^V g) \uparrow^V 2, \\ \tilde{d}_j &= (d_j^H \star^V h) \uparrow^V 2 + (d_j^D \star^V g) \uparrow^V 2, \\ a_{j-1} &= (\tilde{a}_j \star^H h) \uparrow^H 2 + (\tilde{d}_j \star^H g) \uparrow^H 2. \end{aligned}$$

- **Output:** $f = a_J$.

4.6 Wavelet Design

To be able to compute the wavelet coefficients using the FWT algorithm, it remains to know how to compute the scaling and wavelet functions. The FWT only makes use of the filters h and g , so instead of explicitly knowing the functions φ and ψ , one can only know these filters. Indeed, most of the known wavelets do not have explicit formula, and are implicitly defined through the cascade of the FWT algorithm.

This section shows what are the constraints h and g should satisfy, and gives practical examples. Furthermore, it shows that the knowledge of h determines g under the constraint of having quadrature filters, which is the most usual choice for wavelet analysis.

4.6.1 Low-pass Filter Constraints

We introduce the following three conditions on a filter h

$$\hat{h}(0) = \sqrt{2} \tag{C1}$$

$$|\hat{h}(\omega)|^2 + |\hat{h}(\omega + \pi)|^2 = 2, \tag{C2}$$

$$\inf_{\omega \in [-\pi/2, \pi/2]} |\hat{h}(\omega)| > 0. \tag{C*}$$

Here we are using the Fourier series associated to a filter $h \in \mathbb{R}^{\mathbb{Z}}$

$$\forall \omega \in \mathbb{R}/2\pi\mathbb{Z}, \quad \hat{h}(\omega) \stackrel{\text{def.}}{=} \sum_{n \in \mathbb{Z}} h_n e^{-in\omega}.$$

If $h \in \ell^1(\mathbb{Z})$, this defines a continuous periodic function $\hat{h} \in \mathcal{C}^0(\mathbb{R}/2\pi\mathbb{Z})$, and this definition can be extended to $h \in \ell^2(\mathbb{Z})$ and defines $\hat{h} \in L^2(\mathbb{R}/2\pi\mathbb{Z})$.

Theorem 23. *If φ defines multi-resolution approximation spaces, then (C_1) and (C_2) holds for h defined in (4.6). Conversely, if (C_1) , (C_2) and (C^*) holds, then there exists a φ defining multi-resolution approximation spaces so that associated filter is h as defined in (4.6).*

Proof. We only prove the first statement of the theorem. The converse statement is much more difficult to prove.

We now prove condition (C_1) . The refinement equation reads like a discrete-continuous convolution (or equivalently a convolution with a distribution)

$$\frac{1}{\sqrt{2}}\varphi\left(\frac{t}{2}\right) = \sum_{n \in \mathbb{Z}} h_n \varphi(t - n). \quad (4.12)$$

Denoting $h \star \varphi$ such a convolution, assuming $h \in \ell_1(\mathbb{Z})$ and $\varphi \in L^1(\mathbb{R})$, one check that one can apply Fubini and that $h \star \varphi \in L^1(\mathbb{R})$ and then

$$\begin{aligned} \mathcal{F}\left(\sum_{n \in \mathbb{Z}} h_n \varphi(t - n)\right)(\omega) &= \int_{\mathbb{R}} \sum_{n \in \mathbb{Z}} h_n \varphi(t - n) e^{-i\omega t} dt = \sum_{n \in \mathbb{Z}} h_n \int_{\mathbb{R}} \varphi(t - n) e^{-i\omega t} dt \\ &= \sum_{n \in \mathbb{Z}} h_n e^{-in\omega} \int_{\mathbb{R}} \varphi(x) e^{-i\omega x} dx = \hat{\varphi}(\omega) \hat{h}(\omega) \end{aligned}$$

where we made the change of variable $x = t - n$. Note that here, $h(\omega)$ is the 2π -periodic Fourier transform (i.e. Fourier series) of infinite filters defined in (4.6.1), whereas $\hat{\varphi}(\omega)$ is the Fourier transform of function. This is thus a product of a 2π -periodic function \hat{h} and a non-periodic function $\hat{\varphi}$. We recall that $\mathcal{F}(f(\cdot/s)) = s\hat{f}(s\cdot)$. Over the Fourier domain, equation (4.12) thus reads

$$\hat{\varphi}(2\omega) = \frac{1}{\sqrt{2}} \hat{h}(\omega) \hat{\varphi}(\omega). \quad (4.13)$$

One can show that $\hat{\varphi}(0) \neq 0$ (actually, $|\hat{\varphi}(0)| = 1$), so that this relation implies the first condition (C_1) .

We now prove condition (C_2) . The orthogonality of $\varphi(\cdot - n)\}_{n \in \mathbb{Z}}$ is rewritten using a continuous convolution as (see also Proposition 9)

$$\forall n \in \mathbb{Z}, \quad \varphi \star \bar{\varphi}(n) = \delta_0$$

where $\bar{\varphi}(x) = \varphi(-x)$, and thus over the Fourier domain, using (4.13) which shows $\hat{\varphi}(\omega) = \frac{1}{\sqrt{2}} \hat{h}(\omega/2) \hat{\varphi}(\omega/2)$

$$1 = \sum_k |\hat{\varphi}(\omega + 2k\pi)|^2 = \frac{1}{2} \sum_k |\hat{h}(\omega/2 + k\pi)|^2 |\hat{\varphi}(\omega/2 + k\pi)|^2.$$

Since \hat{h} is 2π -periodic, one can split even and odd k and obtain

$$2 = |h(\omega/2)|^2 \sum_k |\hat{\varphi}(\omega/2 + 2k\pi)|^2 + |h(\omega/2 + \pi)|^2 \sum_k |\hat{\varphi}(\omega/2 + 2k\pi + \pi)|^2$$

This leads to condition (C_2) . Re-using the fact that $\sum_k |\hat{\varphi}(\omega + 2k\pi)|^2 = 1$ for $\omega' = \omega/2$ in place of ω , one thus has

$$|h(\omega')|^2 + |h(\omega' + \pi)|^2 = 2.$$

We do not prove the converse statement, which requires to “create” a function φ from the filter h . The intuition is that iterating (4.13) leads informally to

$$\varphi(\omega) = \prod_{k<0} \frac{\hat{h}(\omega/2^k)}{\sqrt{2}}. \quad (4.14)$$

Condition (C^*) can be shown to imply that this infinite product converge, and define a (non-periodic) function in $L^2(\mathbb{R})$. \square

Note that for the converse statement of this theorem to holds, condition (C^*) imposes a control on the behavior of \hat{h} near 0.

4.6.2 High-pass Filter Constraints

We now introduce the following two conditions on a pair of filter (g, h)

$$|\hat{g}(\omega)|^2 + |\hat{g}(\omega + \pi)|^2 = 2 \quad (C_3)$$

$$\hat{g}(\omega)\hat{h}(\omega)^* + \hat{g}(\omega + \pi)\hat{h}(\omega + \pi)^* = 0. \quad (C_4)$$

Theorem 24. *If (φ, ψ) defines a multi-resolution analysis, then (C_3) and (C_4) holds for (h, g) defined in (4.6). Conversely, if (C_1) to (C_4) hold, then there exists a (φ, ψ) defining multi-resolution analysis so that associated filters are (h, g) as defined in (4.6). Furthermore,*

$$\hat{\psi}(\omega) = \frac{1}{\sqrt{2}}\hat{g}(\omega/2)\hat{\varphi}(\omega/2). \quad (4.15)$$

Proof. We prove condition (C_3) . The refinement equation for the wavelet reads

$$\frac{1}{\sqrt{2}}\psi\left(\frac{t}{2}\right) = \sum_{n \in \mathbb{Z}} g_n \varphi(t - n)$$

and thus over the Fourier domain

$$\hat{\psi}(2\omega) = \frac{1}{\sqrt{2}}\hat{g}(\omega)\hat{\varphi}(\omega). \quad (4.16)$$

The orthogonality of $\{\psi(\cdot - n)\}_n$ is re-written

$$\forall n \in \mathbb{Z}, \quad \psi \star \bar{\psi}(n) = \delta_0$$

and thus over the Fourier domain (using Poisson formula, see also Proposition 9)

$$\sum_k |\hat{\psi}(\omega + 2k\pi)|^2 = 1.$$

Using the Fourier domain refinement equation (4.16), similarely to the proof of Theorem 23 for (C_1) , this is equivalent to condition (C_3) . Figure 4.20 shows the Fourier transform of two filters that satisfy this complementary condition.

We now prove condition (C_4) . The orthogonality between $\{\psi(\cdot - n)\}_n$ and $\{\varphi(\cdot - n)\}_n$ is written as

$$\forall n \in \mathbb{Z}, \quad \psi \star \bar{\varphi}(n) = 0$$

and hence over the Fourier domain (using Poisson formula, similarly to Proposition 9)

$$\sum_k \hat{\psi}(\omega + 2k\pi)\hat{\varphi}^*(\omega + 2k\pi) = 0.$$

Using the Fourier domain refinement equations (4.13) and (4.16), this is equivalent to condition (C_4) . \square

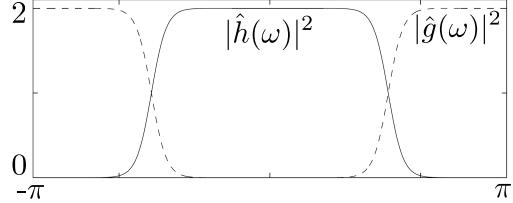


Figure 4.20: Complementarity between a low pass and a high pass wavelet filters h and g that satisfy condition (C_3) .

Quadrature mirror filters. Quadrature mirror filters (QMF) defines g as a function of h so that the conditions of Theorem 24 are automatically satisfied. This choice is the natural choice to build wavelet filters, and is implicitly assumed in most constructions (other choices leading to the same wavelet function anyway, since it satisfies (4.15)).

Proposition 12. *For a filter $h \in \ell^2(\mathbb{Z})$ satisfying (C_1) , defining $g \in \ell^2(\mathbb{Z})$ as*

$$\forall n \in \mathbb{Z}, \quad g_n = (-1)^{1-n} h_{1-n} \quad (4.17)$$

satisfies conditions (C_3) and (C_4) .

Proof. One indeed has that

$$\hat{g}(\omega) = e^{-i\omega} \hat{h}(\omega + \pi)^*, \quad (4.18)$$

so that

$$|\hat{g}(\omega)|^2 + |\hat{g}(\omega + \pi)|^2 = |e^{-i\omega} \hat{h}(\omega + \pi)^*|^2 + |e^{-i(\omega+\pi)} \hat{h}(\omega + 2\pi)^*|^2 = |\hat{h}(\omega + \pi)|^2 + |\hat{h}(\omega + 2\pi)|^2 = 2.$$

where we used the fact that \hat{h} is 2π -periodic, and also

$$\begin{aligned} \hat{g}(\omega) \hat{h}(\omega)^* + \hat{g}(\omega + \pi) \hat{h}(\omega + \pi)^* &= e^{-i\omega} \hat{h}(\omega + \pi)^* \hat{h}(\omega)^* + e^{-i(\omega+\pi)} \hat{h}(\omega + 2\pi)^* \hat{h}(\omega + \pi)^* \\ &= (e^{-i\omega} + e^{-i(\omega+\pi)}) \hat{h}(\omega + \pi)^* \hat{h}(\omega)^* = 0. \end{aligned}$$

□

4.6.3 Wavelet Design Constraints

According to the previous sections, the construction of a multi-resolution analysis (i.e. of functions (φ, ψ)) is obtained by designing a filter h satisfying conditions (C_1) and (C_2) . The function φ is obtained by an infinite cascade of filtering, or equivalently in the Fourier domain by (4.14), there is in general (put aside special case such as the Haar multiresolution) no closed form expression for φ . Once φ is defined, ψ is automatically defined by the relation (4.15) (and g can be defined as (4.18)).

There exists only one Fourier transform, but there is a large choice of different mother wavelet functions ψ . They are characterized by

- Size of the support.
- Number of oscillations (the so called number p of vanishing moments).
- Symmetry (only possible for non-orthogonal bases).
- Smoothness (number of derivatives).

We now detail how these constraints are integrated together with conditions (C_1) - (C_4) .

Vanishing moments. A wavelet ψ has p vanishing moments if

$$\forall k \leq p-1, \quad \int_{\mathbb{R}} \psi(x) x^k dx = 0. \quad (4.19)$$

This ensures that $\langle f, \psi_{j,n} \rangle$ is small if f is C^α , $\alpha < p$ on $\text{Supp}(\psi_{j,n})$.

This condition can be equivalently expressed over Fourier as followed.

Proposition 13. Assuming enough regularity of ψ , and using the QMF construction (4.18), it has p vanishing moments if and only if

$$\forall k \leq p-1, \quad \frac{d^k \hat{h}}{d\omega^k}(\pi) = \frac{d^k \hat{g}}{d\omega^k}(0) = 0. \quad (4.20)$$

Proof. Since ψ is regular, one has that $\hat{\psi}^{(k)} = \mathcal{F}((-i\cdot)^k \psi(\cdot))$, so that

$$(-i)^k \int_{\mathbb{R}} x^k \psi(x) dx = \hat{\psi}^{(k)}(0).$$

Relation (4.15) and (4.18) implies

$$\hat{\psi}(2\omega) = \hat{h}(\omega + \pi)^* \rho(\omega) \quad \text{where} \quad \rho(\omega) \stackrel{\text{def.}}{=} \frac{1}{\sqrt{2}} e^{-i\omega} \hat{\varphi}(\omega).$$

and differentiating this relation shows

$$2\hat{\psi}^{(1)}(2\omega) = \hat{h}(\omega + \pi)^* \rho^{(1)}(\omega) + \hat{h}^{(1)}(\omega + \pi)^* \rho(\omega)$$

which shows that, since $\rho(0) = \hat{\varphi}(0) \neq 0$, $\psi^{(1)}(0) = 0 \Leftrightarrow \hat{h}^{(1)}(\pi)^* = 0$. Recursing this argument and iterating the derivative, one obtains that $\psi^{(k)}(0) = 0 \Leftrightarrow \hat{h}^{(k)}(\pi)^* = 0$ (assuming this hold for previous derivatives). \square

Note that conditions (C₁) and (C₂) implies that $\hat{h}(\pi) = 0$, so that an admissible wavelet necessarily has 1 vanishing moment, i.e. $\int \psi = 0$. Condition (4.20) shows that having more vanishing moment is equivalent to having a Fourier transform \hat{h} which is “flatter” arround $\omega = \pi$.

Support. Figure 4.21 shows the wavelet coefficients of a piecewise smooth signal. Coefficients of large magnitude are clustered near the singularities, because the wavelet ψ has enough vanishing moments.

To avoid that many wavelets create large coefficients near singularities, one should choose ψ with a small support. One can show that the size of the support of ψ is proportional to the size of the support of h . This requirement is however contradictory with the vanishing moment property (4.19). Indeed, one can prove that for an orthogonal wavelet basis with p vanishing moments

$$|\text{Supp}(\psi)| \geq 2p - 1,$$

where $\text{sup}(a)$ is the largest closed interval outside of which the function f is zero.

Chapter 6 studies in details the tradeoff of support size and vanishing moment to perform non-linear approximation of piecewise smooth signals.

Smoothness. In compression or denoising applications, an approximate signals is recovered from a partial set I_M of coefficients,

$$f_M = \sum_{(j,n) \in I_M} \langle f, \psi_{j,n} \rangle \psi_{j,n}.$$

This approximation f_M has the same smoothness as ψ .

To avoid visually unpleasant artifacts, one should thus choose a smooth wavelet function ψ . This is only for cosmetic reasons, since increasing smoothness does not leads to a better approximation. However, for most wavelet family, increasing the number of vanishing moments also increases the smoothness of the wavelets. This is for instance the case of the Daubechies family exposed in the next section.

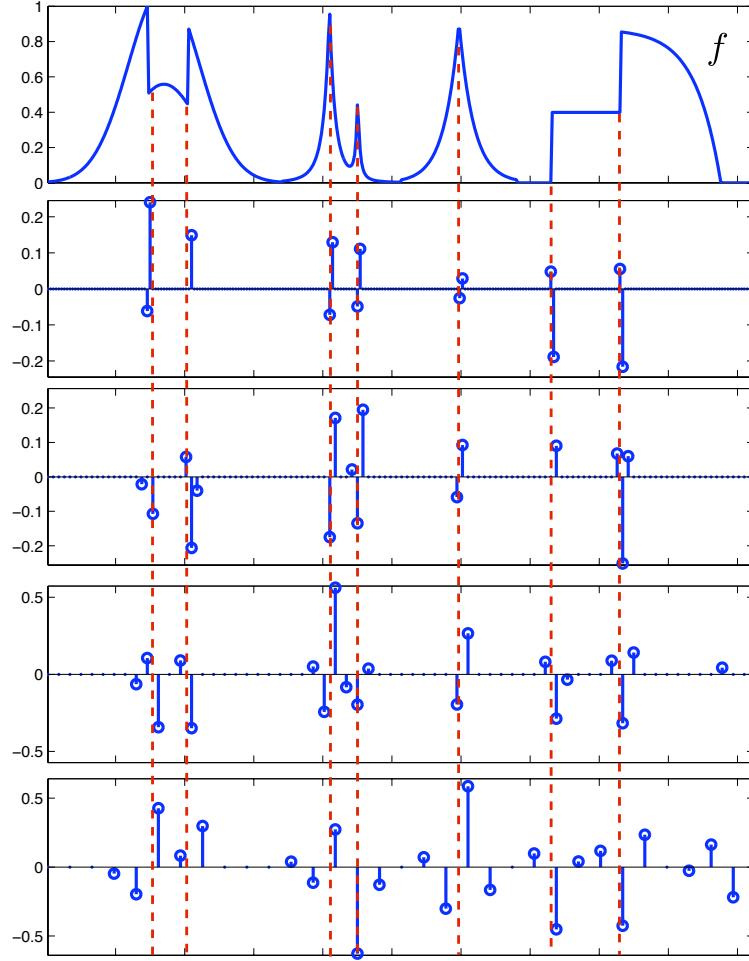


Figure 4.21: Location of large wavelet coefficients.

4.6.4 Daubechies Wavelets

To build a wavelet ψ with a fixed number p of vanishing moments, one designs the filter h , and use the quadrature mirror filter relation (4.18) to compute g . One thus look for h such that

$$|\hat{h}(\omega)|^2 + |\hat{h}(\omega + \pi)|^2 = 2, \quad \hat{h}(0) = \sqrt{2}, \quad \text{and} \quad \forall k < p, \frac{d^k \hat{h}}{d\omega^k}(\pi) = 0.$$

This corresponds to algebraic relationships between the coefficients of h , and it turns out that they can be solved explicitly using the Euclidean division algorithm for polynomials.

This leads to Daubechies wavelets with p vanishing moments, which are orthogonal wavelets with a minimum support length of $2p - 1$.

For $p = 1$, it leads to the Haar wavelet, with

$$h = [h_0 = 0.7071; 0.7071].$$

For $p = 2$, one obtains the celebrated Daubechies 4 filter

$$h = [0.4830; h_0 = 0.8365; 0.2241; -0.1294],$$

and for $p = 3$,

$$h = [0; 0.3327; 0.8069; h_0 = 0.4599; -0.1350; -0.0854; 0.0352].$$

Wavelet display. Figure 4.22 shows examples of Daubechies mother wavelet functions with an increasing number of vanishing moments. These displays are obtained by computing in fact a discrete wavelet $\bar{\psi}_{j,n}$ defined in (??) for a very large number of samples N . This discrete wavelet is computed by applying the inverse wavelet transform to the coefficients $d_{j',n'} = \delta_{j-j'}\delta_{n-n'}$.

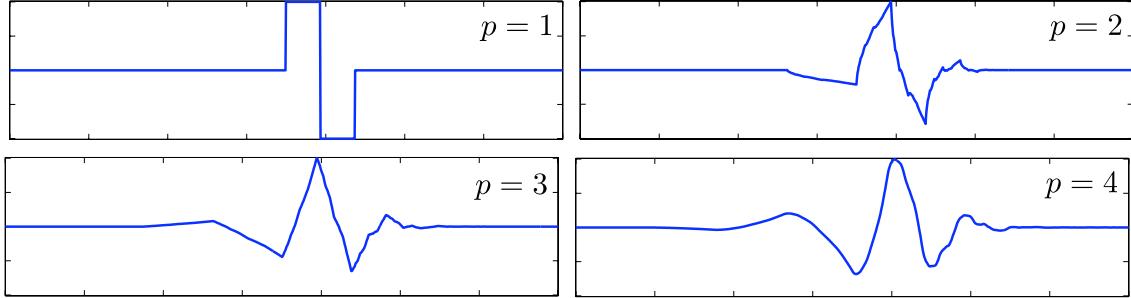


Figure 4.22: Examples of Daubechies mother wavelets ψ with an increasing number p of vanishing moments.

Chapter 5

Multiresolution Mesh Processing

This chapter shows how computations on a mesh can be performed in a multiscale manner, by considering meshes of increasing resolutions. This leads to the notion of subdivision surfaces and wavelet transform, which are two different tools to interpolate and decompose functions on meshes. Both methods rely on a special kind of meshes whose triangulations can be obtained by applying a regular refinement rule.

5.1 Semi-regular Meshes

5.1.1 Nested Multiscale Grids.

In order to perform multiscale mesh processing, one needs to pack the vertices V of a topological mesh $M = (V, E, F)$ in sets of increasing resolution. As explained in section 3.1.2, it is important to remember that this construction is purely combinatorial, in that no geometrical information (such as actual positions of the vertices in \mathbb{R}^3) is required to build the set of multi-resolution meshes. In fact these multiscale grids can be used to actually process the geometrical realization \mathcal{M} of the mesh M as three real valued functions (the three coordinates of the points).

We thus consider a set of nested indexes

$$V_0 \subset V_{-1} \subset \dots \subset V_L = V$$

which are split according to

$$V_j = V_{j+1} \cup H_{j+1}.$$

Next section describes how to actually compute this set of nested grids using a triangular split, but most of the mathematical tools are in fact valid for arbitrary set of indices, as long as they are embedded in one each other through scales.

For mesh processing, an index $\ell \in V_j$ corresponds to a vertex $x_\ell \in \mathcal{V} \subset \mathbb{R}^3$. The signals to be processed are vectors $f \in \mathbb{R}^n$ of size $n = |V_L|$ defined on the grid V_L . We sometimes write $f \in \ell^2(V_L)$ instead of $f \in \mathbb{R}^n$ to emphasize the domain on which f is indexed. This chapter describes transforms for signals $f \in \ell^2(V_L)$ sampled on the finest grid V_L .

5.1.2 Semi-regular Triangulation.

The combinatorial structure of a triangular mesh is defined in section 3.1.2. This chapter considers only a certain class of meshes $M = (V, E, F)$ that can be obtained by a regular split of faces, starting from an initial coarse triangulation. This splitting leads to a set of multiresolution meshes $M_j = (V_j, E_j, F_j)$ for $J \leq j \leq 0$, where the full mesh is $M_J = M$.

Starting from this coarse triangulation, one defines by subdivision a multiscale triangulation $(V_j, E_j, F_j)_{L \leq j \leq 0}$ where

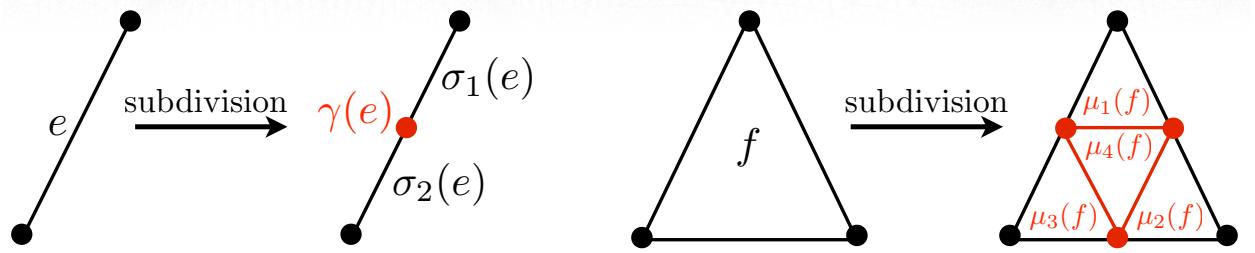


Figure 5.1: Edge-splitting subdivision.

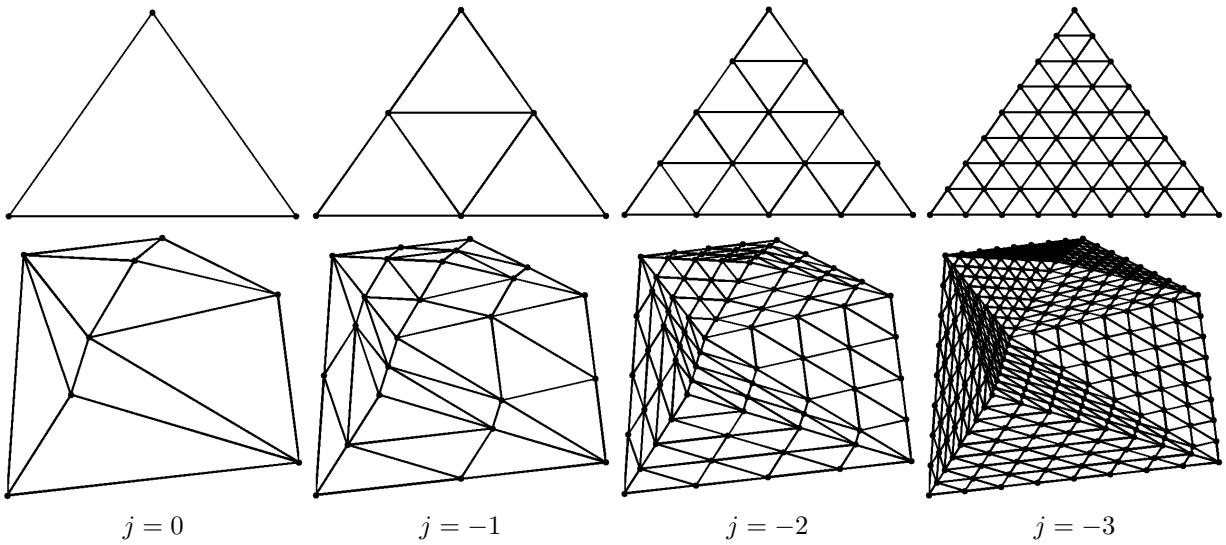


Figure 5.2: Regular subdivision 1:4 of a single triangle. Regular subdivision of a planar triangulation M_0 .

- For each edge $e \in E_j$, a central index $\gamma(e) \in V_{j-1}$ is added to the vertices

$$V_{j-1} = V_j \cup \{\gamma(e) ; e \in E_j\}.$$

- Each edge is subdivided into two finer edges

$$\forall e = (a, b) \in E_j, \quad \sigma_1(e) = (a, \gamma(e)) \quad \text{and} \quad \sigma_2(e) = (b, \gamma(e)).$$

The subdivided set of edges is then

$$E_{j-1} = \{\sigma_i(e) ; i = 1, 2 \quad \text{and} \quad e \in E_j\}.$$

- Each face $f = (a, b, c) \in F_j$ is subdivided into four faces

$$\begin{cases} \mu_1(f) = (a, \gamma(a, b), \gamma(a, c)), \mu_2(f) = (b, \gamma(b, a), \gamma(b, c)), \\ \mu_3(f) = (c, \gamma(c, a), \gamma(c, b)), \mu_4(f) = (\gamma(a, b), \gamma(b, c), \gamma(c, a)). \end{cases}$$

The subdivided set of faces is then

$$F_{j-1} = \{\mu_i(f) ; i = 1, 2, 3, 4 \quad \text{and} \quad f \in F_j\}.$$

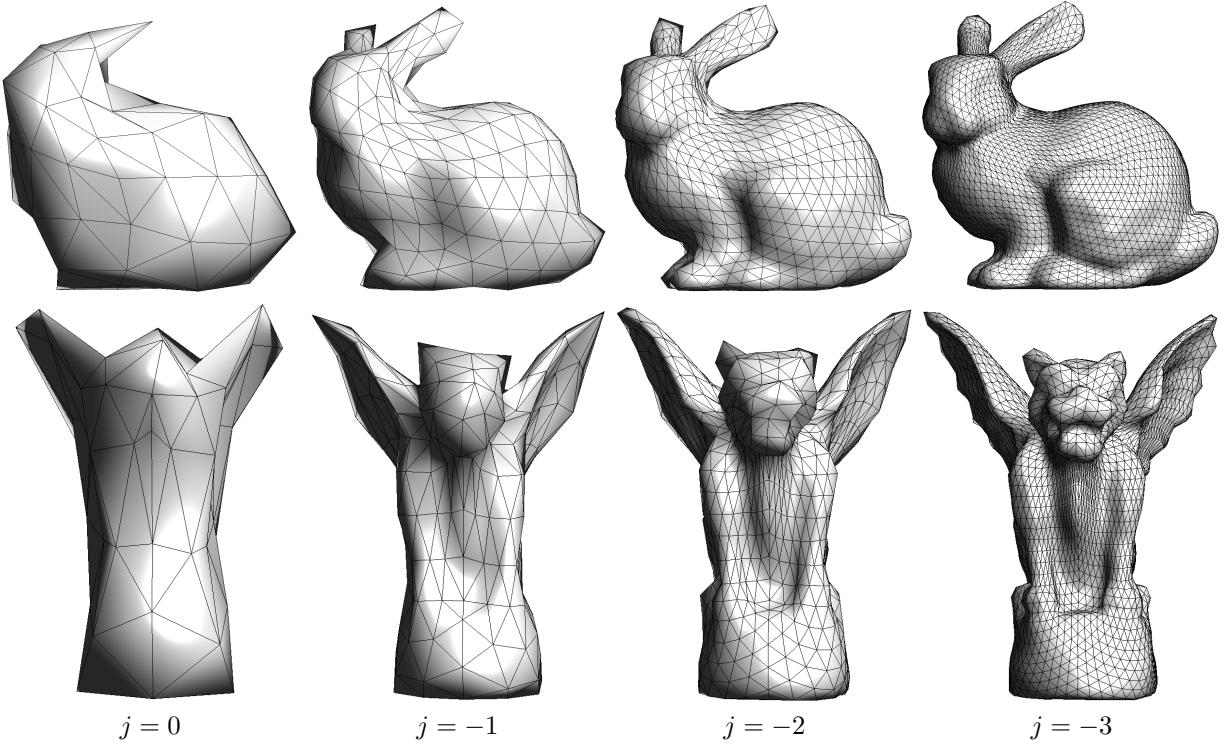


Figure 5.3: Examples of semi-regular meshes $(V_j)_j$ for increasing scale j (from left to right).

Figure 5.1 shows the notations related to the subdivision process. Figure 5.2 shows an example of recursive splitting of a triangle and a coarse triangulation. Figure 5.3 shows examples of semi-regular triangulation using a geometric realization (position of the vertices) to create a 3D surface.

The set of vertices can be classified as

- **Regular vertices** are those who belong neither to the coarse mesh V_0 nor to a boundary of a mesh M_j . These vertices have always 6 neighbors.
- **Extraordinary vertices** are the initial vertices of V_0 . They exhibit arbitrary connectivity.
- **Boundary vertices** are those belonging to a mesh boundary. Boundary vertices not in V_0 always have 4 immediate neighbors.

Obviously not every meshes can be obtained from such a subdivision process. In practice, an arbitrary mesh, obtained from CAD design or range scanning usually does not have any multiscale structure. It is thus necessary to remesh it in order to modify the connectivity of the mesh. During this process, the position of the vertices in \mathbb{R}^3 is modified in order for the geometrical realization to stay close from the original piecewise linear surface. One can see [2] for a survey of various semi-regular remeshing methods.

5.1.3 Spherical Geometry Images

Starting from some input surfaces $\mathcal{S} \subset \mathbb{R}^3$, one typically wants to compute a semi-regular meshes $(M_j)_{j \geq L}$ that approximate \mathcal{S} . In most case, the surface \mathcal{S} is actually given as an arbitrary triangulated mesh and this process corresponds to a semi-regular remeshing. Many algorithm have been devised for surface remeshing and we describe here a method [33] that works for surfaces that have the topology of a sphere. It means that the surface has genus 0, without boundary and without handles.

This methods works by computing several intermediate surface-wise parameterization.

- *Spherical parameterization:* each points of the original triangulation of \mathcal{S} is mapped onto the unit sphere.

This creates a bijective parameterization

$$\varphi_S : S^2 \rightarrow \mathcal{S}.$$

This is a non-linear process that differs from the planar parameterization introduced in section 3.3.7. We do not give the details of such a process, but it requires minimizing the smoothness of the mapping φ_S^{-1} under the constraint that it maps points of \mathcal{S} to unit length vectors (point on the sphere S^2). The algorithm is explained in details in [33].

- *Spherical-tetraedron flattening*: one flattens each quadrant (1/8) of the sphere in order to have a mapping

$$\varphi_T : \text{Octaederon} \rightarrow S^2.$$

One can use for instance a mapping between spherical barycentric coordinate on each quadrant and Euclidean barycentric coordinates on each face of the octahedron.

- *Tetraedron unfolding*: One maps each equilateral face of the octahedron on a rectangular triangle that corresponds to 1/8th of the square $[0, 1]^2$

$$\varphi_U : [0, 1]^2 \rightarrow \text{Octaederon}.$$

- *Regular sampling*: the geometry image is obtained by regularly sampling the square on a uniform grid

$$x_\ell = \varphi_S \circ \varphi_T \circ \varphi_U(\ell/n) \quad \text{for } \ell_i = 0, \dots, n - 1.$$

The mapping $\ell \mapsto x_\ell \in \mathbb{R}^3$ is the geometry image, which can be stored as a 3-channel (color) image.

From such a geometry image x_ℓ , one can easily compute a semi-regular mesh by simply performing a regular 1:4 subdivision of the octahedron. Figure 5.4 shows the steps of the construction of a geometry image, and the resulting semi-regular mesh.

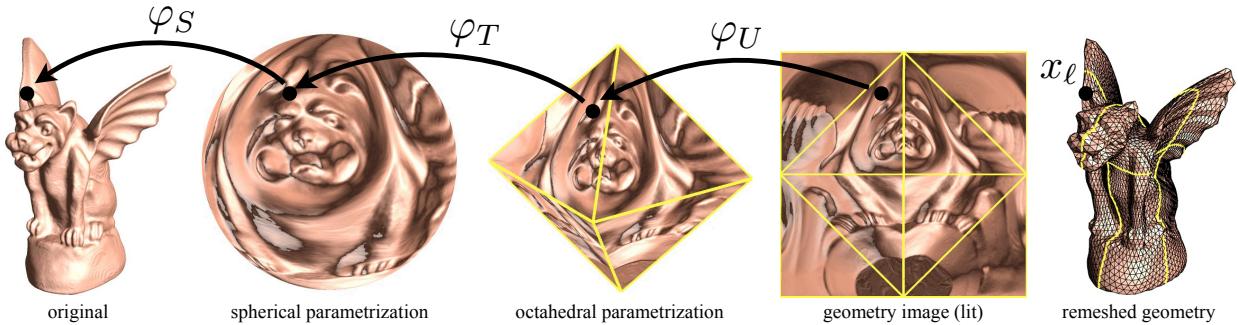


Figure 5.4: Spherical geometry image construction, taken from [33].

5.2 Subdivision Curves

Before getting into the detail of subdivision surfaces, we describe the subdivision process in the simpler setting of 1D signals. This leads to the construction of subdivision of 1D functions and subdivision curves.

In this 1D setting, the grid point indexes are dyadic sub-grids of \mathbb{Z}

$$\forall j \geq L, \quad V_j = \{\ell 2^{j-L} ; 0 \leq \ell < s_0 2^{-j}\},$$

where $s_0 = |V_0|$ is the size of the initial vector f_0 to be subdivided.

Each subdivision step computes, from a set $f_j(\ell) \in \ell^2(V_j)$ of coarse values, a refined vector $f_{j-1} \in \ell^2(V_{j-1})$ defined by

$$\begin{cases} \forall k \in H_j, f_{j-1}(k) = \sum_t f_j((k-1)/2 + t)h(t), \\ \forall \ell \in V_j, f_{j-1}(\ell) = \sum_t f_j(\ell + t)\tilde{h}(t). \end{cases}$$

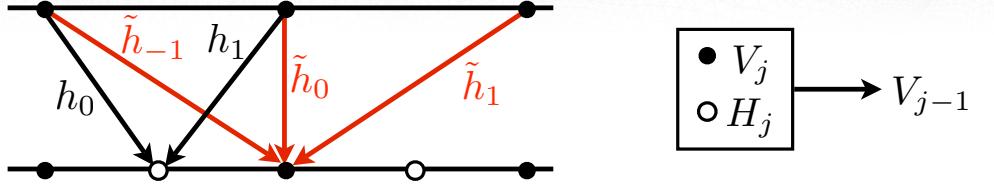


Figure 5.5: 1D subdivision scheme with filters h and \tilde{h} . The red curve represent the original signal f^0 .

where the set of weights h and \tilde{h} acts as local averaging operators. This averaging should be corrected at the boundary, and we use here cyclic boundary conditions which identifies 0 and $s_0 2^{-j}$ in V_j . Figure 5.5 shows a graphical display of these averaging operators.

One can write this subdivision steps as convolution by introducing the global set of weights

$$g = [\dots, \tilde{h}(-1), h(0), \tilde{h}(0), h(1), \tilde{h}(1), \dots]$$

since one has

$$f_{j-1} = (f_j \uparrow 2) * g \quad \text{where } a \uparrow 2 = [\dots, 0, a(-1), 0, a(0), 0, a(1), 0, \dots].$$

This corresponds to the traditional description of the wavelet low-pass filtering [27].

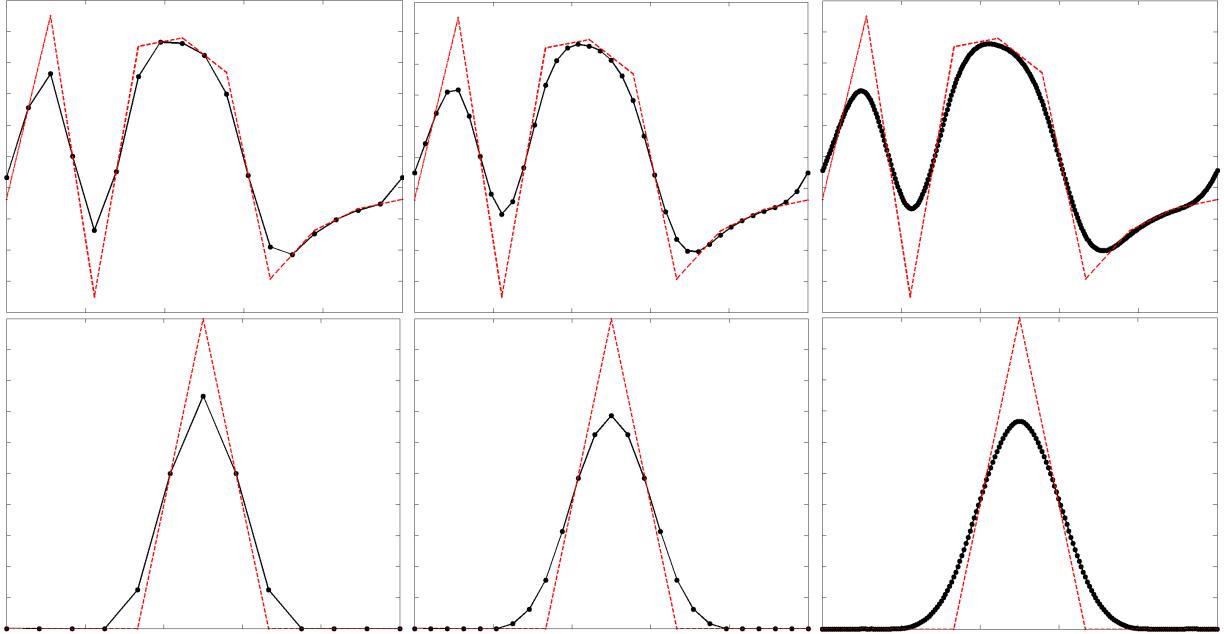


Figure 5.6: 1D subdivision of a signal. Bottom row shows the subdivision from an impulse signal, converging to the scaling function φ .

Figure 5.6 shows several steps of subdivision, starting from an initial vector of size $|V_0| = 10$.

One can apply this subdivision of functions to a pair of signals

$$(X_0, Y_0) : V_0 \rightarrow \mathbb{R}^2$$

which is a control polygon composed of points located in the plane. The subdivision curve converges to the limiting curve

$$(X_j, Y_j) \xrightarrow{j \rightarrow -\infty} (X(t), Y(t))_{t=0}^1 \subset \mathbb{R}^2.$$

An interesting property is that this curve is included in the convex hull of the control polygon

$$(X(t), Y(t))_t \subset \text{Conv}(X_0, Y_0).$$

Figure 5.7 shows examples of subdivision curves.

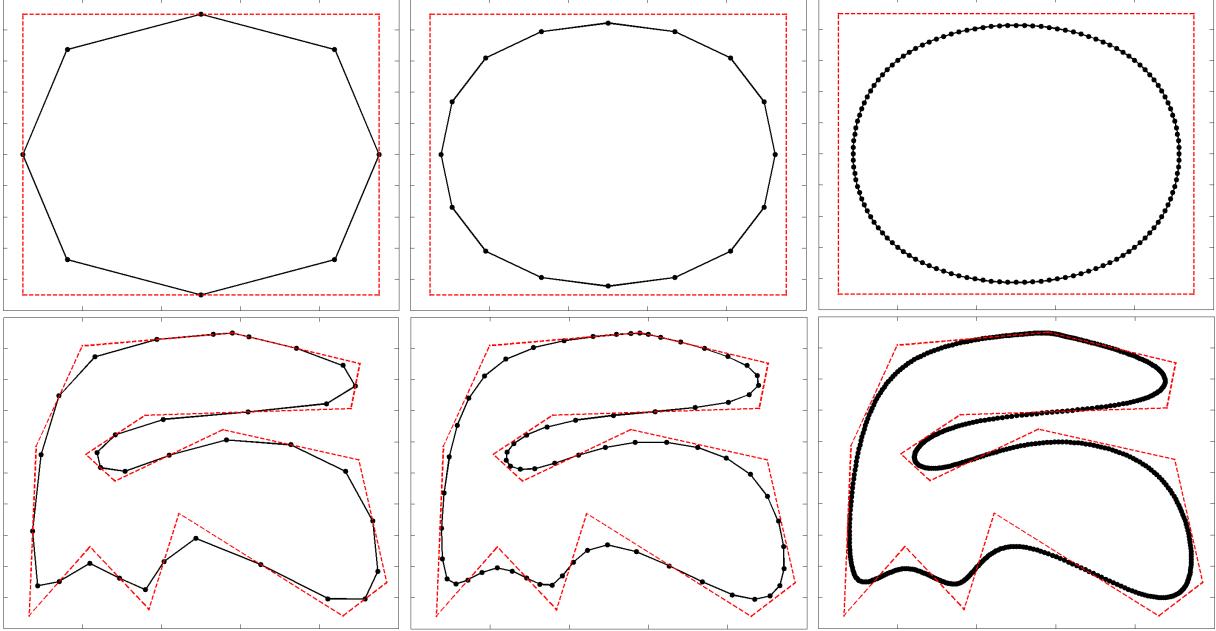


Figure 5.7: Two examples of subdivision curves. The red curve represent the original curve (X^0, Y^0) .

5.3 Subdivision Surfaces

Subdivision schemes allows to compute a set of progressively refined vectors on a semi-regular mesh. More precisely, from an initial vector $f_0 \in \mathbb{R}^{|V_0|}$ defined on the coarse mesh M_0 , local interpolation kernels computes iteratively vectors $f_j \in \mathbb{R}^{|V_j|}$ of finer resolution. When applied to 3 function $(f_i^j)_{i=1,2,3}$ defining the geometrical position of points in \mathbb{R}^3 , this hierarchical construction defines a subdivision surface. These subdivisions surfaces are used extensively in computer aided geometry and computer graphics. One can see [15] for a survey of subdivision surfaces and their applications.

5.3.1 Interpolation Operators

In order to refine a vector $f_j \in \mathbb{R}^{|V_j|}$ defined on the vertex V_j of the mesh M_j , one uses two interpolators

$$P_j : \ell^2(V_j) \longrightarrow \ell^2(H_j) \quad \text{and} \quad \tilde{P}_j : \ell^2(V_j) \longrightarrow \ell^2(V_j). \quad (5.1)$$

A new refined function $f_{j-1} \in \mathbb{R}^{|V_{j-1}|}$ defined on the vertices $V_{j-1} = V_j \cup H_j$ of M_{j-1} is defined by applying these two refinement operators:

$$\forall \ell \in V_{j-1}, \quad f_{j-1}(\ell) = \begin{cases} (P_j f_j)(\ell) & \text{if } \ell \in V_j, \\ (\tilde{P}_j f_j)(\ell) & \text{if } \ell \in H_j. \end{cases}$$

Since $V_j \subset V_{j-1}$, the operator \tilde{P}_j only modify slightly the value at vertex in V_j . On the other hand, the operator P_j creates new value at the vertices of H_j that are inserted between V_j and V_{j-1} .

In practical applications, these interpolating operators are local, meaning that the value of $(P_j f_j)(\ell)$ and $(\tilde{P}_j f_j)(\ell)$ depends only on values $f_j(\ell')$ for $\ell' \in V_j$ being close to $\ell \in V_{j-1}$, typically in the 1-ring or 2-ring vertex neighborhood.

A particularly important setting for subdivision scheme is when one apply the subdivision steps in parallel to three vectors (X_j, Y_j, Z_j) starting from three initial vectors describing the position in 3D space of a coarse mesh M_0 . This allows to defines finer and finer spacial localization for the vertex of the refined meshes M_j . Figure 5.9 shows an example of such a subdivision surface. In order for the resulting infinitely refined surface to have good properties such as being continuous and even smooth, one needs to design carefully the interpolation operators. Next section gives examples of such operators.

5.3.2 Some Classical Subdivision Stencils

In order to define the interpolation operators P_j and \tilde{P}_j of equation (5.1), one needs to use a naming convention for the neighborhoods of vertices.

For a vertex $\ell \in V_j$, the one ring neighborhood V_ℓ has already been defined in equation (3.1). It is the set of vertices adjacent to ℓ . In a regular point (that does not belongs to V_0 and not on a boundary of the mesh), its size is $|V_\ell| = 6$ since a point has 6 neighbors. This 1-ring is used to define \tilde{P}_j .

For a vertex $k \in H_j \subset V_{j-1}$, the butterfly neighborhood is a set of vertices in V_j close to k . This neighborhood is used to define P_j . The two immediate neighbors are

$$(v_k^1, v_k^2) \stackrel{\text{def.}}{=} \{v \in V_j ; (v, k) \in E_{j-1}\}.$$

Two other vertices (w_k^1, w_k^2) are defined using the two faces adjacent to edge $(v_k^2, v_k^2) \in E_j$

$$f_k^1 = (v_k^1, v_k^2, w_k^1) \in F_j \quad \text{and} \quad f_k^2 = (v_k^1, v_k^2, w_k^2) \in F_j.$$

For edges E_j on the boundary of M_j , one face is available, in which case we implicitly assume that $f_1 = f_2$ (reflecting boundary conditions). The four last vertices are defined using faces adjacent to f_1 and f_2 :

$$\forall i, j = 1, 2, \quad f_k^{i,j} \stackrel{\text{def.}}{=} (z_k^{i,j}, v_k^j, w_k^j) \in F_j \quad \text{with} \quad f_k^{i,j} \neq f_j.$$

Once again, reflecting boundary condition are applied for faces on the boundary of the mesh. The butterfly neighborhood is depicted on figure 5.8.

Linear Interpolating Scheme The simplest subdivision rule compute values along edge mide point using a simple linear interpolation as follow

$$\begin{cases} \forall k \in H_j, & (P_j f_j)(k) = \frac{1}{2}(f(v_k^1) + f(v_k^2)), \\ \forall \ell \in V_j, & (\tilde{P}_j f_j)(\ell) = f_j(\ell). \end{cases} \quad (5.2)$$

Since \tilde{P}_j is the identity operator, this scheme is called interpolating. It means that value of f_0 on points of the coarse triangulation are kept during iteration of the subdivision.

Butterfly Interpolating Scheme The linear scheme creates function that are piecewise linear on each face of the coarse triangulation F_0 . In order to create smooth surface, one needs to use more points in the butterfly neighborhood as follow

$$\begin{cases} \forall k \in H_j, & (P_j f_j)(k) = \frac{1}{2} \sum_{i=1}^2 f(v_k^i) + \frac{1}{8} \sum_{i=1}^2 f(w_k^i) - \frac{1}{16} \sum_{i,j=1}^2 f(z_k^{i,j}), \\ \forall \ell \in V_j, & (\tilde{P}_j f_j)(\ell) = f_j(\ell). \end{cases} \quad (5.3)$$

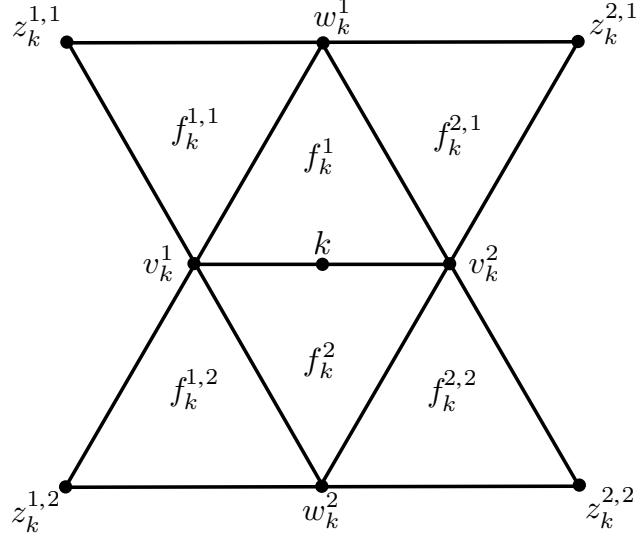


Figure 5.8: The butterfly neighborhood of a vertex $k \in H_j$.

Loop Approximating Scheme In order to gain flexibility in the subdivision design, one can also modify points in V_j during the iterations. This means that \tilde{P}_j is not any more the identity, and that all the values will evolves during the iterations. The question of whether these iterated modification actually converge to a limit value is studied in the next section.

The Loop subdivision rule is defined as

$$\begin{cases} \forall k \in H_j, & (P_j f_j)(k) = \frac{3}{8} \sum_{i=1}^2 f(v_k^i) + \frac{1}{8} \sum_{i=1}^2 f(w_k^i), \\ \forall \ell \in V_j, & (\tilde{P}_j f_j)(\ell) = (1 - |V_\ell| \beta_{|V_\ell|}) f_j(\ell) + \beta_{|V_\ell|} \sum_{\ell' \in V_\ell} f_j(\ell'). \end{cases}$$

where the weights depends on the number of neighbors and are defined as

$$\beta_m \stackrel{\text{def.}}{=} \frac{1}{m} \left(\frac{5}{8} - \left(\frac{3}{8} + \frac{1}{4} \cos(2\pi/m) \right)^2 \right).$$

Other schemes. It is possible to define subdivision schemes using rules that do not involve a regular 1:4 splitting of each coarse face. For instance, in dual schemes such as the one depicted in figure 5.11, the faces of F_j are not included in F_{j-1} but only in F_{j-2} .

5.3.3 Invariant Neighborhoods

In order to study the convergence of subdivision schemes, one needs to consider independently each vertex $x \in V_{j_0(x)}$, where $j_0(x)$ is the coarser scale at which x appears

$$j_0(x) = \max \{j ; x \in V_j\}.$$

Original vertices satisfy $j_0(x) = 0$ and are the only one (except boundary vertices) that have a non-regular connectivity.

The vertex x belongs to the mesh $M_{j_0(x)}$ which is going to be refined through scales $j < j_0(x)$. In order to analyze this refinement, one needs to define an invariant neighborhood $V_j^x \subset V_j$ of x for each scale $j \leq j_0(x)$. These neighborhood are the set of points that are required to compute the operators P_j and \tilde{P}_j .

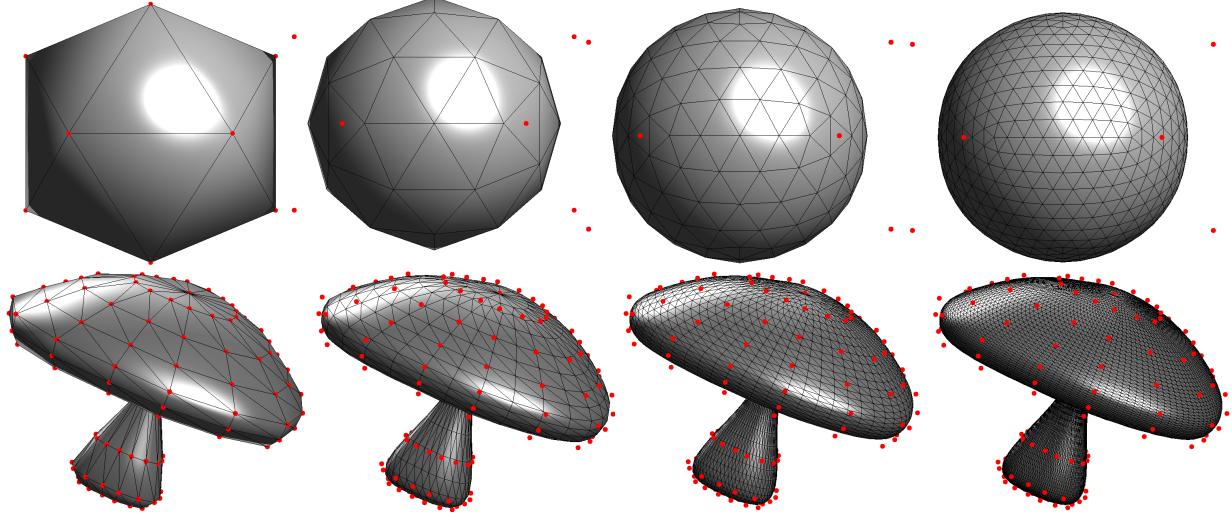


Figure 5.9: Examples of iterative subdivision using Loop scheme. The points (X_0, Y_0, Z_0) of the initial coarse mesh M_0 are shown in red.

More precisely, given a vector $f \in \ell^2(V_{j-1})$, the neighborhoods are required to satisfy

$$\begin{cases} \forall \ell \in V_{j-1}^x \cap V_j, & (\tilde{P}_j f)(\ell) \text{ depends only on } V_j^x \\ \forall k \in V_{j-1}^x \cap H_j, & (P_j f)(k) \text{ depends only on } V_j^x. \end{cases}$$

We further impose that all the invariant neighborhoods have the same size

$$\forall j \leq j_0(x), \quad \#V_j^x = m_x.$$

Figure 5.12 shows an example of invariant neighborhood which corresponds to the 2-ring $V_\ell^{(2)}$, as defined in (3.2).

Thanks to the invariance of these neighborhood systems, one can restrict the predictors around x and define

$$P_j^x : V_j^x \longrightarrow V_{j-1}^x \cap V_j \quad \text{and} \quad \tilde{P}_j^x : V_j^x \longrightarrow V_{j-1}^x \cap H_j.$$

The subdivision matrix $S_j^x \in \mathbb{R}^{m_x \times m_x}$ is then defined as matrix of the following mapping

$$(\tilde{P}_j^x, P_j^x) : V_x^j \longrightarrow V_x^{j-1}.$$

All the subdivision schemes studied in this chapter are invariant, meaning that the subdivision rule does not change through the scales j . This impose that the subdivision matrices are constant $S_j^x = S^x$. In fact, in all the examples given in the previous section, they only depends on the number $|V_x|$ of neighbors in the one ring of x .

5.3.4 Convergence of Subdivisions

The value at $x \in V_{j_0(x)}$ of a function $f_j \in \ell^2(V_j)$ obtained by subdividing at scale $j \leq j_0(x)$ an initial vector $f_0 \in \ell^2(V_0)$ can be computed as

$$f_j(x) = (S^x f_{j+1}^x)(x) = ((S^x)^{j_0(x)-j} f_{j_0(x)}^x)(x),$$

where the vector $f_i^x \in \mathbb{R}^{m_x}$ is the restriction of f_i to the set V_i^x .

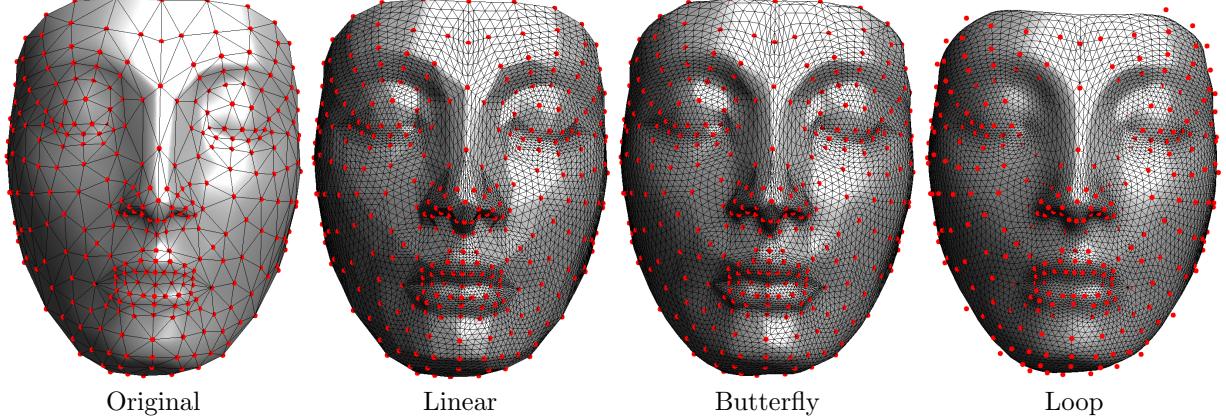


Figure 5.10: Examples of subdivision schemes. The points (X_0, Y_0, Z_0) of the initial coarse mesh M_0 are shown in red. Since the linear and butterfly scheme are interpolating, these points actually belongs to the limiting surface.

In order to analyze the limiting function resulting from an infinite number of subdivision, one can use the eigen vector decomposition of the matrix S^x

$$S^x = \tilde{\Phi} V \Lambda \Phi^T \quad \text{where} \quad \begin{cases} \Phi^T = \tilde{\Phi}^{-1}, \\ \Lambda = \text{diag}(\lambda_i), \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_{m_x}. \end{cases}$$

Since the subdivision matrix S_x is not symmetric, some of the eigenvalues might be complex, and we shall ignore this difficulty here. The fact that P_j and \tilde{P}_j are predictor implies that the subdivision matrix has to satisfy $S^x 1 = 1$, meaning that $\tilde{\varphi}_1 = 1$ is an eigenvector associated to the eigenvalue 1. In the following we further makes the following assumption

$$1 = \lambda_1 < \lambda \stackrel{\text{def.}}{=} \lambda_2 = \lambda_3 < \lambda_4. \quad (5.4)$$

This hypothesis is satisfied by all the subdivision rules introduced in the previous section.

If one write $\Phi = (\varphi_i)_{i=1}^{m_x}$ and $\tilde{\Phi} = (\tilde{\varphi}_i)_{i=1}^{m_x}$, one has the following decomposition of a vector $f \in \mathbb{R}^{m_x}$

$$f = \sum_{i=1}^{m_x} \langle f, \varphi_i \rangle \tilde{\varphi}_i \quad \text{and} \quad (S^x)^k(x) = \sum_{i=1}^{m_x} \lambda_i^k \langle f, \varphi_i \rangle \tilde{\varphi}_i.$$

One thus has the following asymptotic expansion

$$\frac{1}{\lambda^k} (f - \langle f, \varphi_1 \rangle 1) = \langle f, \varphi_2 \rangle \tilde{\varphi}_2 + \langle f, \varphi_3 \rangle \tilde{\varphi}_3 + o(1). \quad (5.5)$$

This expression describes the asymptotic behavior of the subdivision scheme at zero order (position) and first order (tangents).

Theorem 25 (Convergence of the subdivision scheme). *If the subdivision matrix S^x of a point x satisfies (5.4) then the subdivision process converges at x to the value*

$$f^j(x) \xrightarrow{j \rightarrow -\infty} \langle f_{j_0(x)}, \varphi_1 \rangle.$$

The smoothness of the resulting function is more difficult to analyze. A particularly important setting is when one computes the subdivision of 3 function $p_0 = (X_0, Y_0, Z_0) \in \ell^2(V_0)^3$ corresponding to the position

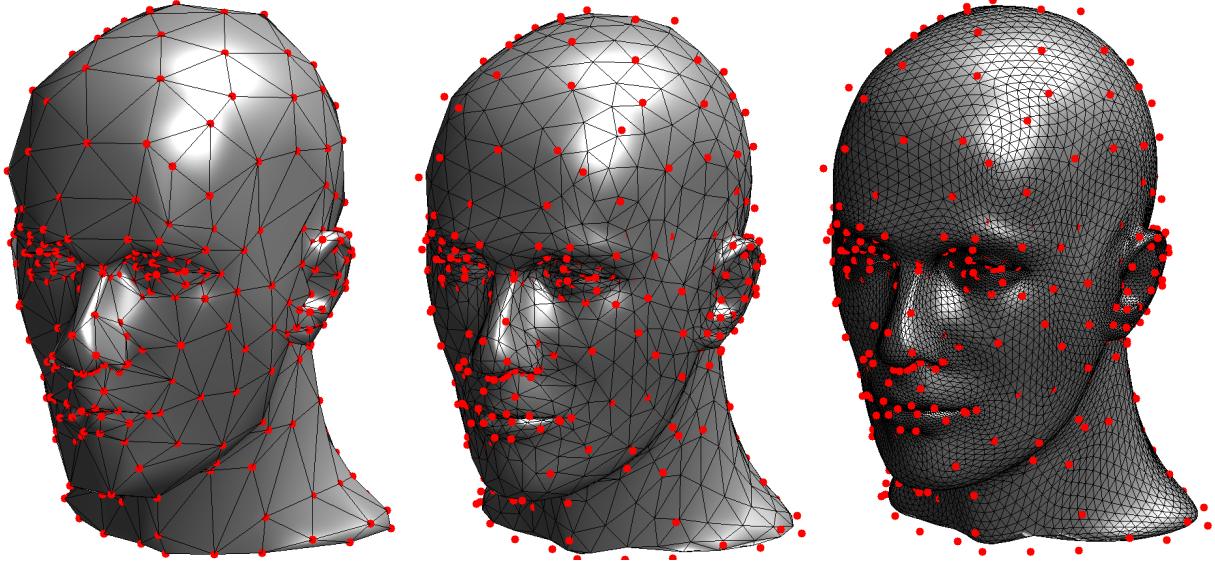


Figure 5.11: Surface after 0, 1 and 3 step of $\sqrt{3}$ subdivision [25].

in \mathbb{R}^3 (geometrical realization) of a coarse mesh M_0 . In this case, the subdivided functions $p_j = (X_j, Y_j, Z_j)$ gives refined 3D meshes that converge uniformly to a continuous surfaces

$$p(x) = (X(x), Y(x), Z(x)) = (\langle X_{j_0}^x, \varphi_1 \rangle, \langle Y_{j_0}^x, \varphi_1 \rangle, \langle Z_{j_0}^x, \varphi_1 \rangle).$$

Condition (5.4) nearly implies that the resulting surface is smooth. Indeed, the asymptotic expansion (5.5) shows that for a point x' near x in the subdivision domain, the differential vector can be well approximated as a projection on a 2D plane

$$p(x) - p(x') + o(1) \in \text{Span}(\tau_2^x, \tau_3^x) \quad \text{where} \quad \tau^i(x) \stackrel{\text{def.}}{=} (\langle X_{j_0}^x, \varphi_i \rangle, \langle Y_{j_0}^x, \varphi_i \rangle, \langle Z_{j_0}^x, \varphi_i \rangle).$$

If the vectors τ_2^x and τ_3^x are linearly independent, they form a basis of the tangent plane at $p(x)$.

Example of the Loop subdivision. For the Loop interpolation operators defined in equation (5.3.2), the invariant neighborhood V_j^x correspond to the 2-ring of x in the triangulation G_j , as shown in figure 5.12. For a vertex with k neighbors, $|V_x| = k$, the size of these invariant neighborhood is $m_x = 3k + 1$. A particular neighboring for $k = 3$ is depicted in figure 5.12, together with an indexing in $\{0, \dots, 3k = 9\}$ of the points in V_j^x and V_{j-1}^x . For this indexing, the subdivision matrix reads

$$\begin{pmatrix} 7 & & & & 3 & 3 & 3 \\ 1 & 1 & & 1 & 1 & 10 & 1 & 1 \\ 1 & & 1 & 1 & 1 & 1 & 10 & 1 \\ 1 & & & 1 & 1 & 1 & 1 & 10 \\ 1 & & & & 1 & 3 & 3 & \\ 1 & & & & & 1 & 3 & 3 \\ 1 & & & & & & 1 & 3 \\ 1 & & & & & & & 1 \\ 1 & & & & & & & 1 \\ 1 & & & & & 2 & 1 & 3 \end{pmatrix}$$

where the 0's have been omitted and where the rows should be rescaled to sum to 1. The eigenvalues of this matrix satisfy $\lambda_1 = 1$ and $\lambda_2 = \lambda_3 = 1/3 > \lambda_4$.

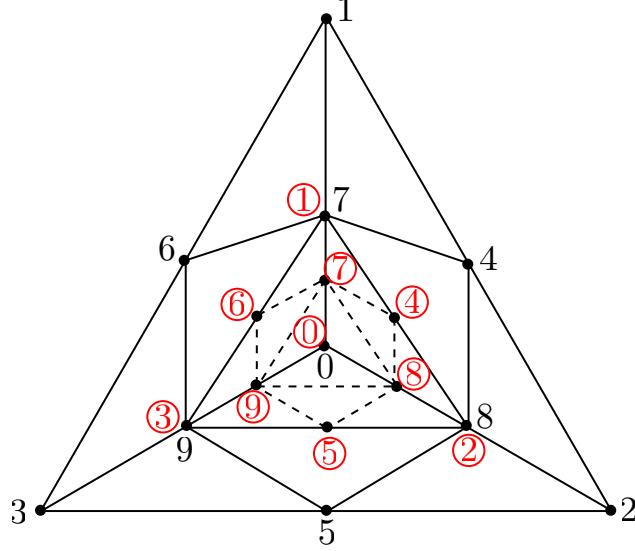


Figure 5.12: Invariant neighborhood V_j^x and V_{j-1}^x (indexing with red circles) of the Loop subdivision scheme for a vertex of valence $|V_\ell| = 0$. The number in $\{0, \dots, 9\}$ refers to the numbering of the vertices in V_j^x and V_{j-1}^x

5.4 Wavelets on Meshes

5.4.1 Multiscale Biorthogonal Bases on Meshes

The transforms considered in this section are multiscale and indexed by the set of nested grids $(V_j)_{L < j \leq J}$. This corresponds to computing a set of coefficients $(d_j)_{L < j \leq J} \cup f_J$ from an initial input signal f . These coefficients corresponds to inner products with basis vectors

$$\begin{cases} d_j \in \ell^2(H_j) & \text{where } \forall k \in H_j, d_j(k) = \langle f, \psi_{j,k} \rangle, \\ f_J \in \ell^2(V_J) & \text{where } \forall \ell \in V_J, f_J(\ell) = \langle f, \varphi_{J,\ell} \rangle. \end{cases}$$

By analogy with the wavelet setting, the vectors $\psi_{j,k} \in \mathbb{R}^n$ corresponds to primal wavelets and are intended to capture the details present in the signal f at a scale j , whereas the scaling vectors $\varphi_{J,k} \in \mathbb{R}^n$ capture the missing coarse approximation of f at scale J . This decomposition is stopped at any coarse scale $L < J \leq 0$.

In order to reconstruct the function f from this set of transformed coefficients, one needs to use a set of bi-orthogonal basis vectors

$$f = \sum_{L < j \leq J, k \in H_j} d_j(k) \tilde{\psi}_{j,k} + \sum_{\ell \in V_J} f_J(\ell) \tilde{\varphi}_{J,\ell}.$$

If this reconstruction formula holds for any scale $L < J \leq 0$, the set of vectors

$$(\psi_{j,k}, \varphi_{j,\ell})_{k \in H_j, \ell \in V_j}^{L < j \leq 0} \quad \text{and} \quad (\tilde{\psi}_{j,k}, \tilde{\varphi}_{j,\ell})_{k \in H_j, \ell \in V_j}^{L < j \leq 0}, \quad (5.6)$$

is said to be a pair of primal and dual multiscale bases (together with their scaling functions).

The following paragraph shows how one can modify such a pair of multiscale bases while still maintaining the biorthogonality property. This lifting process is useful to design multiscale bases with various properties on complicated domains.

5.4.2 The Lifting Scheme

The lifting scheme is a construction of multiscale biorthogonal bases introduced by Sweldens [41, 42]. It extends the traditional construction of wavelets in two main directions:

- As explained in [17], it allows to implement already existing filter banks more efficiently by splitting the computation into elementary blocks. This computational gain is described at the end of the section together with the factorization of wavelets into lifting steps.
- It allows to define multiscale transforms over domains that are not translation invariant. This section gives two examples of such transforms: a non-separable 2D wavelet transform and wavelets on triangulated meshes.

In order to build wavelets on triangulation, one can specialize the lifting scheme to a particular setting where only two lifting steps are applied.

Forward lifting scheme. The forward algorithm performs the transform

$$(f_{j-1}(\ell))_{\ell \in V_{j-1}} \longrightarrow (d_j(k))_{k \in H_j} \cup (f_j(\ell))_{\ell \in V_j}$$

by applying the following steps

- *Splitting:* this corresponds selecting the coefficient of $f_{j-1}(\ell)$ that are in V_j or in H_j

$$(f_{j-1}(\ell))_{\ell \in V_{j-1}} = (f_j(\ell))_{\ell \in V_j} \cup (f_j(\ell))_{\ell \in H_j}.$$

These two sets of coefficients are treated differently in the two remaining steps of the transform.

- *Predict step:* creates wavelets coefficients d_j by computing local differences between each coefficient in V_j and its neighbors in H_j

$$\forall k \in H_j, \quad d_j(k) = f_{j-1}(k) - \sum_{\ell \in V_j} p_j(k, \ell) f_{j-1}(\ell).$$

The coefficients $p_j(k, \ell)$ are weights that determine the predict operator

$$P_j : \begin{cases} \ell^2(V_j) & \longrightarrow \ell^2(H_j) \\ g & \mapsto h = P_j g \end{cases} \quad \text{where} \quad h(k) = \sum_{k \in H_j} p_j(k, \ell) g(\ell).$$

- *Update step:* enhance the properties of each remaining low pass coefficients $f_{j-1}(\ell)$ for $\ell \in V_j$ by pooling locally the wavelets coefficients $d_j(k)$ for k around ℓ

$$\forall \ell \in V_j, \quad f_j(\ell) = f_{j-1}(\ell) + \sum_{k \in H_j} u_j(\ell, k) d_j(k).$$

The coefficients $u_j(\ell, k)$ are weights that determine the update operator

$$U_j : \begin{cases} \ell^2(H_j) & \longrightarrow \ell^2(V_j) \\ h & \mapsto g = U_j h \end{cases} \quad \text{where} \quad g(\ell) = \sum_{k \in H_j} u_j(\ell, k) h(k).$$

Figure 5.13, top row, shows the block diagram associated to this forward lifting wavelet transform.

The iterations of the forward lifting transform can also be written in vector and operator format

$$\begin{cases} d_j = f_{j-1}^{H_j} - P_j f_{j-1}^{V_j}, \\ f_j = f_{j-1}^{V_j} + U_j d_j = (\text{Id}_{V_j} - U_j P_j) f_{j-1}^{V_j} + U_j f_{j-1}^{H_j}, \end{cases}$$

where g^A is the restriction of some vector g to the set A .

Backward lifting scheme. The backward transform algorithm does the reverse computation

$$(d_j(k))_{k \in H_j} \cup (f_j(\ell))_{\ell \in V_j} \longrightarrow (f_{j-1}(\ell))_{\ell \in V_{j-1}}$$

One of the main feature of the lifting scheme is that this is achieved by simply reversing the order of the lifting steps and interchanging $+/-$ signs.

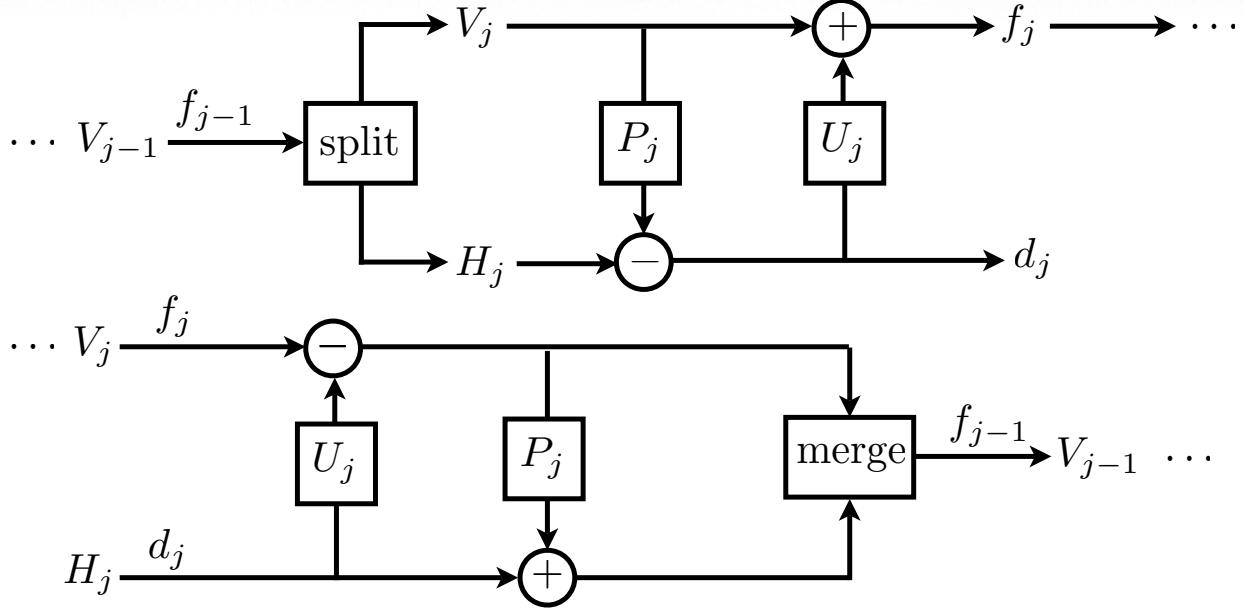


Figure 5.13: Block diagrams for the forward and backward lifting scheme.

– *Inverse update step:*

$$\forall \ell \in V_j, \quad f_{j-1}(\ell) = f_j(\ell) - \sum_{k \in H_j} u_j(\ell, k) d_j(k).$$

– *Inverse predict step:*

$$\forall k \in H_j, \quad f_{j-1}(k) = d_j(k) + \sum_{\ell \in V_j} p_j(k, \ell) f_{j-1}(\ell).$$

– *Merging:* makes the union of the coefficients computed in the two previous steps

$$(f_{j-1}(\ell))_{\ell \in V_{j-1}} = (f_j(\ell))_{\ell \in V_j} \cup (f_j(\ell))_{\ell \in H_j}.$$

Figure 5.13, top row, shows the block diagram associated to this backward lifting wavelet transform.

The lifting scheme is more general than the algorithm described in this section since several passes of predict/update steps can be applied to further enhance the properties of the resulting transform. However, the steps beyond the two initial ones are difficult to analyze, except in the notable exception of points sampled evenly on a 1D axes, where a factorization algorithm [17] allows to recover traditional wavelet filters.

5.4.3 Imposing vanishing moments.

The operator P_j is called a predictor since the values of $P_j f_{j-1}^{V_j}$ should typically be close to $f_{j-1}^{H_j}$ for the wavelet coefficients d_j to be small. Such predictors have already been constructed in equations (5.2), (5.3) and (5.3.2).

The operator U_j is called an update operator since the additional term $U_j d_j$ should enhance the properties of $f_{j-1}^{V_j}$. This update steps does not appear in the theory of subdivision surface and this section considers a local update operator which guaranty the conservation of the mean value when switching from f_{j-1} to f_j .

Polynomial vectors. In order to select predict and update operator that have good properties, one follow the insight gained from the analysis of the wavelet approximation of signal on the real line. In order to do

so, one need analyze the effect of a lifting wavelet transform on polynomials. The most basic constraint enforces one vanishing moment by imposing orthogonality with the constant vector $\Phi_0 = 1$. This constraint does not require to known the spacial location x_ℓ of each index $\ell \in V_L$. In order to impose higher order vanishing moments, one needs to assume some sampling pattern, for instance

$$\forall \ell \in V_L, \quad f(\ell) = \bar{f}(x_\ell) \quad \text{where} \quad x_\ell \in \mathbb{R}^q$$

and where \bar{f} is a function defined on \mathbb{R}^q . For instance, the points x_ℓ might corresponds to a regular sampling of the line (this is the traditional wavelet setting) or to an irregular sampling of a 2D surface embedded in \mathbb{R}^3 . The next paragraphs describe several situations with different sampling grids. Once the precise locations of the samples are known, one can for instance select Φ_s as some monomials of degree (s_1, \dots, s_q) over \mathbb{R}^q .

Vanishing moment and polynomials reproduction. Having defined these polynomial vectors, one requires that the following constraints are fulfilled.

- *Vanishing moments:* the wavelet coefficients of a low order polynomial should be 0, which implies that

$$\forall k \in H_j, \quad \langle \Phi_s, \psi_{j,\ell} \rangle = 0. \quad (5.7)$$

- *Polynomial reproduction:* coarse coefficients f_j computed from a polynomial f_{f-1} should also be polynomials, which implies that

$$\forall \ell \in V_j, \quad \langle \Phi_s, \varphi_{j,\ell} \rangle = \Phi_s(\ell) \quad (5.8)$$

In order for the wavelets and scaling function to satisfy conditions (5.7) and (5.8), the predict operator P_j and update operator U_j should be designed carefully. One can impose these constraint from the fine scale $j = L$ until the coarse scale $j = 0$. Indeed, if $(\varphi_{j-1,\ell}, \psi_{j-1,\ell})_{k,\ell}$ satisfy conditions (5.7) and (5.8), then, for the scale j

$$\forall s \in S, \quad \begin{cases} (5.7) & \iff P_j \Phi_s^{V_j} = \Phi_s^{H_j}, \\ (5.8) & \iff U_j^T \left(\Phi_s^{V_j} + P_j^T \Phi_s^{H_j} \right) = \Phi_s^{H_j}. \end{cases}$$

where $\Phi_s^A \in \ell^2(A)$ is the restriction of Φ_s to A .

In contrast, the constraint (5.8) on the update operator P_j is more involved and the next section shows how to handle it on a triangulation situations for only one vanishing moment $|S| = 1$.

5.4.4 Lifted Wavelets on Meshes

The lifted wavelet bases can be used to process signals $f \in \ell^2(V_L)$ where $\ell \in V_L$ index a sampling x_ℓ of an arbitrary surface. The construction of biorthogonal wavelets on triangulated mesh has been first proposed by Lounsbury et al. [26] and re-casted into the lifting scheme framework by Schroeder and Sweldens [36, 37].

Designing predict operators. The constraints (5.7) on the predictor P_j is easily solved. For instance, for each k , one selects only $|S|$ non vanishing weights $(p_j(k, \ell))_\ell$ and solves a small $|S| \times |S|$ linear system. Furthermore, in the case of a regular triangulation with edges of constant length, predictors with several vanishing moments have been already defined in (5.2), (5.3) and (5.3.2). Figure 5.14 shows the weights for these predictors.

One can choose any of these operators, and creates respectively linear, butterfly and Loop wavelets bases. All these predictors have one vanishing moment since they satisfy $P_j 1^{H_j} = 1^{V_j}$. In fact they have more vanishing moments if one consider polynomials Φ_s sampled at points $x_\ell \in \mathbb{R}^2$ of an hexagonal tiling with constant edge length. In practice, if the triangulation under consideration have edges with smoothly varying length, the resulting predictor are efficient to predict the value of smooth functions on the triangulation.

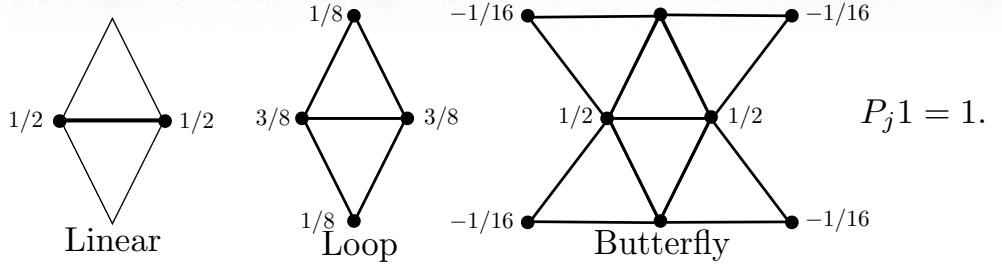


Figure 5.14: Predict operators on a triangulation.

Designing update operators. In order to ensure the reproduction of constant polynomials, we design the update operator so that it depends only on the direct neighbors in H_j of each point in V_j

$$\forall \ell \in V_j, \quad V_\ell = \{\gamma(\ell, \ell') ; (\ell, \ell') \in E_j\}.$$

One then looks for a valid update operator in the following form

$$\forall h \in \ell^2(H_j), \forall \ell \in V_j, \quad (U_j h)(\ell) = \lambda_\ell \sum_{k \in V_\ell} h(k), \quad (5.9)$$

where each λ_ℓ should be fixed in order for condition (5.8) to be satisfied.

In a semi-regular triangulation, $|V_\ell| = 6$ excepted maybe for some points in the coarse grid $\ell \in V_0$. In this setting, the values of λ_ℓ can be computed by a recursion through the scales. In an ideal triangulation where $|V_\ell| = 6$ for all ℓ , one can use a constant weight $\lambda_\ell = \lambda$.

For the predictors defined in (5.2), (5.3) and (5.3.2), one has

$$P_j^T 1^{H_j} = 3 \times 1^{V_j} \quad \text{and} \quad U_j^T 1^{V_j} = 6\lambda 1^{H_j}$$

so setting $\lambda_\ell = 1/24$ solves equation (5.8). Figure 5.15 shows examples of butterfly wavelets on a planar semi-regular triangulation.

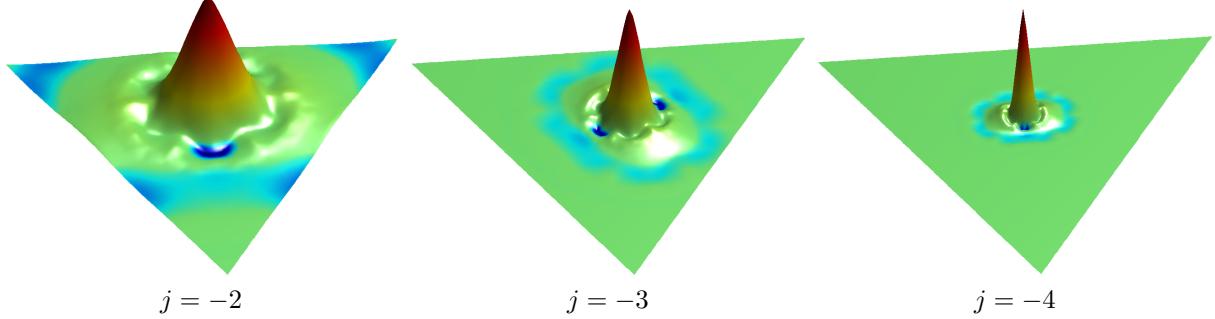


Figure 5.15: Example of wavelets $\psi_{j,k}$ on a semi-regular triangulation. The height over the triangle (together with the color) indicates the value of the wavelet vector.

5.4.5 Non-linear Mesh Compression

These wavelets can be used to perform an approximation of a function $f \in \ell^2(V_L)$ defined on the fine triangulation. For instance a wavelet approximation can be applied to each coordinate $f_i, i = 1, 2, 3$ of the

actual position $x_\ell = (f_1(\ell), f_2(\ell), f_3(\ell)) \in \mathbb{R}^3$ of the surface points, as done in [23, 24]. This leads to a scheme to approximate and compress a 3D surface using the lifted biorthogonal wavelets associated to the semi-regular triangulation. This is possible because these wavelets depend only on the combinatorial grids V_j and not on the precise position of the samples x_ℓ in 3D.

In order to perform a wavelet approximation in this biorthogonal basis, one uses a non-linear thresholding at $T > 0$

$$f = \sum_{(j,k) \in I_T} \langle f, \psi_{j,k} \rangle \tilde{\psi}_{j,k}$$

where $I_T = \{(j, k) ; k \in H_j \text{ and } |\langle f, \psi_{j,k} \rangle| > T |\text{supp}(\psi_{j,k})|^{-1/2}\}.$

Note that for each coefficient the threshold T is scaled according to the size of the support of the wavelet in order to approximately normalize the wavelets in $\ell^2(V_L)$ norm.

Figure 5.16 shows an example of compression of the position of a vertex in 3D spaces as 3 functions defined on a semi-regular mesh. Figure 5.17 shows an example of compression of a spherical texture map which is a single function defined at each vertex of a semi-regular mesh obtained by subdividing an icosaedron.

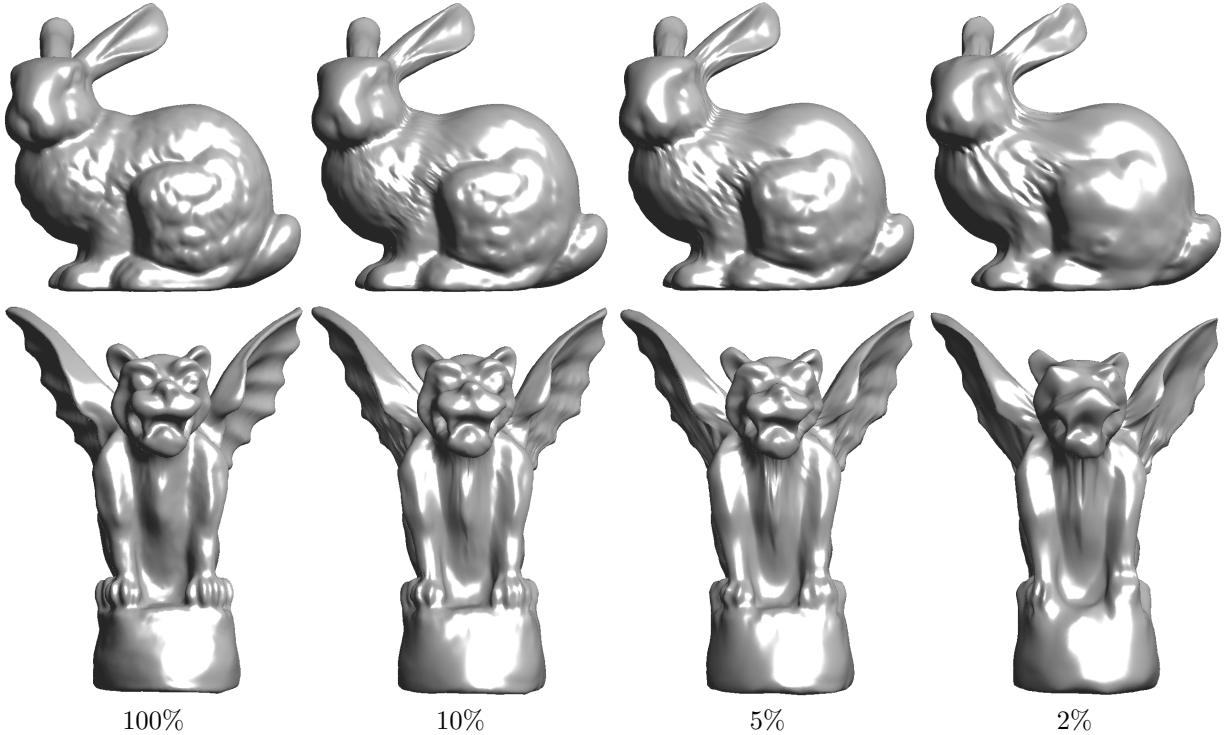


Figure 5.16: Non-linear wavelet mesh compression with a decreasing number of coefficients.

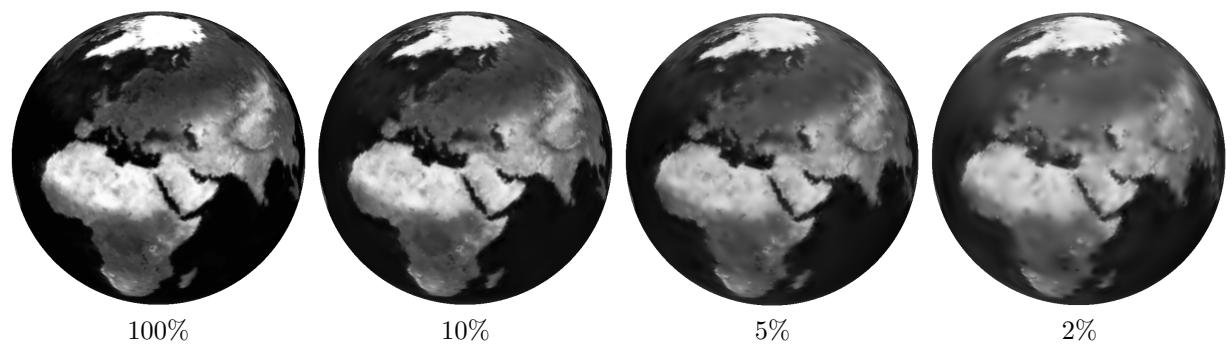


Figure 5.17: Non-linear spherical wavelet compression with a decreasing number of coefficients.

Chapter 6

Linear and Non-linear Approximation

This chapter studies the theory of signal and image approximation, and gives an application to lossy compression. This theoretical analysis is performed for continuous functions $f \in L^2([0, 1]^d)$ for $d = 1, 2$. This analysis is important to studies the performance of compression, denoising, and super-resolution applications.

6.1 Approximation

6.1.1 Approximation in an Ortho-basis

We consider an orthogonal basis $\mathcal{B} = \{\psi_m\}_m$ of $L^2([0, 1]^d)$, with for instance $d = 1$ (signals) or $d = 2$ (images). We recall that the decomposition of a signal in an orthonormal basis

$$f = \sum_{m \in \mathbb{Z}} \langle f, \psi_m \rangle \psi_m$$

gives back the original signal and thus produces no error. Processing algorithms modify the coefficients $\langle f, \psi_m \rangle$ and introduce some error.

The simplest processing computes an approximation by considering only a sub-set $I_M \subset \mathbb{Z}$ of M coefficients and performing the reconstruction from this subset

$$f_M \stackrel{\text{def.}}{=} \sum_{m \in I_M} \langle f, \psi_m \rangle \psi_m, \quad \text{where } M = |I_M|.$$

The reconstructed signal f_M is the orthogonal projection of f onto the space

$$V_M \stackrel{\text{def.}}{=} \text{Span} \{ \psi_m ; m \in I_M \}.$$

Since V_M might depend on f , this projection $f \mapsto f_M$ might be non-linear.

Since the basis is orthogonal, the approximation error is

$$\|f - f_M\|^2 = \sum_{m \notin I_M} |\langle f, \psi_m \rangle|^2.$$

The important question is now to choose the set I_M , which might depend on the signal f itself.

6.1.2 Linear Approximation

Linear approximation is obtained by fixing once for all I_M , and thus using the same set of coefficients for all f . The mapping $f \mapsto f_M$ is a linear orthogonal projection on V_M , and it satisfies

$$(f + g)_M = f_M + g_M$$

For the Fourier basis, one usually selects the low-frequency atoms

$$I_M = \{-M/2 + 1, \dots, M/2\}.$$

For a 1-D wavelet basis, one usually selects the coarse wavelets

$$I_M = \{m = (j, m) ; j \geq j_0\}$$

where j_0 is selected such that $|I_M| = M$.



Figure 6.1: Linear versus non-linear wavelet approximation.

Figure 6.1, center, shows an example of such a linear approximation with wavelets. Linear approximation tends to perform poorly near singularities, because they introduce some blurring.

6.1.3 Non-linear Approximation

A non-linear approximation is obtained by choosing I_M depending on f . In particular, one would like to choose I_M to minimize the approximation error $\|f - f_M\|$. Since the basis is orthogonal, this is achieved by selecting the M largest coefficients in magnitude

$$I_M = \{M \text{ largest coefficients } |\langle f, \psi_m \rangle|\}.$$

This can be equivalently obtained using a thresholding

$$I_M = \{m ; |\langle f, \psi_m \rangle| > T\}$$

where T depends on the number of coefficients M ,

$$M = \# \{m ; |\langle f, \psi_m \rangle| > T\}.$$

Computation of the threshold. There is a bijective 1:1 mapping between T and M obtained by ordering the coefficient magnitudes $|\langle f, \psi_m \rangle|$ by decaying order,

$$T = d_M \quad \text{where} \quad \{d_m\}_{m=0}^{N-1} = \{|\langle f, \psi_m \rangle|\}_{0}^{N-1} \quad \text{and} \quad d_m \geq d_{m+1}. \quad (6.1)$$

Figure 6.2 shows this mapping between M and T .

The following proposition shows that the decay of the ordered coefficients is linked to the non-linear approximation decay.

Proposition 14. One has

$$d_m = O(m^{-\frac{\alpha+1}{2}}) \iff \|f - f_M\|^2 = O(M^{-\alpha}). \quad (6.2)$$

Proof. One has

$$\|f - f_M\|^2 = \sum_{m>M} d_m^2$$

and

$$d_M^2 \leq \frac{2}{M} \sum_{m=M/2+1}^M d_m^2 \leq \frac{2}{M} \sum_{m>M/2} d_m^2 = \frac{2}{M} \|f - f_{M/2}\|^2.$$

□

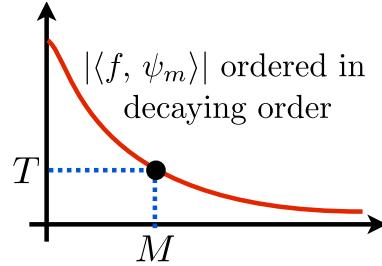


Figure 6.2: Decay of the ordered coefficients and determination of the threshold for non-linear approximation.

Hard thresholding. The non-linear approximation is re-written as

$$f_M = \sum_{|\langle f, \psi_m \rangle| > T} \langle f, \psi_m \rangle \psi_m = \sum_m S_T(\langle f, \psi_m \rangle) \psi_m, \quad (6.3)$$

where

$$S_T(x) = \begin{cases} x & \text{if } |x| > T \\ 0 & \text{if } |x| \leq T \end{cases} \quad (6.4)$$

is the hard thresholding, that is displayed in Figure 6.3.

6.2 Signal and Image Modeling

A signal model is a constraint $f \in \Theta$, where $\Theta \subset L^2([0, 1]^d)$ is a set of signals one is interested in. Figure 6.4 shows different class of models for images, that we describe in the following paragraph.

6.2.1 Uniformly Smooth Signals and Images

Signals with derivatives. The simplest model is made of uniformly smooth signals, that have bounded derivatives

$$\Theta = \{f \in L^2([0, 1]^d) ; \|f\|_{C^\alpha} \leq C\}, \quad (6.5)$$

where $C > 0$ is a fixed constant, and where in 1-D

$$\|f\|_{C^\alpha} \stackrel{\text{def.}}{=} \max_{k \leq \alpha} \left\| \frac{d^k f}{dt^k} \right\|_\infty.$$

This extends to higher dimensional signals by considering partial derivatives along each direction.

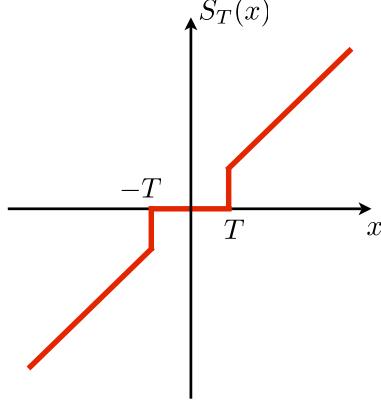


Figure 6.3: Hard thresholding.

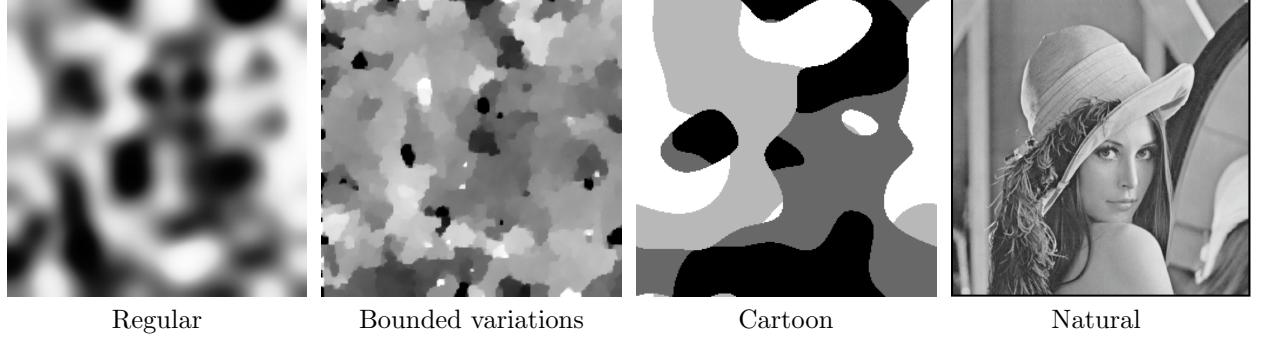


Figure 6.4: Examples of different kinds of image models.

Sobolev smooth signals and images. A smooth C^α signals in (6.5) has derivatives with bounded energy, so that

$$\frac{d^\alpha f}{dt^\alpha}(t) = f^{(\alpha)}(t) \in L^2([0, 1]).$$

Using the fact that

$$\hat{f}_m^{(\alpha)} = (2i\pi m)^\alpha \hat{f}_m$$

where \hat{f} is the Fourier coefficient defined in (2.2), except we are here on \mathbb{R}/\mathbb{Z} in place of $\mathbb{R}/2\pi\mathbb{Z}$,

$$\hat{f}_n \stackrel{\text{def.}}{=} \int_0^1 e^{-2i\pi n x} f(x) dx,$$

one defines a so-called Sobolev functional

$$\|f\|_{Sob(\alpha)}^2 = \sum_{m \in \mathbb{Z}} |2\pi m|^{2\alpha} |\langle f, e_m \rangle|^2, \quad (6.6)$$

that satisfies $\|f\|_{Sob(\alpha)} = \|f^{(\alpha)}\|$ for smooth functions. This Sobolev functional is extended to signals that have derivatives in the sense of distribution in $L^2([0, 1])$.

This definition extends to distributions and signals $f \in L^2([0, 1]^d)$ of arbitrary dimension d as

$$\|f\|_{Sob(\alpha)}^2 = \sum_{m \in \mathbb{Z}^d} (2\pi \|m\|)^{2\alpha} |\langle f, e_m \rangle|^2, \quad (6.7)$$

The C^α -Sobolev model

$$\Theta = \left\{ f \in \ell^2([0, 1]^d) ; \max_{k \leq \alpha} \|f\|_{\text{Sob}(k)}^2 \leq C \right\} \quad (6.8)$$

generalizes the C^α smooth image model (6.5).

Figure 6.5 shows images with an increasing Sobolev norm for $\alpha = 2$.

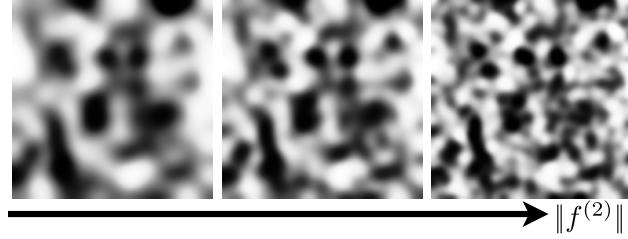


Figure 6.5: Images with increasing Sobolev norm.

6.2.2 Piecewise Regular Signals and Images

Piecewise smooth signals. Piecewise smooth signals in 1-D are functions $f \in L^2([0, 1])$ that are C^α smooth outside a set of less than K pointwise discontinuities

$$\Theta = \left\{ f \in L^2([0, 1]) ; \exists (t_i)_{i=0}^{K-1}, \|f_{(t_i, t_{i+1})}\|_{C^\alpha} \leq C \right\} \quad (6.9)$$

where $f_{(t_i, t_{i+1})}$ is the restriction of f to the open interval (t_i, t_{i+1}) .

Piecewise smooth images. Piecewise smooth images are 2-D functions $f \in L^2([0, 1]^2)$ that are C^α regular outside a set of less than K curves that have a finite perimeter

$$\Theta = \left\{ f \in L^2([0, 1]^2) ; \exists \Gamma = (\gamma_i)_{i=0}^{K-1}, \|f\|_{C^\alpha(\Gamma^c)} \leq C_1 \quad \text{and} \quad |\gamma_i| \leq C_2 \right\} \quad (6.10)$$

where $|\gamma_i|$ is the length of the curve γ_i and where $\|f\|_{C^\alpha(\Gamma^c)}$ is the maximum norm of the derivatives of f outside the set of curves Γ .

Segmentation methods such as the one proposed by Mumford and Shah [29] implicitly assume such a piecewise smooth image model.

6.2.3 Bounded Variation Signals and Images

Signals with edges are obtained by considering functions with bounded variations

$$\Theta = \left\{ f \in L^2(\mathbb{R}^d) ; \|f\|_\infty \leq C_1 \quad \text{and} \quad \|f\|_{\text{TV}} \leq C_2 \right\}. \quad (6.11)$$

For $d = 1$ and $d = 2$, this model generalizes the model of piecewise smooth signals (6.9) and images (6.10).

The total variation of a smooth function is

$$\int \|\nabla f(x)\| dx$$

where

$$\nabla f(x) = \left(\frac{\partial f}{\partial x_i} \right)_{i=0}^{d-1} \in \mathbb{R}^d$$

is the gradient vector at x . The total variation is extended to discontinuous images, that might for instance exhibit jumps across singular curves (the edges of the image). In particular, the total variation of a piecewise smooth image is the sum of the lengths of its level sets

$$\|f\|_{\text{TV}} = \int_{-\infty}^{\infty} |\mathcal{L}_t(f)| dt < +\infty \quad \text{where} \quad \mathcal{L}_t(f) = \{x ; f(x) = t\}, \quad (6.12)$$

and where $|\mathcal{L}_t(f)|$ is the length of $\mathcal{L}_t(f)$. For a set $\Omega \subset \mathbb{R}^2$ with finite perimeter $|\partial\Omega|$, then

$$\|1_\Omega\|_{\text{TV}} = |\partial\Omega|.$$

The model of bounded variation was introduced in image processing by Rudin, Osher and Fatemi [34].

6.2.4 Cartoon Images

The bounded variation image model (6.11) does not constrain the smoothness of the level set curves $\mathcal{L}_t(f)$. Geometrical images have smooth contour curves, which should be taken into account to improve the result of processing methods.

The model of C^α cartoon images is composed of 2-D functions that are C^α regular outside a set of less than K regular edge curves γ_i

$$\Theta = \{f \in L^2([0, 1]^2) ; \exists \Gamma = (\gamma_i)_{i=0}^{K-1}, \|f\|_{C^\alpha(\Gamma^c)} \leq C_1 \quad \text{and} \quad \|\gamma_i\|_{C^\alpha} \leq C_2\} \quad (6.13)$$

where each γ_i is a arc-length parameterization of the curve $\gamma_i : [0, A] \mapsto [0, 1]^2$. Figure 6.6 shows cartoon images with increasing total variation $\|f\|_{\text{TV}}$.



Figure 6.6: Cartoon image with increasing total variation.

Typical images might also be slightly blurred by optical diffraction, so that one might consider a blurred cartoon image model

$$\tilde{\Theta} = \{\tilde{f} = f * h \in L^2([0, 1]^2) ; f \in \Theta \quad \text{and} \quad h \in \mathcal{H}\} \quad (6.14)$$

where Θ is the model of sharp (unblurred) images (6.13) and \mathcal{H} is a set of constraints on the blurring kernel, for instance $h \geq 0$ should be smooth, localized in space and frequency. This unknown blurring makes difficult brute force approaches that detects the edges location Γ and then process the regular parts in $[0, 1]^2 \setminus \Gamma$.

Figure 6.7 shows examples of images in Θ and $\tilde{\Theta}$.

6.3 Efficient approximation

6.3.1 Decay of Approximation Error

To perform an efficient processing of signals or images in Θ , the goal is to design an orthogonal basis such that the non-linear approximation error $\|f - f_M\|$ decays as fast as possible to 0 when M increases.

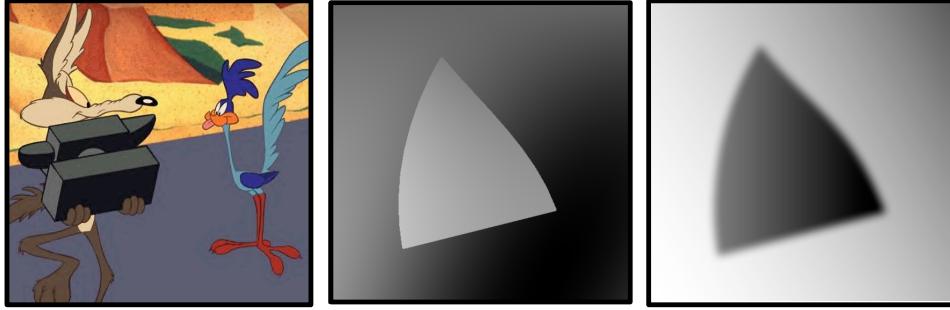


Figure 6.7: Examples of cartoon images: sharp discontinuities (left and center) and smooth discontinuities (right).

Polynomial error decay. This error decay measured using a power law

$$\forall f \in \Theta, \forall M, \quad \|f - f_M\|^2 \leq C_f M^{-\alpha} \quad (6.15)$$

where α is independent of f and should be as large as possible. The parameter α depends on the basis and on Θ . It is a class-complexity parameter that describes the overall complexity of signals in Θ with respect to the orthogonal basis one considers for approximation. The parameter C_f depends on f and describes the complexity of the signal f within its class Θ .

Relevance for compression, denoising and inverse problems. Monitoring the decay of approximation error is not only interesting from a mathematical point of view. Section 7.1 shows that the compression error is close to the non-linear approximation error. Bases that are efficient for approximation are thus also efficient for compression.

Chapter ?? shows that a similar conclusion holds for non-linear denoising with thresholding. Efficient denoisers are obtained by performing a non-linear approximation of the noisy image in a well chosen basis. The average denoising error with respect to a random noise is closely related to the approximation error.

Chapter ?? shows that ill-posed inverse problems such as super-resolution of missing information can be solved by taking advantage of the compressibility of the signal or the image in a well chosen basis. A basis that is efficient for approximation of the high resolution signal is needed to recover efficiently missing information. The performance of these schemes is difficult to analyze, and the basis atoms must also be far enough from the kernel of the operator that removes information.

Comparison of signals. For a fixed basis (for instance wavelets), the decay of $\|f - f_M\|$ allows one to compare the complexity of different images. Figure 6.9 shows that natural images with complicated geometric structures and textures are more difficult to approximate using wavelets.

Since the approximation error often decays in a power-low fashion (6.15), the curves are displayed in a log-log plot, so that

$$\log(\|f - f_M\|^2) = \text{cst} - \alpha \log(M)$$

and hence one should expect an affine curve with slope $-\alpha$. Due to discretization issue, this is only the case for value of $M \ll N$, since the error quickly drops to zero for $M \approx N$.

6.3.2 Comparison of bases.

For a given image f , one can compare different ortho-bases using the decay of $\|f - f_M\|$. Figure 6.11 shows the efficiency of several bases to approximate a fixed natural image with contours and textures. The Fourier basis described in Section ?? is highly inefficient because of periodic boundary artifact and the global

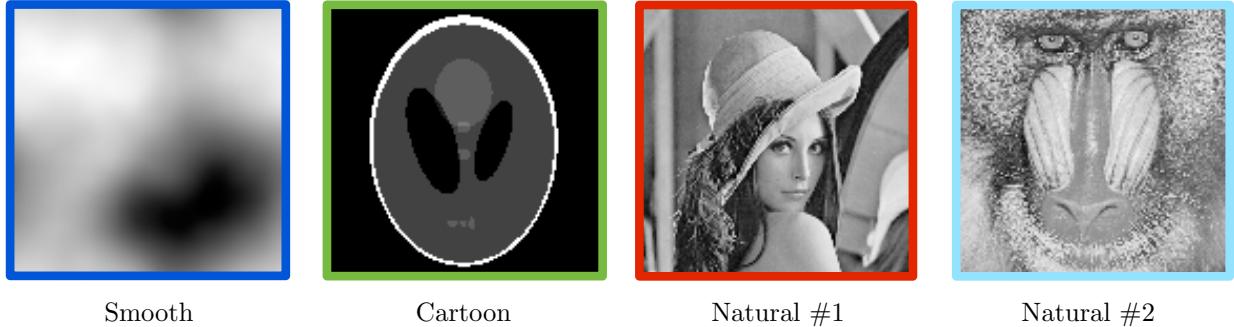


Figure 6.8: Several different test images.

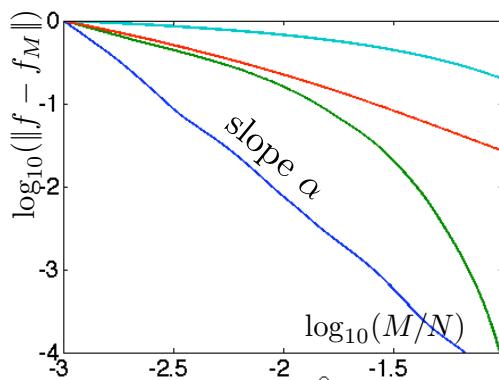


Figure 6.9: Comparison of approximation error decay in wavelets for different images shown in Figure 6.8.

support of its atoms that fail to capture contours. The cosine basis uses symmetric boundary conditions and thus removes the boundary artifacts, but it still not able to resolve efficiently localized features. The local DCT basis corresponds to the union of local cosine bases defined on small square patches. It is more efficient since its atoms have a local support. However, it gives bad approximation for a small number M of kept coefficients, because of blocking artifacts. The isotropic wavelet basis detailed in Section ?? gives the best approximation results because its is both composed of localized atoms and does not have a block structure but rather a multiresolution structure.

Figure 6.12 summarizes the different type of approximation decays for various class of data, which are detailed in the following section.

6.4 Fourier Linear Approximation of Smooth Functions

The smooth signal and image model (6.5) assumed that the analog function have bounded α continuous derivatives. A function f with a large α has more smoothness, and is thus simpler to approximate. Figure 6.5 shows images with increasing smoothness.

6.4.1 1-D Fourier Approximation

A 1-D signal $f \in L^2([0, 1])$ is associated to a 1-periodic function $f(t+1) = f(t)$ defined for $t \in \mathbb{R}$.



Figure 6.10: Comparison of approximation errors for different bases using the same number $M = N/50$ of coefficients.

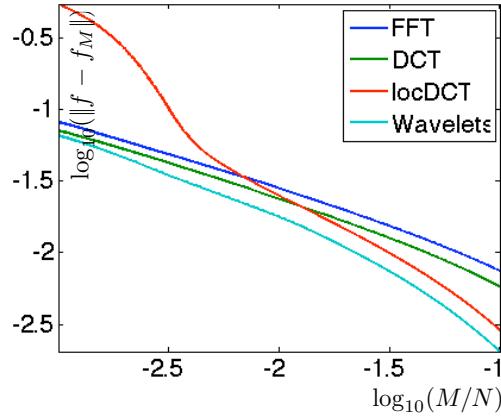


Figure 6.11: Comparison of approximation error decay for different bases.

Low pass approximation. We consider a linear Fourier approximation, that only retains low frequencies

$$f_M^{\text{lin}} = \sum_{m=-M/2}^{M/2} \langle f, e_m \rangle e_m$$

where we use the 1-D Fourier atoms

$$\forall n \in \mathbb{Z}, \quad e_m(t) \stackrel{\text{def.}}{=} e^{2i\pi m t}.$$

We note that f_M actually requires $M + 1$ Fourier atoms.

Figure 6.13 shows examples of such linear approximation for an increasing value of M . Since the original function f is singular (no derivative), this produces a large error and one observe ringing artifacts near singularities.

This low pass approximation corresponds to a filtering, since

$$f_M = \sum_{m=-M/2}^{M/2} \langle f, e_m \rangle e_m = f * h_M \quad \text{where} \quad \hat{h}_M \stackrel{\text{def.}}{=} 1_{[-M/2, M/2]}.$$

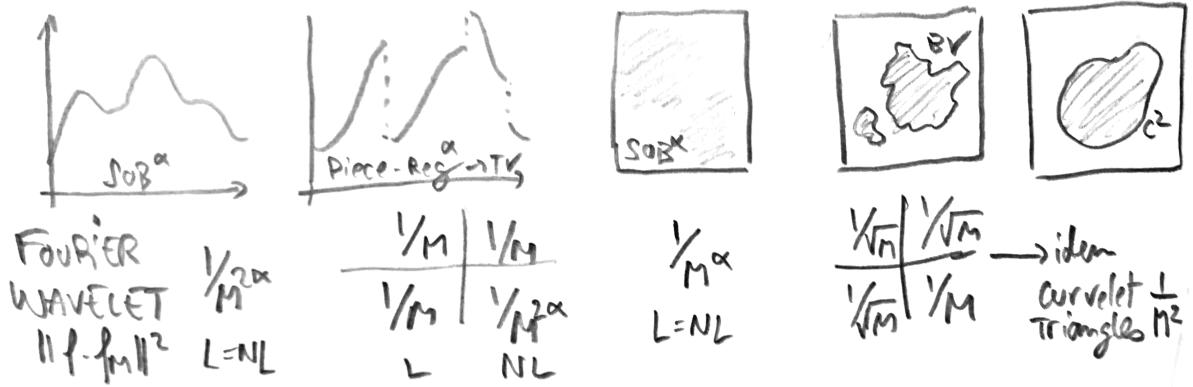


Figure 6.12: Summary of linear and non-linear approximation rate for different class of 1-D signals and images.

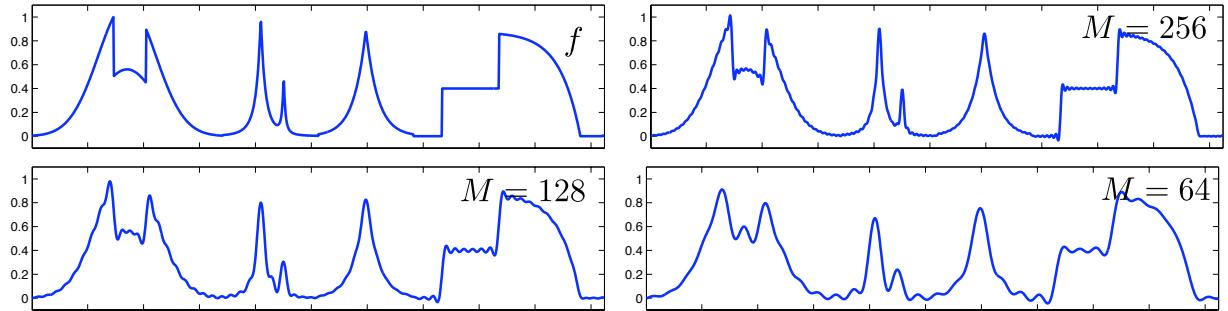


Figure 6.13: Fourier approximation of a signal.

Here, h_M is the so-called Dirichlet kernel (see also the chapter on Fourier).

The following proposition shows that this approximation decays fast for \mathcal{C}^α signals.

Proposition 15. *One has for $f \in \mathcal{C}^\alpha(\mathbb{R}/\mathbb{Z})$*

$$\|f - f_M^{\text{lin}}\| = O(M^{-2\alpha+1}).$$

Proof. Using integration by part, one shows that for a C^α function f in the smooth signal model (6.5), one has

$$|\langle f, e_m \rangle| \leq |2\pi m|^{-\alpha} \|f^{(\alpha)}\|.$$

This implies

$$\|f - f_M^{\text{lin}}\| = \sum_{|m| > M/2} |\langle f, e_m \rangle|^2 \leq \|f^{(\alpha)}\| \sum_{|m| > M/2} |2\pi m|^{-2\alpha} = O(M^{-2\alpha+1}).$$

□

We show next that a slightly modified proof gives a better decay (assuming $f^{(\alpha)}$ is in $L^2(\mathbb{R}/\mathbb{Z})$) and that this conclusion is valid for a larger class of Sobolev functions.

Proposition 16. *For signal f in the Sobolev model (6.8), one has*

$$\|f - f_M^{\text{lin}}\|^2 \leq C \|f^{(\alpha)}\|^2 M^{-2\alpha}. \quad (6.16)$$

Proof. One has

$$\|f^{(\alpha)}\|^2 = \sum_m |2\pi m|^{2\alpha} |\langle f, e_m \rangle|^2 \geq \sum_{|m|>M/2} |2\pi m|^{2\alpha} |\langle f, e_m \rangle|^2 \quad (6.17)$$

$$\geq (\pi M)^{2\alpha} \sum_{m>M/2} |\langle f, e_m \rangle|^2 \geq (\pi M)^{2\alpha} \|f - f_M^{\text{lin}}\|^2. \quad (6.18)$$

□

One can also shows that this asymptotic error decay is optimal, and that the non-linear approximation error in Fourier also decays like $O(M^{-2\alpha})$.

For a signal in the piecewise smooth model (6.9), such as the one shows in Figure 6.13, one only has a slow decay of the linear and non-linear approximation error

$$\|f - f_M\|^2 \leq C_f M^{-1} \quad (6.19)$$

and Fourier atoms are not anymore optimal for approximation.

6.4.2 Sobolev Images

This analysis caries over to images and higher dimensional datasets by considering a Sobolev functional (6.7) for $d > 1$.

The linear and non-linear approximation of an α -regular Sobolev image then satisfy

$$\|f - f_M\|^2 \leq C \|f^\alpha\|^2 M^{-\alpha}.$$

For d -dimensional data $f : [0, 1]^d \rightarrow \mathbb{R}$, one would have an error decay of $O(M^{-2\alpha/d})$.

For an image in the piecewise smooth model (6.10), the linear and non-linear error decays are slow,

$$\|f - f_M\|^2 \leq C_f M^{-1/2}, \quad (6.20)$$

and Fourier atoms are not anymore optimal for approximation.

6.5 Wavelet Approximation of Piecewise Smooth Functions

Wavelet approximation improve significantly over Fourier approximation to capture singularities. This is due to the localized support of wavelets.

6.5.1 Decay of Wavelet Coefficients

To efficiently approximate regular parts of signals and images, one uses wavelet with a large enough number p of vanishing moments

$$\forall k < p, \int \psi(x) x^k dx = 0.$$

This number p should be larger than the regularity α of the signal outside singularities (for instance jumps or kinks).

To quantify the approximation error decay for piecewise smooth signals (6.9) and images (6.10), one needs to treat differently wavelets that are in regular and singular areas. Figure 6.15 shows for a signal and an image the localization of singular and regular parts.

Proposition 17. *If f is C^α on $\text{supp}(\psi_{j,n})$, with $p \geq \alpha$, one has*

$$|\langle f, \psi_{j,n} \rangle| \leq C_f \|\psi\|_1 2^{j(\alpha+d/2)}. \quad (6.21)$$

In general, one always has for bounded f

$$|\langle f, \psi_{j,n} \rangle| \leq \|f\|_\infty \|\psi\|_1 2^{j\frac{d}{2}}.$$



Figure 6.14: Linear (top row) and non-linear (bottom row) Fourier approximation.

Proof. If f is C^α on $\text{supp}(\psi_{j,n})$, with $p \geq \alpha$, then one can perform a Taylor expansion of f around the point $2^j n$

$$f(x) = P(x - 2^j n) + R(x - 2^j n) = P(2^j t) + R(2^j t)$$

where $\deg(P) < \alpha$ and

$$|R(x)| \leq C_f \|x\|^\alpha.$$

One then bounds the wavelet coefficient

$$\langle f, \psi_{j,n} \rangle = \frac{1}{2^{j\frac{d}{2}}} \int f(x) \psi\left(\frac{x - 2^j n}{2^j}\right) dx = 2^{j\frac{d}{2}} \int R(2^j t) \psi(t) dt$$

where we have performed the change of variable $t = \frac{x - 2^j n}{2^j}$. This shows that (6.21) holds. Property (17) is straightforward. \square

6.5.2 1-D Piecewise Smooth Approximation

For 1-D signal in the piecewise regular model (6.9), large wavelets coefficients $\langle f, \psi_{j,n} \rangle$ are clustered around the singularities of the signal. We call $\mathcal{S} \subset [0, 1]$ the finite set of singular points.

Theorem 26. f is in the piecewise smooth signal model (6.9), the non-linear approximation error in wavelet obeys

$$\|f - f_M\|^2 = O(M^{-2\alpha}). \quad (6.22)$$

Proof. The proof is split in several parts.

Step 1. Coefficient segmentation. The singular support at scale 2^j is the set of coefficients corresponding to wavelets that are crossing a singularity

$$\mathcal{C}_j = \{n ; \text{supp}(\psi_{j,n}) \cap \mathcal{S} \neq \emptyset\} \quad (6.23)$$

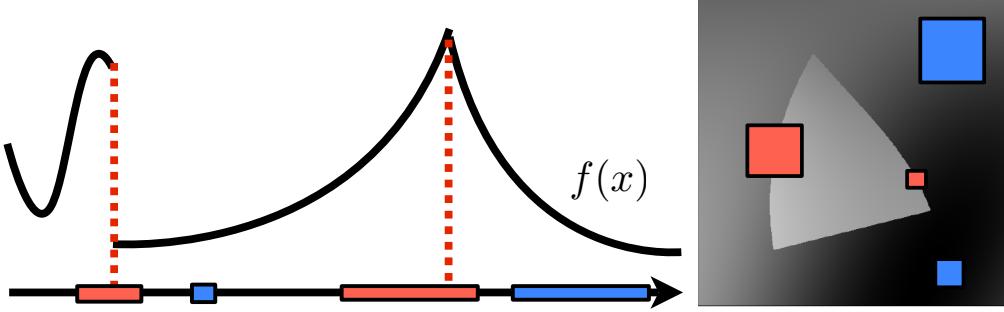


Figure 6.15: Singular pair of signals (left) and image (right).

It has a constant size because of the dyadic translation of wavelets

$$|\mathcal{C}_j| \leq K|\mathcal{S}| = \text{constant}.$$

Using (6.21) for $d = 1$, the decay of regular coefficients is bounded as

$$\forall n \in \mathcal{C}_j^c, \quad |\langle f, \psi_{j,n} \rangle| \leq C 2^{j(\alpha+1/2)}.$$

Using (17) for $d = 1$, the decay of singular coefficients is bounded as

$$\forall n \in \mathcal{C}_j, \quad |\langle f, \psi_{j,n} \rangle| \leq C 2^{j/2}.$$

Once a fixed threshold T is fixed to compute the non-linear approximation, one defines cut-off scales for regular and singular coefficients that depend on T

$$2^{j_1} = (T/C)^{\frac{1}{\alpha+1/2}} \quad \text{and} \quad 2^{j_2} = (T/C)^2.$$

Figure 6.16 shows a schematic segmentation of the set of wavelet coefficients into regular and singular parts, and also using the cut-off scales.

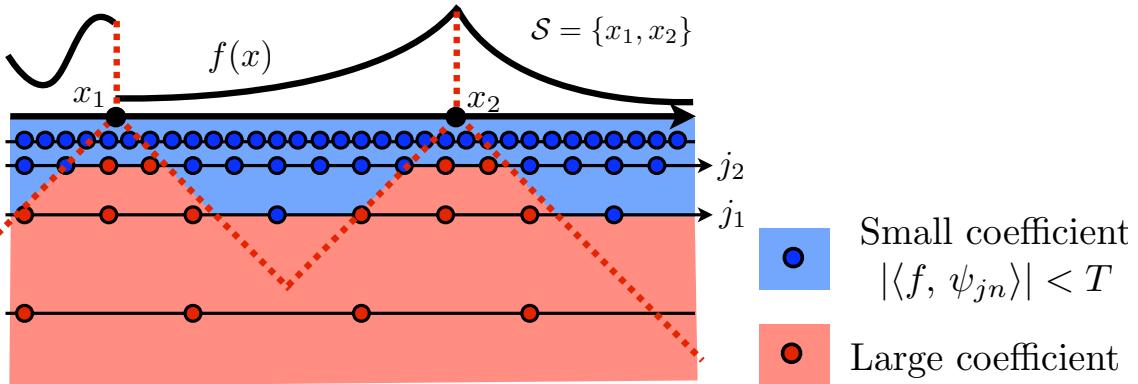


Figure 6.16: Segmentation of the wavelet coefficients into regular and singular parts.

Step 2. Counting the error. These cut-off scales allow us to define a hand-crafted approximation signal

$$\tilde{f}_M = \sum_{j \geq j_2} \sum_{n \in \mathcal{C}_j} \langle f, \psi_{j,n} \rangle \psi_{j,n} + \sum_{j \geq j_1} \sum_{n \in \mathcal{C}_j^c} \langle f, \psi_{j,n} \rangle \psi_{j,n}. \quad (6.24)$$

The approximation error generated by this M -term approximation \tilde{f}_M is larger than the best M -term approximation error, and hence

$$\|f - f_M\|^2 \leq \|f - \tilde{f}_M\|^2 \leq \sum_{j < j_2, n \in \mathcal{C}_j} |\langle f, \psi_{j,n} \rangle|^2 + \sum_{j < j_1, n \in \mathcal{C}_j^c} |\langle f, \psi_{j,n} \rangle|^2 \quad (6.25)$$

$$\leq \sum_{j < j_2} (K|\mathcal{S}|) \times C^2 2^j + \sum_{j < j_1} 2^{-j} \times C^2 2^{j(2\alpha+1)} \quad (6.26)$$

$$= O(2^{j_2} + 2^{2\alpha j_1}) = O(T^2 + T^{\frac{2\alpha}{\alpha+1/2}}) = O(T^{\frac{2\alpha}{\alpha+1/2}}). \quad (6.27)$$

Step 3. Counting the number of measurements. The number of coefficients needed to build the approximating signal \tilde{f}_M is

$$M \leq \sum_{j \geq j_2} |\mathcal{C}_j| + \sum_{j \geq j_1} |\mathcal{C}_j^c| \leq \sum_{j \geq j_2} K|\mathcal{S}| + \sum_{j \geq j_1} 2^{-j} \quad (6.28)$$

$$= O(|\log(T)| + T^{\frac{-1}{\alpha+1/2}}) = O(T^{\frac{-1}{\alpha+1/2}}). \quad (6.29)$$

Step 3. Putting everything together. Putting equations (6.25) and (6.28) together gives the desired result. \square

This theorem improves significantly over the $O(M^{-1})$ decay of Fourier approximation (6.19). Furthermore, this decay is the same as the error decay of uniformly smooth signal (6.16). In 1-D, wavelet approximations do not “see” the singularities. The error decay (6.22) can be shown to be asymptotically optimal.

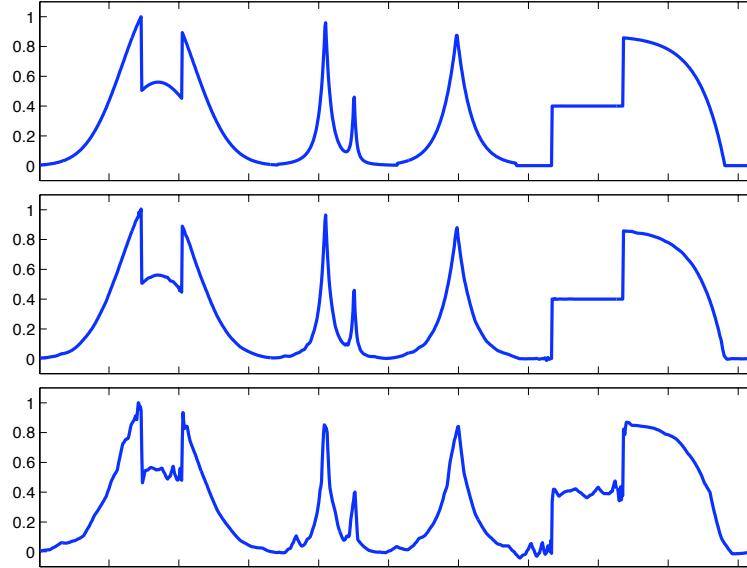


Figure 6.17: 1-D wavelet approximation.

Figure 6.17 shows examples of wavelet approximation of singular signals.

6.5.3 2-D Piecewise Smooth Approximation

We now give the equivalent of Theorem 26 but for 2-D functions.

Theorem 27. f is in the piecewise smooth signal model (6.10), the non-linear approximation error in wavelet obeys

$$\|f - f_M\|^2 = O(M^{-1}). \quad (6.30)$$

Proof. For an image in the piecewise smooth model (6.10), we define the singular support \mathcal{C}_j as in (6.23). The major difference with the 1-D setting, is that for 2-D images, the size of the singular support grows when the scale 2^j goes to zero

$$|\mathcal{C}_j^\omega| \leq 2^{-j} K |\mathcal{S}|,$$

where $|\mathcal{S}|$ is the perimeter of the singular curves \mathcal{S} , and $\omega \in \{V, H, D\}$ is the wavelet orientation. Using (6.21) for $d = 2$, the decay of regular coefficients is bounded as

$$\forall n \in (\mathcal{C}_j^\omega)^c, \quad |\langle f, \psi_{j,n}^\omega \rangle| \leq C 2^{j(\alpha+1)}.$$

Using (17) for $d = 2$, the decay of singular coefficients is bounded as

$$\forall n \in \mathcal{C}_j^\omega, \quad |\langle f, \psi_{j,n}^\omega \rangle| \leq C 2^j.$$

After fixing T , the cut-off scales are defined as

$$2^{j_1} = (T/C)^{\frac{1}{\alpha+1}} \quad \text{and} \quad 2^{j_2} = T/C.$$

We define similarly to (6.24) a hand-made approximation. Similarly to (6.25), we bound the approximation error as

$$\|f - f_M\|^2 \leq \|f - \tilde{f}_M\|^2 = O(2^{j_2} + 2^{2\alpha j_1}) = O(T + T^{\frac{2\alpha}{\alpha+1}}) = O(T)$$

and the number of coefficients as

$$M = O(T^{-1} + T^{\frac{-1}{\alpha+1}}) = O(T^{-1}).$$

This leads to the announced decay of the non-linear wavelet approximation error. \square

This improves significantly over the $O(M^{-1/2})$ decay of Fourier approximation (6.20). This result is however deceiving, since it does not take advantage of the C^α regularity of the image outside the edge curves.

This error decay is still valid for the more general model of images with bounded variations (6.11). One can shows that wavelets are asymptotically optimal to approximate images with bounded variations.



Figure 6.18: 2-D wavelet approximation.

Figure 6.18 shows wavelet approximations of a bounded variation image.

6.6 Cartoon Images Approximation

The square support of wavelet makes them inefficient to approximate geometric images (6.13), whose edges are more regular than the level set of bounded variation images (6.11), which are only assumed to be of finite length.

6.6.1 Wavelet Approximation of Cartoon Images

Result (6.30) shows that wavelet approximation of images in the cartoon models (6.13) decays at least like $O(M^{-1})$. One can show that simple cartoon images like $f = 1_{\Omega}$ where Ω is a disk reach this low decay speed. This is because the square support of wavelets forbid them to take advantage of the regularity of edge curves. The approximation error for the smoothed cartoon model (6.14) is also slow if the width of the blurring kernel is small with respect to the number M of coefficients.

Figure 6.19 shows that many large coefficients are located near edge curves, and retaining only a small number leads to a bad approximation with visually unpleasant artifacts.

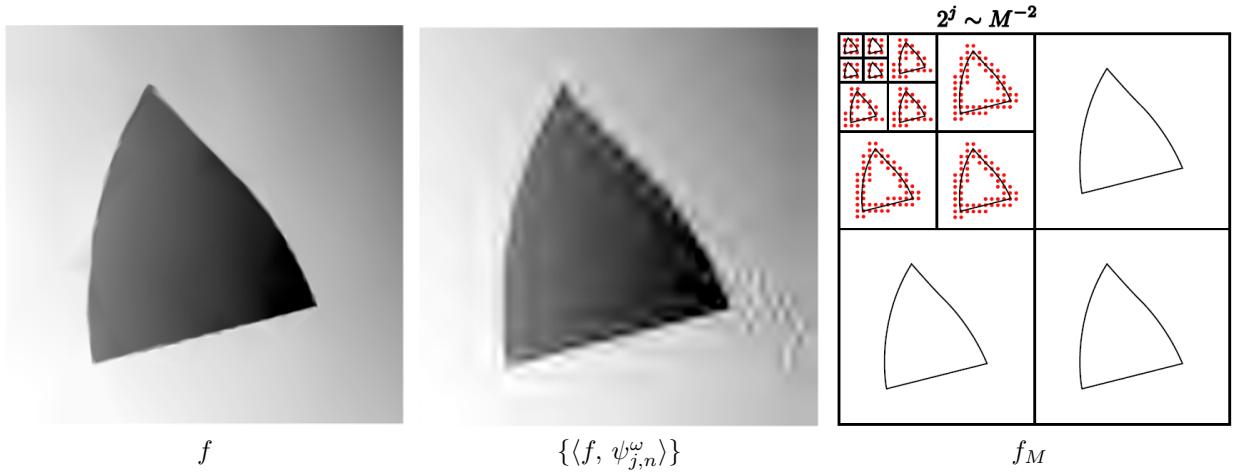


Figure 6.19: Wavelet approximation of a cartoon image.

6.6.2 Finite Element Approximation

To improve over the wavelet approximation, one can design an adapted triangulation that is highly anisotropic near edges. Figure 6.20 shows an example of such a triangulation.

A triangulation is obtain by sampling M points over the image domain $[0, 1]^2$ and then connecting them using triangles. One then defines a piecewise linear interpolation \tilde{f}_M over these triangles.

As shown in Figure 6.21, an efficient approximation of a C^2 -cartoon image (6.13) for $\alpha = 2$ is obtained by seeding $\approx M/2$ approximately equilateral triangles of width $\approx M^{-1/2}$ in the areas where the image is regular. Near the edges, using the C^2 regularity of the singular curve, one can seed $\approx M/2$ anisotropic triangles of length M^{-1} and width $\approx M^{-1/2}$. One can show that such an adaptive triangulation leads to an approximation error

$$\|f - f_M\|^2 = O(M^{-2}), \quad (6.31)$$

which improves over the wavelet approximation error decay (6.30).

This scheme is however difficult to implement in practice, since the edge curves are not known and difficult to find. This is in particular the case for smooth cartoon images when the smoothing kernel h is unknown.

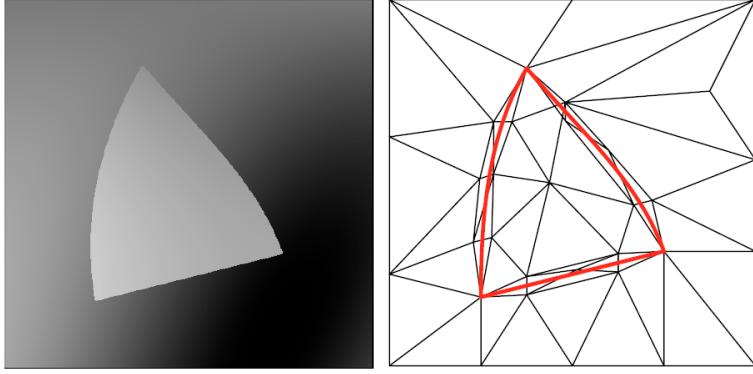


Figure 6.20: Left: cartoon image, right: adaptive triangulation.

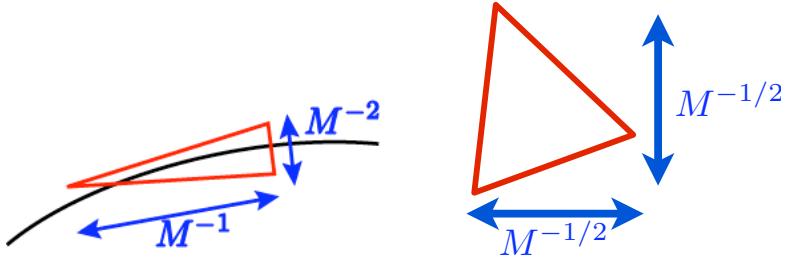


Figure 6.21: Aspect ratio of triangle away from edges (left) and near an edge (right).

There is currently no known algorithm that can automatically produces the error decay (6.31). One thus has to use heuristics and greedy algorithm to find the location of the sampling points and computes the triangles. Figure (6.22) shows an example of compression using such a greedy seeding algorithm, that works well in practice.

6.6.3 Curvelets Approximation

Instead of using an adaptive approximation scheme such as finite elements, one can replace the wavelet basis by a set of oriented anisotropic atoms. The curvelet frame was proposed by Candès and Donoho for this purpose [6].

Curvelets. The curvelet construction starts from a curvelet function c that is oriented along the horizontal direction, and perform stretching

$$c_{2^j}(x_1, x_2) \approx 2^{-3j/4} c(2^{-j/2}x_1, 2^{-j}x_2),$$

translation and rotation

$$c_{2^j,u}^\theta(x_1, x_2) = c_{2^j}(R_\theta(x - u))$$

where R_θ is the rotation of angle θ .

The atoms $c_{2^j,u}^\theta$ is located near u , with an orientation θ , and has an aspect ratio “width \approx length²”, which is the same aspect used to build an adaptive finite element approximation. This aspect ratio is essential to capture the anisotropic regularity near edges for images in the cartoon model (6.13) for $\alpha = 2$.

Figure 6.24 shows the spacial and frequency localization of curvelets.



Figure 6.22: Comparison of adaptive triangulation and JPEG-2000, with the same number of bits.

Parameter discretization. To build an image representation, one need to sample the u and θ parameter. To maintain a stable representation, the sub-sampling of the angles depends on the scale

$$\forall 0 \leq k < 2^{\lceil j/2 \rceil + 2}, \quad \theta_k^{(j)} = k\pi 2^{\lceil j/2 \rceil - 1}$$

and the spacial grid depends on the scale and on the angles

$$\forall m = (m_1, m_2) \in \mathbb{Z}^2, \quad u_m^{(j,\theta)} = R_\theta(2^{j/2}m_1, 2^j m_2).$$

Figure 6.24 shows this sampling pattern.

Curvelet tight frame. This sampling leads to a stable redundant family

$$C_{j,m,k}(x) = c_{2^j,u}^\theta(x) \quad \text{where} \quad \theta = \theta_k^{(j)} \quad \text{and} \quad u = u_m^{(j,\theta)},$$

that obeys a conservation of energy

$$\|f\|^2 = \sum_{j \in \mathbb{Z}} \sum_{k=0}^{2^{-\lceil j/2 \rceil + 2}} \sum_{m \in \mathbb{Z}^2} |\langle f, C_{j,m,k} \rangle|^2$$

and a reconstruction formula

$$f = \sum_{j \in \mathbb{Z}} \sum_{k=0}^{2^{-\lceil j/2 \rceil + 2}} \sum_{m \in \mathbb{Z}^2} \langle f, C_{j,m,k} \rangle C_{j,m,k}$$

that extends the properties of orthogonal basis (tight frame), although the representation is redundant (the atoms are not orthogonal).

A numerical implementation of this tight frame also defines a discrete tight frame for image of N pixels, that is made of $\approx 5N$ atoms [7].

Curvelet approximation. A non-linear M -term approximation in curvelets is defined as

$$f_M = \sum_{|\langle f, C_{j,m,k} \rangle| > T} \langle f, C_{j,m,k} \rangle C_{j,m,k}$$

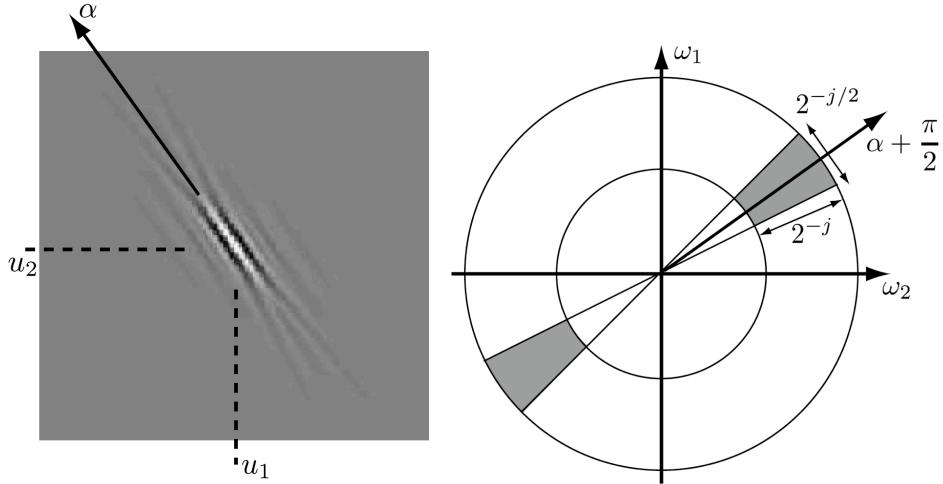


Figure 6.23: Left: a curvelet c_m , right: its Fourier transform localization.

where T is a threshold that depends on M . One should note that f_M is not necessarily the best M -term curvelet approximation since the curvelet frame is not orthogonal.

For position $u_m^{(j,\theta)}$ that are far away from an edges, the vanishing moments of the curvelets create a small coefficient $\langle f, C_{j,m,k} \rangle$. If $u_m^{(j,\theta)}$ is close to an edge curve whose tangent has direction $\tilde{\theta}$, then the coefficient $\langle f, C_{j,m,k} \rangle$ decays very fast to zero when $|\theta - \tilde{\theta}|$ increases. Figure 6.25 shows the principle of this curvelet approximation, and compares it with directional wavelets that have a square support.

Using these two properties together with the sparse sampling of the curvelet in space and orientation leads to the following approximation error decay

$$\|f - f_M\|^2 = O(\log^3(M)M^{-2})$$

for image in the cartoon model (6.13) for $\alpha = 2$. This is close to the decay of adaptive triangulations (6.31), but this time one computes f_M with a fast $O(N \log(N))$ algorithm for an image of N pixels.

In practice, the redundancy of the curvelet frame makes it not suitable for image compression. Its efficiency is however useful for denoising purpose, where it can improve over wavelet to denoise geometric images and textures, see Figure 6.26. The result is obtained by using a thresholding denoiser as detailed in Section 8.3.1.

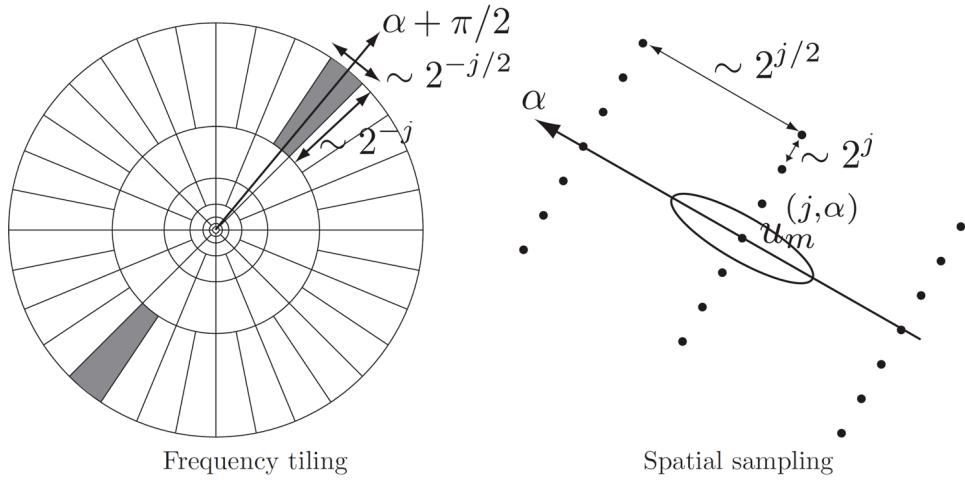


Figure 6.24: Sampling pattern for the curvelet positions.

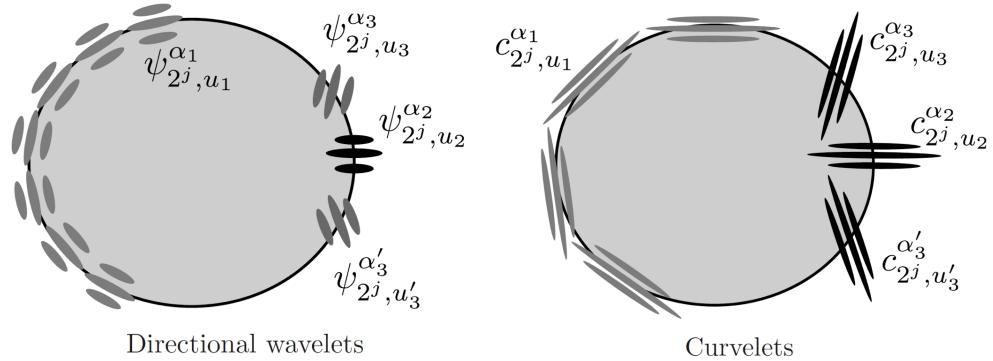


Figure 6.25: Comparison of the principle of wavelets (left) and curvelet (right) approximations of a cartoon image.

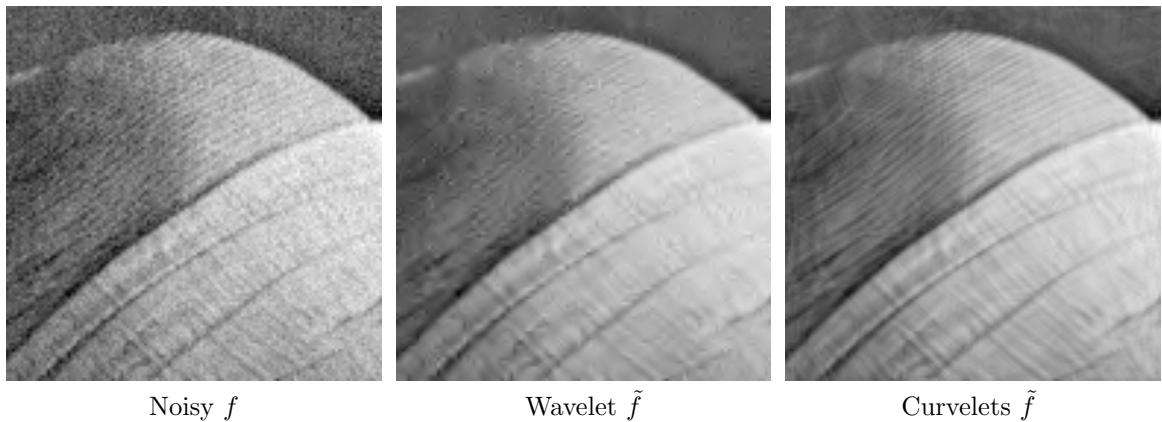


Figure 6.26: Comparison of wavelets (translation invariant) and curvelet denoising.

Chapter 7

Compression

7.1 Transform Coding

7.1.1 Coding

State of the art compression schemes correspond to transform coders, that code quantized coefficients in an ortho-basis. They first computes the coefficients of the decomposition of the signal into an well-chosen basis (for instance wavelets)

$$a_m = \langle f, \psi_m \rangle \in \mathbb{R}.$$

Quantization corresponds to rounding the coefficients to an integer using a step size $T > 0$

$$q_m = Q_T(a_m) \in \mathbb{Z} \quad \text{where} \quad Q_T(x) = \text{sign}(x) \left\lfloor \frac{|x|}{T} \right\rfloor.$$

We note that this quantizer has a twice larger zero bin, so that coefficients in $[-T, T]$ are set to zero.

This quantizer nonlinearity should be compared to the hard thresholding nonlinearity (11.4) used to perform non-linear approximation. The quantizer not only set to zero small coefficients that are smaller than T in magnitude, it also modifies larger coefficients by rounding, see Figure 7.1.

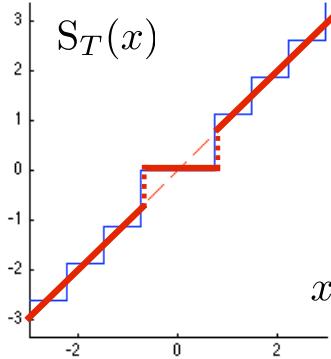


Figure 7.1: Thresholding and quantization non-linearity mappings.

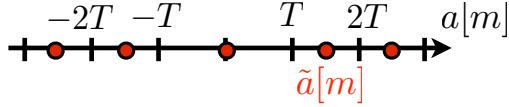
The resulting integer values $(q_m)_m$ are stored into a binary file of length R , which corresponds to a number of bits. Sections 7.1.3 and 7.2 detail two different approach to perform this transformation from integer to bits. The goal is to reduce as much as possible the number R of bits.

7.1.2 De-coding

The decoder retrieves the quantized coefficients q_m from the binary file, and dequantizes the coefficients using

$$\tilde{a}_m = \text{sign}(q_m) \left(|q_m| + \frac{1}{2} \right) T. \quad (7.1)$$

This corresponds to retrieving the value from quantization at the center of quantization bins:



The compressed-decompressed image is then reconstructed as

$$\mathcal{Q}_T(f) \stackrel{\text{def.}}{=} \sum_{m \in I_T} \tilde{a}_m \psi_m = \sum_{m \in I_T} Q_T(\langle f, \psi_m \rangle) \psi_m,$$

thus producing a decompression error $\|f - \mathcal{Q}_T(f)\|$.

This decompression reconstruction (7.1.2) should be compared with the non-linear approximation formula (6.3). One necessarily has $\|f - f_M\| \leq \|f - \mathcal{Q}_T(f)\|$, but in practice these two errors have comparable magnitudes.

Proposition 18. *One has*

$$\|f - \mathcal{Q}_T(f)\|^2 \leq \|f - f_M\|^2 + MT^2/4 \quad \text{where } M = \# \{m ; \tilde{a}_m \neq 0\}. \quad (7.2)$$

Proof. Indeed, the de-quantization formula (7.1) implies that for $|a_m| > T$,

$$|a_m - \tilde{a}_m| \leq \frac{T}{2}.$$

One thus has

$$\|f - \mathcal{Q}_T(f)\|^2 = \sum_m (a_m - \tilde{a}_m)^2 \leq \sum_{|a_m| < T} |a_m|^2 + \sum_{|a_m| \geq T} \left(\frac{T}{2} \right)^2,$$

which leads to the desired bound. \square

7.1.3 Support Coding

To measure how large is the additional error term $MT^2/4$ in (7.2), one needs to choose a method to store the quantized coefficients q_m into a file.

For aggressive compression scenario, where R and M are small with respect to the size N of the image, the support

$$I_M = \{m ; \tilde{a}_m \neq 0\}$$

is highly sparse. It thus make sense to code first this support and then to code the actual value $q_m \neq 0$ for $m \in I_M$.

The remaining of this section proves the following theorem.

Theorem 28. *We assume $\|f - f_M\|^2 \sim M^{-\alpha}$ where f_M is the M -term non-linear approximation of f in $\{\psi_m\}_m$. We also assume that the required number of discrete samples N used can be bounded polynomially with N (see (7.7) bellow for more details). Then for all T there exists a coding strategy of $\mathcal{Q}_T(f)$ using $R = R(T)$ bits such that*

$$\|f - \mathcal{Q}_T(f)\|^2 = O(\log^\alpha(R) R^{-\alpha}).$$

Proof. This proof is split in several parts.

Signals constraints. First, let us notice that, thanks to Proposition 14 the error decay hypothesis $\|f - f_M\|^2 \sim M^{-\alpha}$ is equivalent to imposing a fast decay of the ordered coefficients d_m defined in (6.1) satisfies

$$d_m \sim m^{-\frac{\alpha+1}{2}} \implies T \sim M^{-\frac{\alpha+1}{2}}. \quad (7.3)$$

Thanks to Proposition 18, one thus has

$$\|f - f_M\|^2 \leq \|f - Q_T(f)\|^2 \leq \|f - f_M\|^2 + MT^2/4 \sim M^{-\alpha} \sim \|f - f_M\|^2, \quad (7.4)$$

which shows that the compression error is comparable with the approximation error.

Discrete computation and scaling of N . For the compression from the discrete signals to be the same as a compression of a continuous signal, we impose that N is large enough so that

$$\forall m \geq N, \quad |\langle f, \psi_m \rangle| < T \quad (7.5)$$

so that the coefficients not quantized to 0 are contained within the set $\{\langle f, \psi_m \rangle\}_{0 \leq m < N}$ of the N computed coefficients. For instance, if f is bounded and one considers a wavelet basis, (17) ensures that $|\langle f, \psi_m \rangle| = O(2^{-jd/2})$, and thus choosing $T \sim 2^{-j_{\max}d/2}$ ensures (7.5) where $N = 2^{-j_{\max}d}$, i.e.

$$N = O(1/T^2). \quad (7.6)$$

The other hypothesis beyond (7.3) of Theorem 28 is that the number N of required discrete samples is not too large, and in particular, that there is a polynomial grows of N with respect to the number M of coefficients to code

$$N = O(M^\beta) \quad (7.7)$$

for some $\beta > 0$. For instance, under the decay condition (7.3) and the worse case scaling (7.6) for bounded function in a wavelet basis,

$$N = O(1/T^2) \quad \text{and} \quad T \sim M^{-\frac{\alpha+1}{2}} \implies N = O(M^{\alpha+1}).$$

Support coding. One can simply encode the M indexes of each element $m \in I_M \subset \{1, \dots, N\}$ using $\log_2(N)$ bits, so that the total number of bits for the support I_M is

$$R_{\text{ind}} = M \log_2(N) = O(M \log_2(M)) \quad (7.8)$$

where we used (7.3).

Values coding. The quantized values satisfy $q_m \in \{-A, \dots, A\}$, with

$$A \leq \frac{1}{T} \max_m |\langle f, \psi_m \rangle| = O(T^{-1}),$$

so one can code them using a number of bits

$$R_{\text{val}} = O(M |\log_2(T)|) = O(M \log_2(M)) \quad (7.9)$$

where we have used hypothesis (7.3) that implies $|\log_2(T)| \sim \log_2(M)$.

Total number of bits. Putting (7.8) and (7.9) together, the total number of bits for this support coding approach is thus

$$R = R_{\text{ind}} + R_{\text{val}} = O(M \log_2(M)). \quad (7.10)$$

The function $\varphi(m) = m \log_2(m)$ is strictly increasing, one can invert it, and we now show that $\varphi^{-1}(r) = r / \log(r) + o(1)$ for large r . Indeed, writing $r = \varphi(m)$

$$m = \frac{r}{\log_2(m)} = \frac{r}{\log_2(r) - \log_2 \log_2(m)} \sim \frac{r}{\log_2(r)}$$

where we used the fact that since $m \leq r$, $\log_2 \log_2(m) = o(\log_2(r))$. Inverting this relationship (7.10) thus proves that

$$M \geq C \frac{R}{\log_2(R)} \quad (7.11)$$

for some constant C . Using (7.4) and (7.11), one thus finally arrives to

$$\|f - \mathcal{Q}_T(f)\|^2 \sim M^{-\alpha} = O(\log^\alpha(R) R^{-\alpha}).$$

□

This theorem shows the importance of the study of non-linear approximation, and in particular the design of bases that are efficient for approximation of a given signal model Θ . Let us also insist on the fact that a practical compression algorithm is only capable of dealing with discrete signals of size N . We thus consider that the algorithm has access to N inner products $\{\langle f, \psi_m \rangle\}_{1 \leq m < N}$ that are computed using a decomposition algorithm from a discretized signal or image of size N . For instance, Section ?? details a discrete wavelet transform, and introduces a compatibility condition (??) on the sampling operator for this inner product computation to be possible from discrete data.

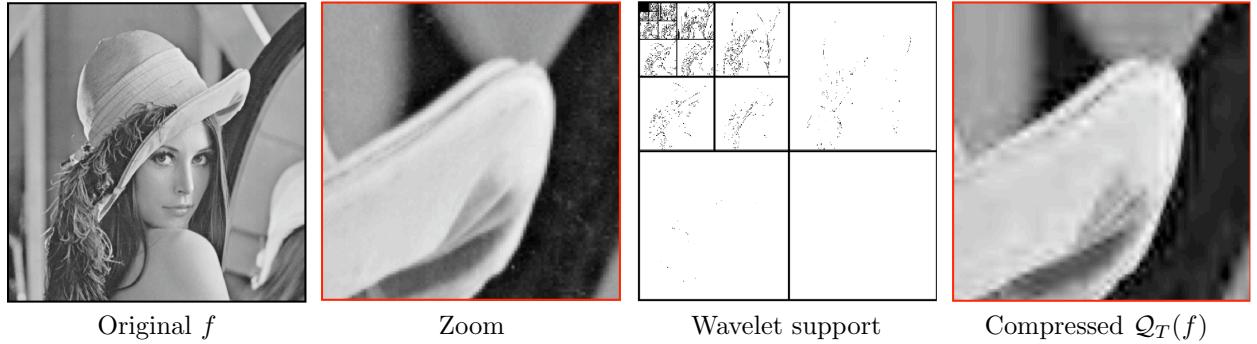


Figure 7.2: Image compression using wavelet support coding.

7.2 Entropic Coding

To further reduce the file size R (in bits), one can use an entropic coder to transform the integer values q_m into bits. Such coding scheme makes use of the statistical redundancy of the quantized values, that have a large number of zero entries and very few large entries. The theory of coding was formalized by Shannon [38].

We refer to Section 1.3 for the theoretical foundation associated to what we now describe.

Probabilistic modeling. The quantized coefficients $q_m \in \{-A, \dots, A\}$ are assumed to take values in an alphabet of $Q = 2A + 1$ elements. A coding scheme performs the transformation

$$\{q_m\}_m \longmapsto \{0, 1, 1, \dots, 0, 1\} \in \{0, 1\}^R.$$

To reduce the average value of R , one makes use of a statistical model, which assumes that the q_m are drawn independently at random from a known probability distribution

$$\mathbb{P}(q_m = i) = p_i \in [0, 1].$$

Huffman code. A Huffman code is a code with variable length, since it performs a mapping from symbols to binary strings

$$q_m = i \in \{-A, \dots, A\} \longmapsto c_i \in \{0, 1\}^{|c_i|}$$

where $|c_i|$ is the length of the binary code word c_i , that should be larger if p_i is small. A Huffman tree algorithm is able to build a code such that

$$|c_i| \leq \lceil \log_2(p_i) \rceil$$

so that

$$R \leq (\mathcal{E}(p) + 1)N$$

where \mathcal{E} is the entropy of the distribution, defined as

$$\mathcal{E}(p) = - \sum_i p_i \log_2(p_i).$$

Figure 7.3 shows different probability distributions. The entropy is small for highly sparse distribution. Wavelet coefficients of natural images tend to have a low entropy because many coefficients are small.

The Huffman scheme codes symbols independently, leading to a sub-optimal code if some of the p_i are large, which is usually the case for wavelet coefficients. One usually prefers arithmetic coding schemes, that codes groups of symbols, and are able to get close to the entropy bound $R \approx \mathcal{E}(p)N$ for large N .

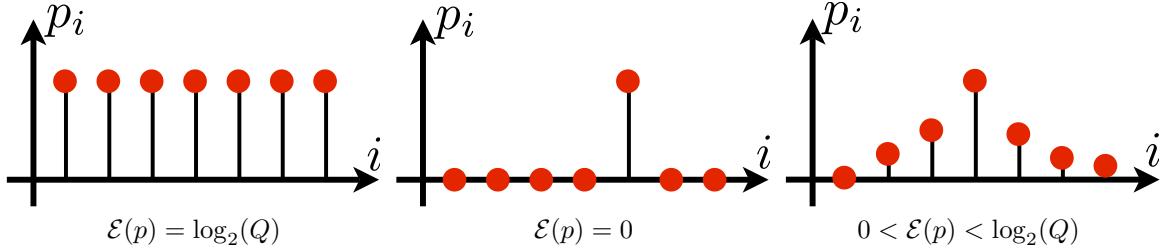


Figure 7.3: Three different probability distributions.

7.3 JPEG-2000

JPEG-2000 is the latest still image compression standard. It corresponds to a wavelet transform coder that performs a clever adaptive entropy coding that makes use of the statistical redundancy of wavelet coefficients of natural images. The wavelet transform is not orthogonal, it is a symmetric 7/9 biorthogonal, with symmetric boundary condition and a lifting implementation. This transform is however close to orthogonality, so that the previous discussion about orthogonal approximation and coding is still relevant.

Figure 7.4 shows an overview of JPEG-2000 architecture. Figure 7.5 shows a comparison between JPEG and JPEG-2000 compressors. JPEG is based on a local DCT transform, and suffers from blocking artifacts at low bit rates, which is not the case of JPEG-2000. This new standard also comes with several important features, such as regions of interest, which allows to refine the coding in some specific parts of the image.

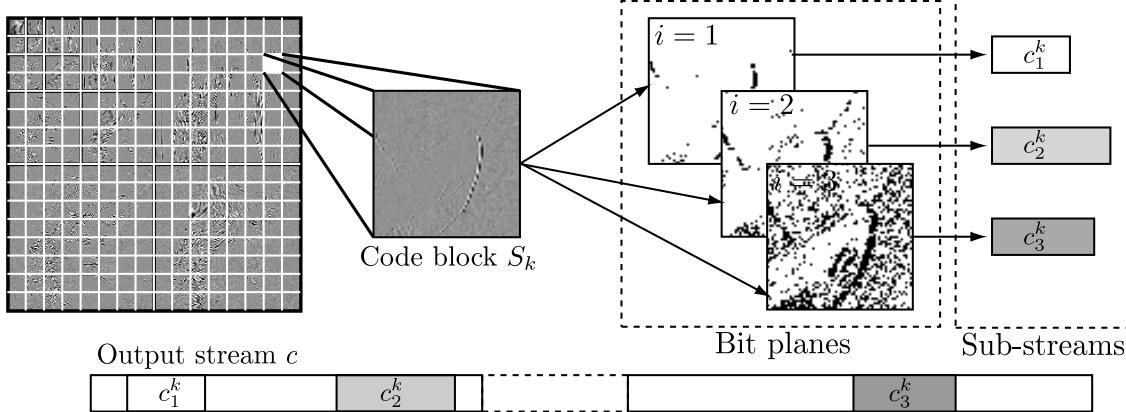


Figure 7.4: JPEG-2000 coding architecture.

Dyadic quantization. The wavelet coefficients are quantized with a varying quantization step $T_i = 2^{-i}T_0$. This allows one to progressively increase the precision of the coded coefficients. For each i , a bit plane coding pass produces new bits to refine the value of the coefficients when i increases.

Steam packing. The bits obtained using the bit plane pass with quantizer $T_i = 2^{-i}T_0$ are entropy coded using a contextual coder. This coder processes square blocks S_k of coefficients. This local coding enhances the parallelization of the method. This local block coding produces a bit stream c_i^k , and these streams are optimally packed into the final coded file to reduce the distortion $\|f - \mathcal{Q}_T(f)\|$ for almost every possible number R of bits. This stream packing ensures scalability of the output bit stream. It means that one can receive only the R first bits of a large coded file and get a low resolution decoded image $\mathcal{Q}_T(f)$ that has an almost minimal distortion $\|f - \mathcal{Q}_T(f)\|$.

Bit plane coding pass. For each threshold T_i , for each scale and orientation $(j, \omega \in \{V, H, D\})$, for each coefficient location $n \in S_k$, JPEG-2000 coder encodes several bit reflecting the value of the wavelet coefficient $d_j^\omega[n]$. In the following we drop the dependancy on (j, ω) for simplicity.

- If $d_j^\omega[n] < T_{i-1}$, the coefficient was not significant at bit-plane $i-1$. It thus encodes a significance bit $b_i^1[n]$ to tell whether $d_j^\omega[n] \geq T_i$ or not.
- If $b_i^1[n] = 1$, meaning that the coefficient has became significant, it codes its sign as a bit $b_i^2[n]$.
- For every position n that was previously significant, meaning $d_j^\omega[n] \geq T_{i-1}$, it codes a value refinement bit $b_i^3[n]$ to tell whether $d_j^\omega[n] \geq T_i$ or not.

Contextual coder. The final bits streams c_i^k are computed from the produced bits $\{b_i^s[n]\}_{s=1}^3$ for $n \in S_k$ using a contextual coder. The contextual coding makes use of spacial redundancies in wavelet coefficients, especially near edges and geometric singularities that create clusters of large coefficients. The coefficients $n \in S_k$ are traversed in zig-zag order as shown on Figure 7.6.

For each coefficient location $n \in S_k$, the context value $v_i^s[n]$ of the bit $b_i^s[n]$ to code at position x is an integer computed over a 3×3 window

$$w_n = \{(n_1 + \varepsilon_1, n_2 + \varepsilon_2)\}_{\varepsilon_i=\pm 1}.$$

This local context $v_i^s[n]$ integrates in a complicated way the previous bit plane values $\{b_{i-1}^s[\tilde{n}]\}_{\tilde{n} \in w_n}$, and neighboring bits at plane $\{b_i^s[\tilde{n}]\}_{\tilde{n} \in w_n, \tilde{n}}$ coded that have already been coded.

The bit value $b_i^s[n]$ is then coded with an arithmetic coding by making use of the conditional probability distribution $\mathbb{P}(b_i^s[n]|v_i^s[n])$. The choice made for the computation $v_i^s[n]$ allows to reduce significantly the



Figure 7.5: Comparison of JPEG (left) and JPEG-2000 (right) coding.

entropy of this conditional probability condition with respect to the original distribution $\mathbb{P}(b_i^s[n])$, thus reducing the overall number of bits.

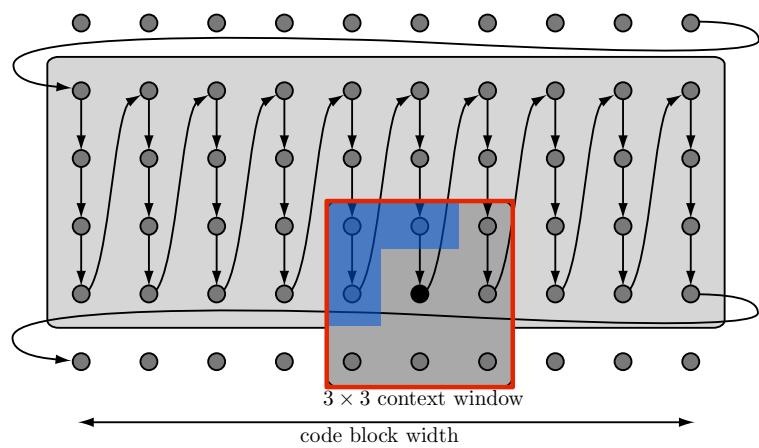


Figure 7.6: Coding order and context for JPEG-2000 coding.

Chapter 8

Denoising

Together with compression, denoising is the most important processing application, that is pervasive in almost any signal or image processing pipeline. Indeed, data acquisition always comes with some kind of noise, so modeling this noise and removing it efficiently is crucial.

8.1 Noise Modeling

8.1.1 Noise in Images

Image acquisition devices always produce some noise. Figure 8.1 shows images produced by different hardware, where the regularity of the underlying signal and the statistics of the noise is very different.

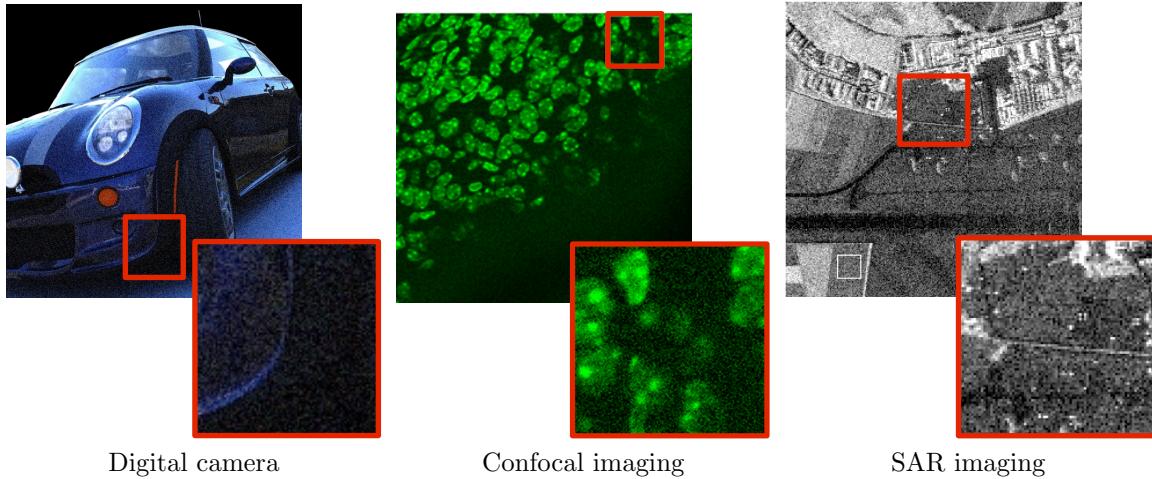


Figure 8.1: Example of noise in different imaging device.

One should thus model both the acquisition process and the statistics of the noise to fit the imaging process. Then one should also model the regularity and geometry of the clean signal to choose a basis adapted to its representation. This chapter describes how thresholding methods can be used to perform denoising in some specific situations where the noise statistics are close to being Gaussian and the mixing operator is a sum or can be approximated by a sum.

Since noise perturbs discrete measurements acquired by some hardware, in the following, we consider only finite dimensional signal $f \in \mathbb{C}^N$.

8.1.2 Image Formation

Figure 8.2 shows an idealized view of the image formation process, that mixes a clean image f_0 with a noise w to obtain noisy observations $f = f_0 \oplus w$, where \oplus might for instance be a sum or a multiplication.

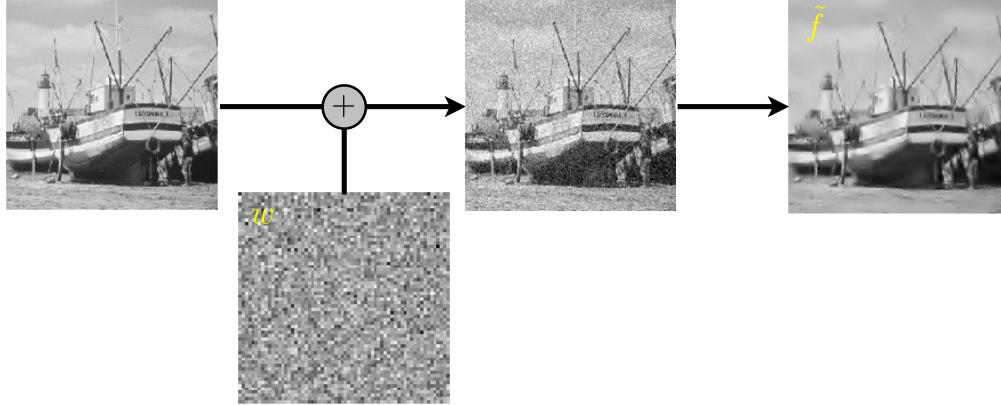


Figure 8.2: Image formation with noise modeling and denoising pipeline.

Statistical modeling considers w as a random vector with known distribution, while numerical computation are usually done on a single realization of this random vector, still denoted as w .

Additive Noise. The simplest model for such image formation consists in assuming that it is an additive perturbation of a clean signal f_0

$$f = f_0 + w$$

where w is the noise residual. Statistical noise modeling assume that w is a random vector, and in practice one only observes a realization of this vector. This modeling thus implies that the image f to be processed is also a random vector. Figure 8.3 and 8.4 show examples of noise addition to a clean signal and a clean image.

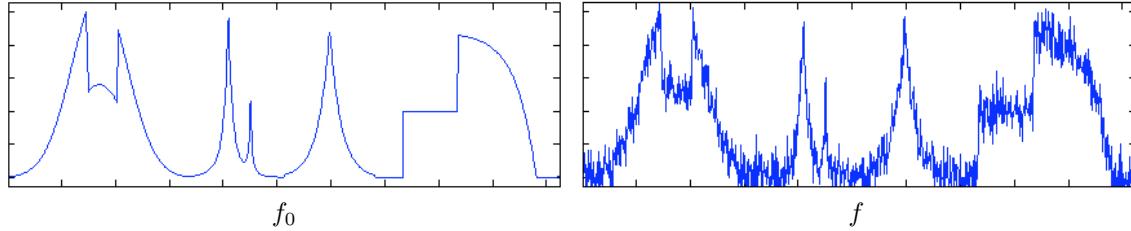


Figure 8.3: 1-D additive noise example.

The simplest noise model assumes that each entry w_n of the noise is a Gaussian random variable of variance σ^2 , and that the w_n are independent, i.e. $w \sim \mathcal{N}(0, \text{Id}_N)$. This is the white noise model.

Depending on the image acquisition device, one should consider different noise distributions, such as for instance uniform noise $w_n \in [-a, a]$ or Impulse noise

$$\mathbb{P}(w_n = x) \propto e^{-|x/\sigma|^\alpha} \quad \text{where} \quad \alpha < 2$$

In many situations, the noise perturbation is not additive, and for instance its intensity might depend on the intensity of the signal. This is the case with Poisson and multiplicative noises considered in Section 8.4.

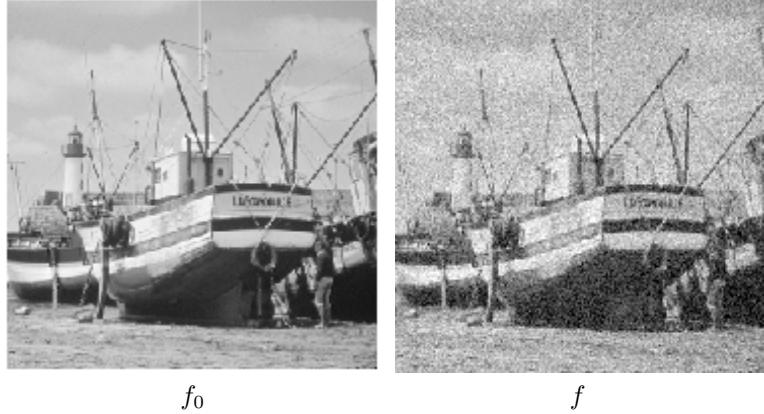


Figure 8.4: 2-D additive noise example.

8.1.3 Denoiser

A denoiser (also called estimator) is an estimation \tilde{f} of f_0 computed from the observation f alone. It is thus also a random vector that depends on the noise w . Since f is a random vector of mean f_0 , the numerical denoising process corresponds to the estimation of the mean of a random vector from a single realization. Figure 8.5 shows an example of denoising.

The quality of a denoiser is measured using the average mean square risk $\mathbb{E}_w(\|f_0 - \tilde{f}\|^2)$, where \mathbb{E}_w is the esperance (averaging) with respect to the noise w . Since f_0 is unknown, this corresponds to a theoretical measure of performance, that is bounded using a mathematical analysis. In the numerical experiments, one observes a single realization $f^r \sim f_0 + w$, and the performance is estimated from this single denoising using the SNR

$$\text{SNR}(\tilde{f}^r, f_0) = -20 \log_{10}(\|\tilde{f}^r - f_0\|/\|f_0\|),$$

where \tilde{f}^r should be computed from the single realization f^r of f . In the following, with an abuse of notation, when displaying single realization, we ignore the exponent r . The SNR is expressed in “decibels”, denoted dB. This measure of performance requires the knowledge of the clean signal f_0 , and should thus only be considered as an experimentation tool, that might not be available in a real life denoising scenario where clean data are not available. Furthermore, the use of an ℓ^2 measure of performance is questionable, and one should also observe the result to judge of the visual quality of the denoising.

8.2 Linear Denoising using Filtering

8.2.1 Translation Invariant Estimators

A linear estimator $\mathcal{E}(f) = \tilde{f}$ of f_0 depends linearly on f , so that $\mathcal{E}(f + g) = \mathcal{E}(f) + \mathcal{E}(g)$. A translation invariant estimator commutes with translation, so that $\mathcal{E}(f_\tau) = \mathcal{E}(f)_\tau$, where $f_\tau(t) = f(t - \tau)$. Such a denoiser can always be written as a filtering

$$\tilde{f} = f \star h$$

where $h \in \mathbb{R}^N$ is a (low pass) filter, that should satisfy at least

$$\sum_n h_n = \hat{h}_0 = 1$$

where \hat{h} is the discrete Fourier transform.

Figure 8.6 shows an example of denoising using a low pass filter.

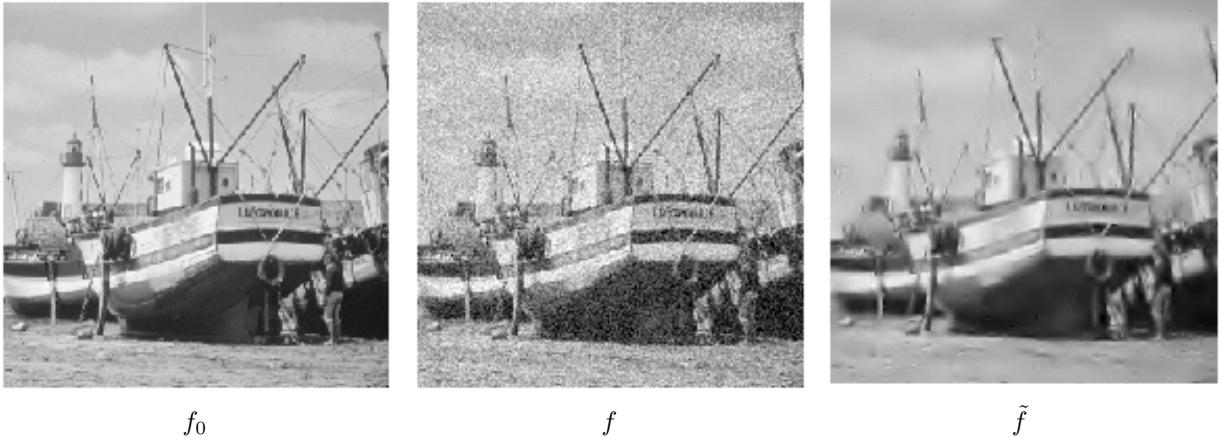


Figure 8.5: Left: clean image, center: noisy image, right: denoised image.

The filtering strength is usually controlled the width s of h . A typical example is the Gaussian filter

$$\forall -N/2 < i \leq N/2, \quad h_{s,i} = \frac{1}{Z_s} \exp\left(-\frac{i^2}{2s^2}\right) \quad (8.1)$$

where Z_s ensures that $\sum_i h_{s,i} = 1$ (low pass). Figure 8.6 shows the effect of Gaussian filtering over the spacial and Fourier domains.

Figure 8.7 shows the effect of low pass filtering on a signal and an image with an increasing filter width s . Linear filtering introduces a blur and are thus only efficient to denoise smooth signals and image. For signals and images with discontinuities, this blur deteriorates the signal. Removing a large amount of noise necessitates to also smooth significantly edges and singularities.

8.2.2 Optimal Filter Selection

The selection of an optimal filter is a difficult task. Its choice depends both on the regularity of the (unknown) data f_0 and the noise level σ . A simpler option is to optimize the filter width s among a parametric family of filters, such as for instance the Gaussian filters defined in (8.1).

The denoising error can be decomposed as

$$\|\tilde{f} - f_0\| \leq \|h_s * f_0 - f_0\| + \|h_s * w\|$$

The filter width s should be optimized to perform a tradeoff between removing enough noise ($\|h_s * w\|$ decreases with s) and not smoothing too much the singularities ($\|h_s * f_0 - f_0\|$ increases with s).

Figure (8.8) shows the oracle SNR performance, defined in (??).

Figure 8.9 and 8.10 show the results of denoising using the optimal filter width s^* that minimizes the SNR for a given noisy observation.

These optimal filtering appear quite noisy, and the optimal SNR choice is usually quite conservative. Increasing the filter width introduces a strong blurring that deteriorates the SNR, although it might look visually more pleasant.

8.2.3 Wiener Filter

If one has a random model both for the noise $w \sim W$ and for the signal $f_0 \sim F$, one can derive an optimal filters in average over both the noise and the signal realizations. One further assumes that w and

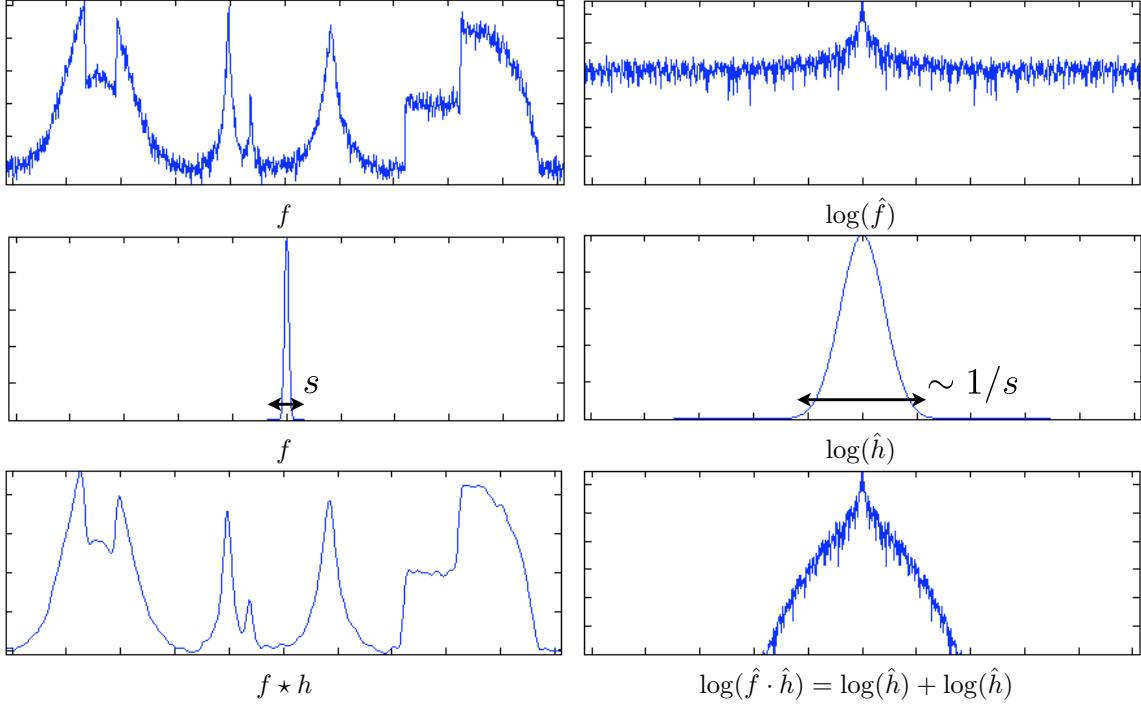


Figure 8.6: Denoising by filtering over the spacial (left) and Fourier (right) domains.

f_0 are independent realization. The optimal h thus minimizes

$$\mathbb{E}_{W,F}(\|h \star (F + W) - F\|^2)$$

If both F is wide-sense stationary, and W is a Gaussian white noise of variance σ^2 , then the optimal filer is known as the Wiener filter

$$\hat{h}_\omega = \frac{|\hat{F}_\omega|^2}{|\hat{F}_\omega|^2 + \sigma^2}$$

where $|\hat{F}|^2$ is the power spectrum of F ,

$$\hat{F}_\omega = \hat{C}_\omega \quad \text{where} \quad C_n = \mathbb{E}(\langle F, F[\cdot + n] \rangle),$$

the Fourier transform of an infinite vector is defined in Section 2.3.

In practice, one rarely has such a random model for the signal, and interesting signals are often not stationary. Most signals exhibit discontinuities, and are thus poorly restored with filtering.

8.2.4 Denoising and Linear Approximation

In order to study linear (and also non-linear, see the section bellow) denoising without assuming a random signal model, one should use approximation theory as studied in Chapter 6. We thus consider an ortho-basis $\mathcal{B} = (\psi_m)_m$ of \mathbb{R}^N , and consider a simple denoising obtained by keeping only the M first term elements of the approximation of the noisy observation in \mathcal{B}

$$\tilde{f} \stackrel{\text{def.}}{=} \sum_{m=1}^M \langle f, \psi_m \rangle \psi_m. \quad (8.2)$$

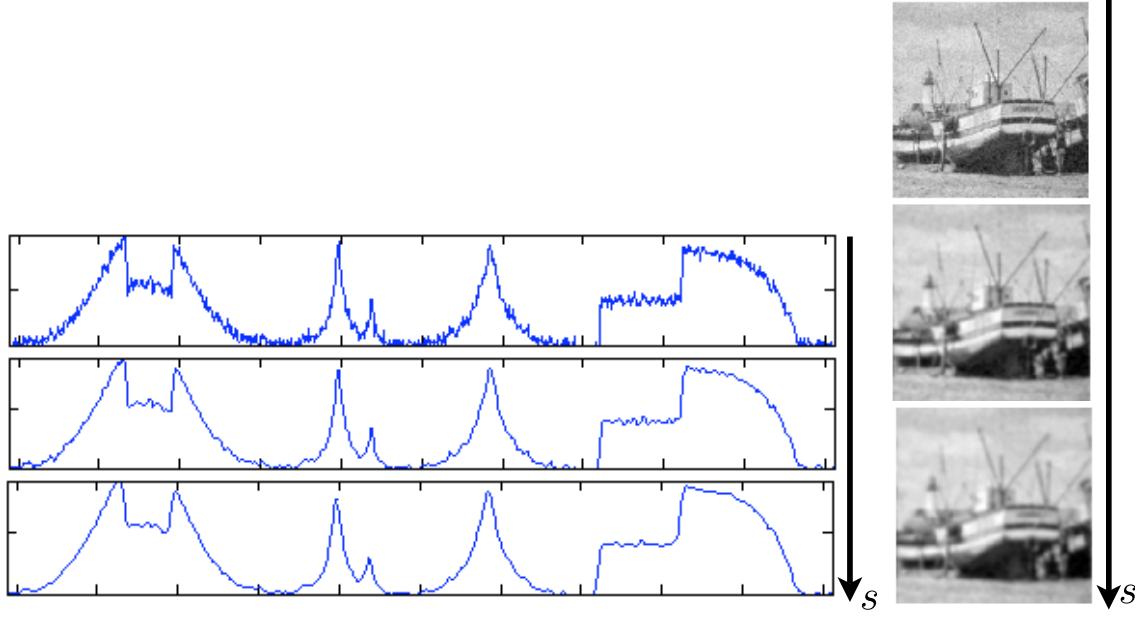


Figure 8.7: Denoising using a filter of increasing width s .

This is a linear projection on the space spanned by $(\psi_m)_{m=1}^M$. This denoising scheme is thus parameterized by some integer $M > 0$, increasing M increases the denoising strength. For instance, when \mathcal{B} is the discrete Fourier basis, this corresponds to an ideal low-pass filter against a (discretized) Dirichlet kernel.

More advanced linear denoising operator can be designed by computing weighted average $\sum_m \lambda_m \langle f, \psi_m \rangle \psi_m$, and (8.2) is retrieved when using binary weights $\alpha_n = 1$ for $n \leq M$, and $\alpha_n = 0$ otherwise. The asymptotic theoretical performances described by the following theorem are however not improved by using non-binary weights.

Theorem 29. *We assume that $f_0 \in \mathbb{R}^N$ has a linear approximation error decay that satisfies*

$$\forall M, \quad \|f_0 - f_{0,M}^{lin}\|^2 \leq CM^{-2\beta} \quad \text{where} \quad f_{0,M}^{lin} \stackrel{\text{def.}}{=} \sum_{m=1}^M \langle f_0, \psi_m \rangle \psi_m$$

for some constant C . Then the linear denoising error using (8.2) satisfies

$$\mathbb{E}(\|f_0 - \tilde{f}\|^2) \leq 2C^{\frac{1}{2\beta+1}} \sigma^{2-\frac{1}{\beta+1/2}},$$

when choosing

$$M = C^{\frac{1}{2\beta+1}} \sigma^{-\frac{2}{2\beta+1}}. \tag{8.3}$$

Proof. One has, thanks to the ortho-normality of $(\psi_m)_m$

$$\begin{aligned} \mathbb{E}(\|f_0 - \tilde{f}\|^2) &= \mathbb{E}(\sum_m \langle f_0 - \tilde{f}, \psi_m \rangle^2) = \mathbb{E}(\sum_{m=1}^M \langle f_0 - f, \psi_m \rangle^2 + \sum_{m>M} \langle f_0, \psi_m \rangle^2) \\ &= \mathbb{E}\left(\sum_{m=1}^M \langle f_0, \psi_m \rangle^2\right) + \sum_{m>M} \langle f_0, \psi_m \rangle^2 = M\sigma^2 + \|f_0 - f_{0,M}^{lin}\|^2 \\ &\leq M\sigma^2 + CM^{-2\beta}. \end{aligned}$$

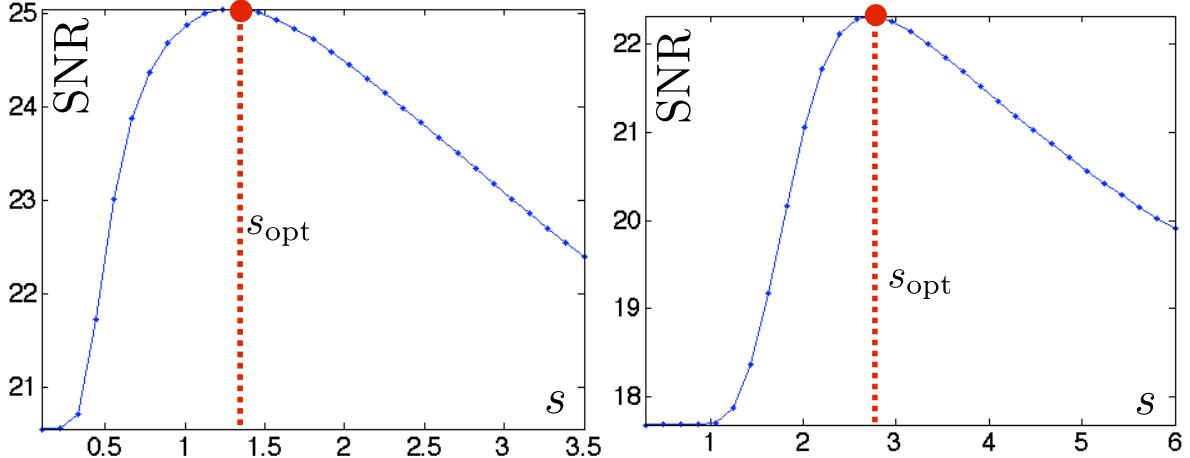


Figure 8.8: Curves of SNR as a function of the filtering width in 1-D (left) and 2-D (right).

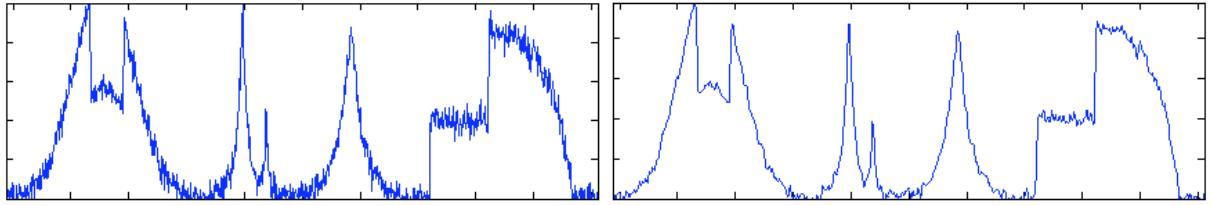


Figure 8.9: Noisy image (left) and denoising (right) using the optimal filter width.

Here we use the fundamental fact that $(\langle w, \psi_m \rangle)_m$ is also $\mathcal{N}(0, \sigma^2 \text{Id}_N)$. Choosing M such that $M\sigma^2 = CM^{-2\beta}$, i.e. $M = C^{\frac{1}{2\beta+1}}\sigma^{-\frac{2}{2\beta+1}}$ leads to

$$\mathbb{E}(\|f_0 - \tilde{f}\|^2) = 2CM^{-2\beta} = 2CC^{-\frac{2\beta}{2\beta+1}}\sigma^{\frac{4\beta}{2\beta+1}} = 2C^{\frac{1}{2\beta+1}}\sigma^{2-\frac{1}{\beta+1/2}}.$$

□

There are several important remarks regarding this simple but important result:

- Thanks to the decay of the linear approximation error, the denoising error $\mathbb{E}(\|f_0 - \tilde{f}\|^2)$ is bounded *independently* of the sampling size N , although the input noise level $\mathbb{E}(\|w\|^2) = N\sigma^2$ grows with N .
- If the signal is well approximated linearly, i.e. if β is large, then the denoising error decays fast when the noise level σ drops to zero. The upper bound approaches the optimal rate σ^2 by taking β large enough.
- This theory is finite dimensional, i.e. this computation makes only sense when introducing some discretization step N . This is natural because random noise vectors of finite energy are necessarily finite dimensional. For the choice (8.3) to be realizable, one should however have $M \leq N$, i.e. $N \geq C^{\frac{1}{2\beta+1}}\sigma^{-\frac{2}{2\beta+1}}$. Thus N should increase when the noise diminishes for the denoising effect to kick-in.
- Section 6.3.1 bounds the linear approximation error for infinite dimensional signal and image model. This theory can be applied provided that the discretization error is smaller than the denoising error, i.e. once again, one should use N large enough.

A typical setup where this denoising theorem can be applied is for the Sobolev signal and image model detailed in Section 6.2.1. In the discrete setting, where the sampling size N is intended to grow (especially if σ diminishes), one can similarly consider a “Sobolev-like” model, and similarly as for Proposition 16, this model implies a decay of the linear approximation error.



Figure 8.10: Noisy image (left) and denoising (right) using the optimal filter width.

Proposition 19. *Assuming that*

$$\sum_{m=1}^N m^{2\alpha} |\langle f_0, \psi_m \rangle|^2 \leq C \quad (8.4)$$

then

$$\forall M, \quad \|f_0 - f_{0,M}^{\text{lin}}\|^2 \leq CM^{-2\alpha}$$

Proof.

$$C \geq \sum_{m=1}^N m^{2\alpha} |\langle f_0, \psi_m \rangle|^2 \geq \sum_{m>M} m^{2\alpha} |\langle f_0, \psi_m \rangle|^2 \geq M^{2\alpha} \sum_{m>M} |\langle f_0, \psi_m \rangle|^2 \geq M^{2\alpha} \|f_0 - f_{0,M}^{\text{lin}}\|^2.$$

□

If ψ_m is the discrete Fourier basis defined in (2.8), then this discrete Sobolev model (8.4) is equivalent to the continuous Sobolev model of Section 6.2.1, up to a discretization error which tends to 0 as N increase. Choosing N large enough shows that smooth signals and image are thus efficiently denoised by a simple linear projection on the first M element of the Fourier basis.

8.3 Non-linear Denoising using Thresholding

8.3.1 Hard Thresholding

We consider an orthogonal basis $\{\psi_m\}_m$ of \mathbb{C}^N , for instance a discrete wavelet basis. The noisy coefficients satisfy

$$\langle f, \psi_m \rangle = \langle f_0, \psi_m \rangle + \langle w, \psi_m \rangle. \quad (8.5)$$

Since a Gaussian white noise is invariant under an orthogonal transformation, $\langle w, \psi_m \rangle$ is also a Gaussian white noise of variance σ^2 . If the basis $\{\psi_m\}_m$ is efficient to represent f_0 , then most of the coefficients $\langle f_0, \psi_m \rangle$ are close to zero, and one observes a large set of small noisy coefficients, as shown on Figure 8.11. This idea of using thresholding estimator for denoising was first systematically explored by Donoho and Johnstone [18].

A thresholding estimator removes these small amplitude coefficients using a non-linear hard thresholding

$$\tilde{f} = \sum_{|\langle f, \psi_m \rangle| > T} \langle f, \psi_m \rangle \psi_m = \sum_m S_T(\langle f, \psi_m \rangle) \psi_m.$$

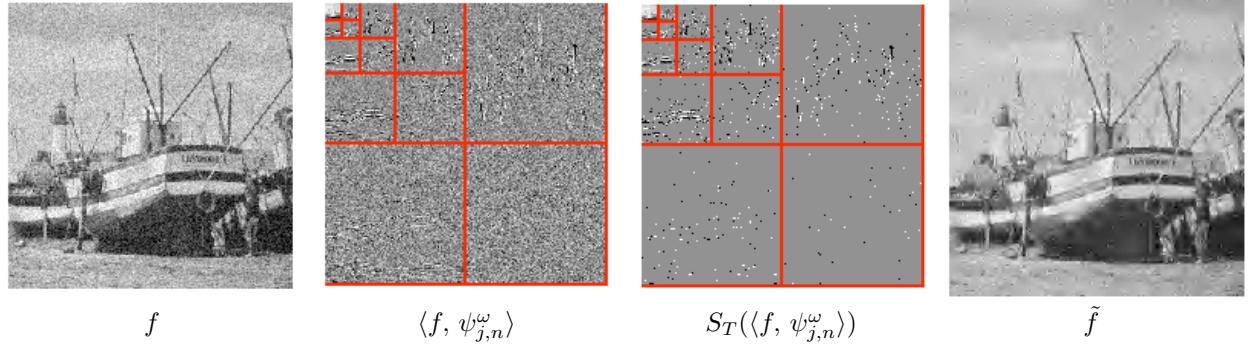


Figure 8.11: Denoising using thresholding of wavelet coefficients.

where S_T is defined in (11.4). This corresponds to the computation of the best M -term approximation $\tilde{f} = f_M$ of the noisy function f . Figure 8.11 shows that if T is well chose, this non-linear estimator is able to remove most of the noise while maintaining sharp features, which was not the case with linear filtering estimatiors.

8.3.2 Soft Thresholding

We recall that the hard thresholding operator is defined as

$$S_T(x) = S_T^0(x) = \begin{cases} x & \text{if } |x| > T, \\ 0 & \text{if } |x| \leq T. \end{cases} \quad (8.6)$$

This thresholding performs a binary decision that might introduces artifacts. A less aggressive nonlinearity is the soft thresholding

$$S_T^1(x) = \max(1 - T/|x|, 0)x. \quad (8.7)$$

Figure 8.12 shows the 1-D curves of these 1-D non-linear mapping.

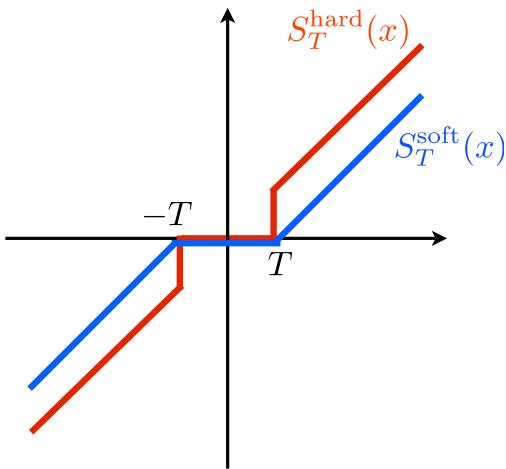


Figure 8.12: Hard and soft thresholding functions.

For $q = 0$ and $q = 1$, these thresholding defines two different estimators

$$\tilde{f}^q = \sum_m S_T^q(\langle f, \psi_m \rangle) \psi_m \quad (8.8)$$

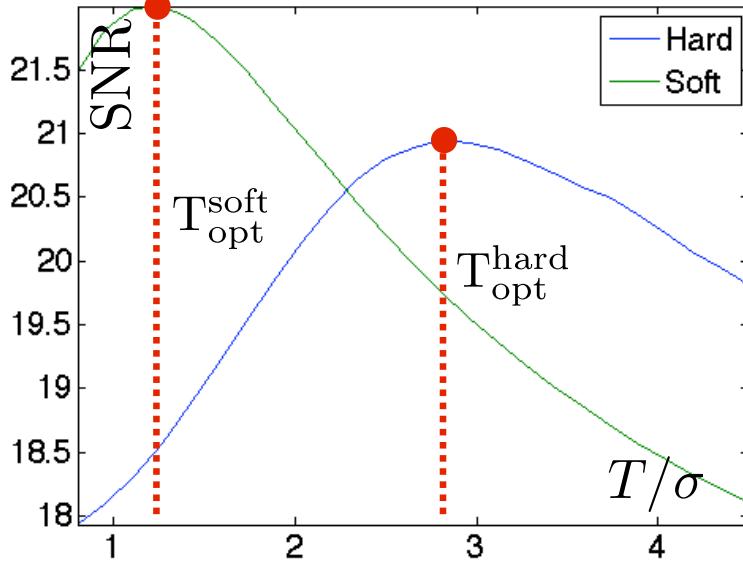


Figure 8.13: Curves of SNR with respect to T/σ for hard and soft thresholding.

Coarse scale management. The soft thresholded S_T^1 introduces a bias since it diminishes the value of large coefficients. For wavelet transforms, it tends to introduce unwanted low-frequencies artifacts by modifying coarse scale coefficients. If the coarse scale is 2^{j_0} , one thus prefers not to threshold the coarse approximation coefficients and use, for instance in 1-D,

$$\tilde{f}^1 = \sum_{0 \leq n < 2^{-j_0}} \langle f, \varphi_{j_0, n} \rangle \varphi_{j_0, n} + \sum_{j=j_0}^0 \sum_{0 \leq n < 2^{-j}} S_T^1(\langle f, \psi_{j_0, n} \rangle) \psi_{j_0, n}.$$

Empirical choice of the threshold. Figure 8.13 shows the evolution of the SNR with respect to the threshold T for these two estimators, for a natural image f_0 . For the hard thresholding, the best result is obtained around $T \approx 3\sigma$, while for the soft thresholding, the optimal choice is around $T \approx 3\sigma/2$. These results also show that numerically, for thresholding in orthogonal bases, soft thresholding is slightly superior than hard thresholding on natural signals and images.

Although these are experimental conclusions, these results are robust across various natural signals and images, and should be considered as good default parameters.

8.3.3 Minimax Optimality of Thresholding

Sparse coefficients estimation. To analyze the performance of the estimator, and gives an estimate for the value of T , we first assume that the coefficients

$$a_{0,m} = \langle f_0, \psi_m \rangle \in \mathbb{R}^N$$

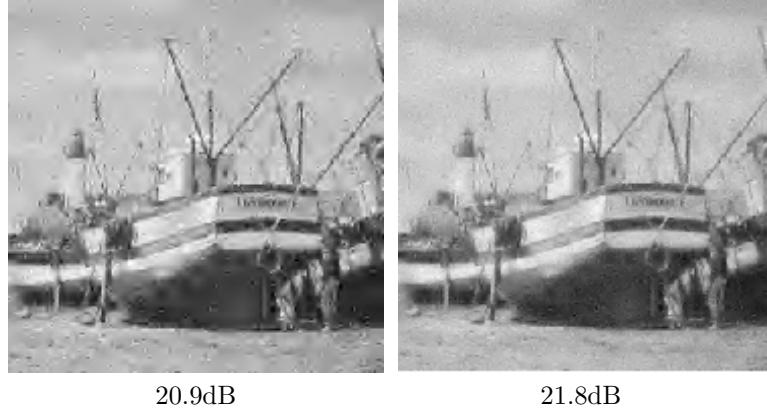


Figure 8.14: Comparison of hard (left) and soft (right) thresholding.

are sparse, meaning that most of the $a_{0,m}$ are zero, so that its ℓ^0 norm

$$\|a_0\|_0 = \#\{m ; a_{0,m} \neq 0\}$$

is small. As shown in (8.5), noisy coefficients

$$\langle f, \psi_m \rangle = a_m = a_{0,m} + z_m$$

are perturbed with an additive Gaussian white noise of variance σ^2 . Figure 8.15 shows an example of such a noisy sparse signal.

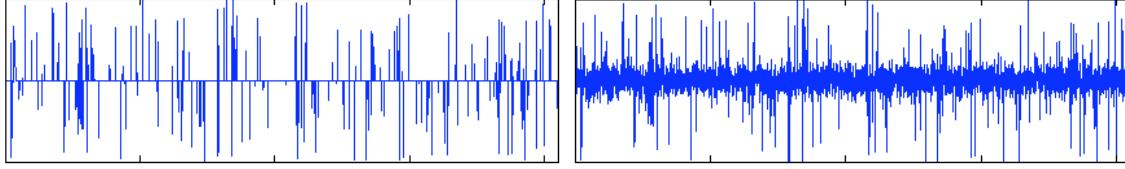


Figure 8.15: Left: sparse signal a , right: noisy signal.

Universal threshold value. If

$$\min_{m:a_{0,m} \neq 0} |a_{0,m}|$$

is large enough, then $\|f_0 - \tilde{f}\| = \|a_0 - S_T(a)\|$ is minimum for

$$T \approx \tau_N = \max_{0 \leq m < N} |z_m|.$$

τ_N is a random variable that depends on N . One can show that its mean is $\sigma\sqrt{2\log(N)}$, and that as N increases, its variance tends to zero and τ_N is highly concentrated close to its mean. Figure 8.16 shows that this is indeed the case numerically.

Asymptotic optimality. Donoho and Johnstone [18] have shown that the universal threshold $T = \sigma\sqrt{2\log(N)}$ is a good theoretical choice for the denoising of signals that are well approximated non-linearly in $\{\psi_m\}_m$. The obtain denoising error decay rate with σ can also be shown to be in some sense optimal.

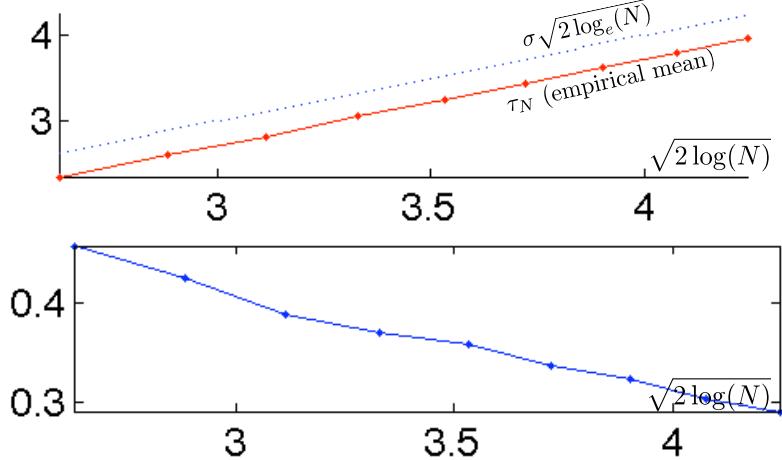


Figure 8.16: Empirical estimation of the mean of Z_n (top) and standard deviation of Z_n (bottom)

Theorem 30. We assume that $f_0 \in \mathbb{R}^N$ has a non-linear approximation error decay that satisfies

$$\forall M, \quad \|f_0 - f_{0,M}^{nlin}\|^2 \leq CM^{-2\beta} \quad \text{where} \quad f_{0,M}^{nlin} \stackrel{\text{def}}{=} \sum_{r=1}^M \langle f_0, \psi_{m_r} \rangle \psi_{m_r}$$

for some constant C , where here $(\langle f_0, \psi_{m_r} \rangle)_r$ are the coefficient sorted by decaying magnitude. Then the non-linear denoising error using (8.2) satisfies

$$\mathbb{E}(\|f_0 - \tilde{f}^q\|^2) \leq C' \ln(N) \sigma^{2-\frac{1}{\beta+1/2}},$$

for some constant C' , when choosing $T = \sqrt{2\ln(N)}$, where \tilde{f}^q is defined in (8.8) for $q \in \{0, 1\}$.

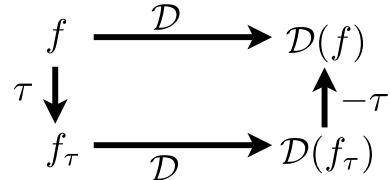
This universal threshold choice $T = \sqrt{2\ln(N)}$ is however very conservative since it is guaranteed to remove almost all the noise. In practice, as shown in Figure 8.14, better results are obtained on natural signals and images by using $T \approx 3\sigma$ and $T \approx 3\sigma/2$ for hard and soft thresholdings.

8.3.4 Translation Invariant Thresholding Estimators

Translation invariance. Let $f \mapsto \tilde{f} = \mathcal{D}(f)$ by a denoising method, and $f_\tau(x) = f(x - \tau)$ be a translated signal or image for $\tau \in \mathbb{R}^d$, ($d = 1$ or $d = 2$). The denoising is said to be translation invariant at precision Δ if

$$\forall \tau \in \Delta, \quad \mathcal{D}(f) = \mathcal{D}(f_\tau)_{-\tau}$$

where Δ is a lattice of \mathbb{R}^d . The denser Δ is, the more translation invariant the method is. This corresponds to the fact that \mathcal{D} computes with the translation operator.



Imposing translation invariance for a fine enough set Δ is a natural constraint, since intuitively the denoising results should not depend on the location of features in the signal or image. Otherwise, some locations might be favored by the denoising process, which might result in visually unpleasant denoising artifacts.

For denoising by thresholding

$$\mathcal{D}(f) = \sum_m S_T(\langle f, \psi_m \rangle) \psi_m.$$

then translation invariance is equivalent to asking that the basis $\{\psi_m\}_m$ is translation invariant at precision Δ ,

$$\forall m, \forall \tau \in \Delta, \exists m, \exists \lambda \in \mathbb{C}, \quad (\psi_{m'})_\tau = \lambda \psi_m$$

where $|\lambda| = 1$.

The Fourier basis is fully translation invariant for $\Delta = \mathbb{R}^d$ over $[0, 1]^d$ with periodic boundary conditions and the discrete Fourier basis is translation invariant for all integer translations $\Delta = \{0, \dots, N_0 - 1\}^d$ where $N = N_0$ is the number of points in 1-D, and $N = N_0 \times N_0$ is the number of pixels in 2-D.

Unfortunately, an orthogonal wavelet basis

$$\{\psi_m = \psi_{j,n}\}_{j,n}$$

is not translation invariant both in the continuous setting or in the discrete setting. For instance, in 1-D,

$$(\psi_{j',n'})_\tau \notin \{\psi_{j,n}\} \quad \text{for } \tau = 2^j/2.$$

Cycle spinning. A simple way to turn a denoiser Δ into a translation invariant denoiser is to average the result of translated images

$$\mathcal{D}_{\text{inv}}(f) = \frac{1}{|\Delta|} \sum_{\tau \in \Delta} \mathcal{D}(f_\tau)_{-\tau}. \quad (8.9)$$

One easily check that

$$\forall \tau \in \Delta, \quad \mathcal{D}_{\text{inv}}(f) = \mathcal{D}_{\text{inv}}(f_\tau)_{-\tau}$$

To obtain a translation invariance up to the pixel precision for a data of N samples, one should use a set of $|\Delta| = N$ translation vectors. To obtain a pixel precision invariance for wavelets, this will result in $O(N^2)$ operations.

Figure 8.17 shows the result of applying cycle spinning to an orthogonal hard thresholding denoising using wavelets, where we have used the following translation of the continuous wavelet basis $\Delta = \{0, 1/N, 2/N, 3/N\}^2$, which corresponds to discrete translation by $\{0, 1, 2, 3\}^2$ on the discretized image. The complexity of the denoising scheme is thus 16 wavelet transforms. The translation invariance brings a very large SNR improvement, and significantly reduces the oscillating artifacts of orthogonal thresholding. This is because these artifacts pop-out at random locations when τ changes, so that the averaging process reduces significantly these artifacts.

Figure 8.18 shows that translation invariant hard thresholding does a slightly better job than translation invariant soft thresholding. The situation is thus reversed with respect to thresholding in an orthogonal wavelet basis.

Translation invariant wavelet frame. An equivalent way to define a translation invariant denoiser is to replace the orthogonal basis $\mathcal{B} = \{\psi_m\}$ by a redundant family of translated vectors

$$\mathcal{B}_{\text{inv}} = \{(\psi_m)_\tau\}_{m,\tau \in \Delta}. \quad (8.10)$$

One should be careful about the fact that \mathcal{B}_{inv} is not any more an orthogonal basis, but it still enjoys a conservation of energy formula

$$\|f\|^2 = \frac{1}{|\Delta|} \sum_{m,\tau \in \Delta} |\langle f, (\psi_m)_\tau \rangle|^2 \quad \text{and} \quad f = \frac{1}{|\Delta|} \sum_{m,\tau \in \Delta} \langle f, (\psi_m)_\tau \rangle (\psi_m)_\tau.$$

This kind of redundant family are called tight frames.

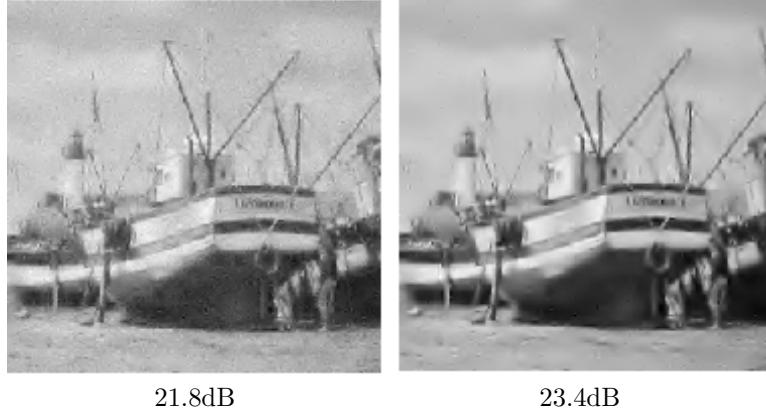


Figure 8.17: Comparison of wavelet orthogonal soft thresholding (left) and translation invariant wavelet hard thresholding (right).

One can then define a translation invariant thresholding denoising

$$\mathcal{D}_{\text{inv}}(f) = \frac{1}{|\Delta|} \sum_{m, \tau \in \Delta} S_T(\langle f, (\psi_m)_\tau \rangle)(\psi_m)_\tau. \quad (8.11)$$

This denoising is the same as the cycle spinning denoising defined in (8.9).

The frame \mathcal{B}_{inv} might contain up to $|\Delta||\mathcal{B}|$ basis element. For a discrete basis of signal with N samples, and a translation lattice of $|\Delta| = N$ vectors, it corresponds to up to N^2 elements in \mathcal{B}_{inv} . Hopefully, for a hierarchical basis such as a discrete orthogonal wavelet basis, one might have

$$(\psi_m)_\tau = (\psi_{m'})_{\tau'} \quad \text{for } m \neq m' \quad \text{and} \quad \tau \neq \tau',$$

so that the number of elements in \mathcal{B}_{inv} might be much smaller than N^2 . For instance, for an orthogonal wavelet basis, one has

$$(\psi_{j,n})_{k2^j} = \psi_{j,n+k},$$

so that the number of basis elements is $|\mathcal{B}_{\text{inv}}| = N \log_2(N)$ for a 2-D basis, and $3N \log_2(N)$ for a 3-D basis. The fast translation invariant wavelet transform, also called “a trou” wavelet transform, computes all the inner products $\langle f, (\psi_m)_\tau \rangle$ in $O(N \log_2(N))$ operations. Implementing formula (8.11) is thus much faster than applying the cycle spinning (8.9) equivalent formulation.

Translation invariant wavelet coefficients are usually grouped by scales in $\log_2(N)$ (for $d = 1$) or by scales and orientations $3 \log_2(N)$ (for $d = 2$) sets of coefficients. For instance, for a 2-D translation invariant transform, one consider

$$\forall n \in \{0, \dots, 2^j N_0 - 1\}^2, \forall k \in \{0, \dots, 2^{-j}\}^2, \quad d_j^\omega[2^{-j}n + k] = \langle f, (\psi_{j,n})_{k2^j} \rangle$$

where $\omega \in \{V, H, D\}$ is the orientation. Each set d_j^ω has N coefficients and is a band-pass filtered version of the original image f , as shown on Figure 8.19.

Figure 8.20 shows how these set of coefficients are hard thresholded by the translation invariant estimator.

8.3.5 Exotic Thresholdings

It is possible to devise many thresholding nonlinearities that interpolate between the hard and soft thresholding. We present here two examples, but many more exist in the literature. Depending on the statistical distribution of the wavelet coefficients of the coefficients of f in the basis, these thresholders might produce slightly better results.

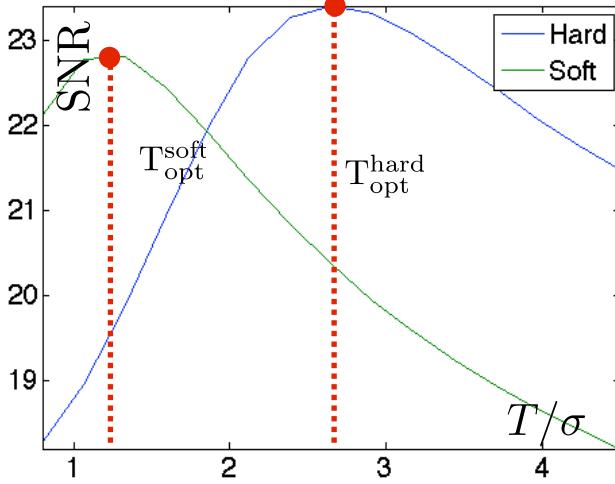


Figure 8.18: Curve of SNR with respect to T/σ for translation invariant thresholding.

Semi-soft thresholding. One can define a family of intermediate thresher that depends on a parameter $\mu > 1$

$$S_T^\theta(x) = g_{\frac{1}{1-\theta}}(x) \quad \text{where} \quad g_\mu(x) = \begin{cases} 0 & \text{if } |x| < T \\ x & \text{if } |x| > \mu T \\ \text{sign}(x) \frac{|x|-T}{\mu-1} & \text{otherwise.} \end{cases}$$

One thus recovers the hard thresholding as S_T^0 and the soft thresholding as S_T^1 . Figure 8.21 display an example of such a non-linearity.

Figure 8.22 shows that a well chosen value of μ might actually improves over both hard and soft thresholder. The improvement is however hardly noticeable visually.

Stein thresholding. The Stein thresholding is defined using a quadratic attenuation of large coefficients

$$S_T^{\text{Stein}}(x) = \max \left(1 - \frac{T^2}{|x|^2}, 0 \right) x.$$

This should be compared with the linear attenuation of the soft thresholding

$$S_T^1(x) = \max \left(1 - \frac{T}{|x|}, 0 \right) x.$$

The advantage of the Stein thresher with respect to the soft thresholding is that

$$|S_T^{\text{Stein}}(x) - x| \rightarrow 0 \quad \text{whereas} \quad |S_T^1(x) - x| \rightarrow T,$$

where $x \rightarrow \pm\infty$. This means that Stein thresholding does not suffer from the bias of soft thresholding.

For translation invariant thresholding, Stein and hard thresholding perform similarly on natural images.

8.3.6 Block Thresholding

The non-linear thresholding method presented in the previous section are diagonal estimators, since they operate a coefficient-by-coefficient attenuation

$$\tilde{f} = \sum_m A_T^q(\langle f, \psi_m \rangle) \langle f, \psi_m \rangle \psi_m$$

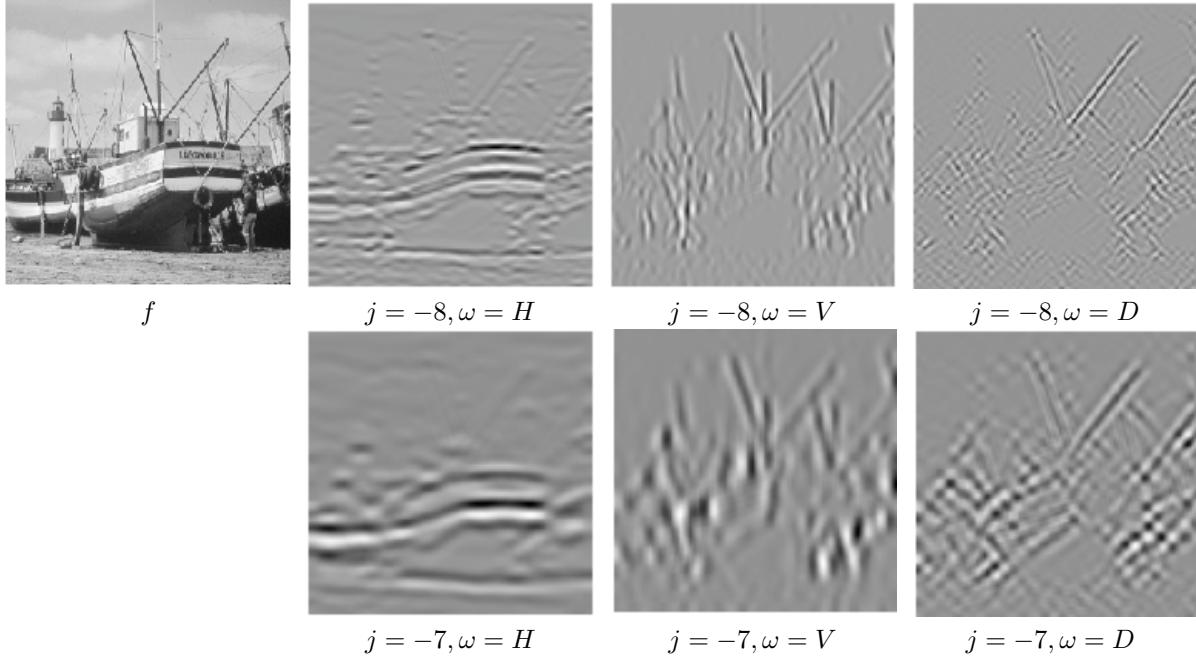


Figure 8.19: Translation invariant wavelet coefficients.

where

$$A_T^q(x) = \begin{cases} \max(1 - x^2/T^2, 0) & \text{for } q = \text{Stein} \\ \max(1 - |x|/T, 0) & \text{for } q = 1 \text{ (soft)} \\ 1_{|x|>T} & \text{for } q = 0 \text{ (hard)} \end{cases}$$

Block thresholding takes advantage of the statistical dependency of wavelet coefficients, by computing the attenuation factor on block of coefficients. This is especially efficient for natural images, where edges and geometric features create clusters of high magnitude coefficients. Block decisions also help to remove artifacts due to isolated noisy large coefficients in regular areas.

The set of coefficients is divided into disjoint blocks, and for instance for 2-D wavelet coefficients

$$\{(j, n, \omega)\}_{j, n, \omega} = \bigcup_k B_k,$$

where each B_k is a square of $s \times s$ coefficients, where the block size s is a parameter of the method. Figure 8.24 shows an example of such a block.

The block energy is defined as

$$B_k = \frac{1}{s^2} \sum_{m \in B_k} |\langle f, \psi_m \rangle|^2,$$

and the block thresholding

$$\tilde{f} = \sum_m S_T^{\text{block}, q}(\langle f, \psi_m \rangle) \psi_m$$

makes use of the same attenuation for all coefficients within a block

$$\forall m \in B_k, \quad S_T^{\text{block}, q}(\langle f, \psi_m \rangle) = A_T^q(E_k) \langle f, \psi_m \rangle.$$

for $q \in \{0, 1, \text{stein}\}$. Figure 8.24 shows the effect of this block attenuation, and the corresponding denoising result.

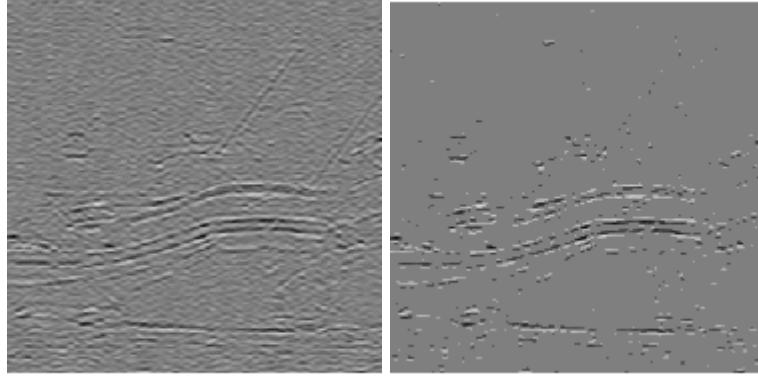


Figure 8.20: Left: translation invariant wavelet coefficients, for $j = -8, \omega = H$, right: thresholded coefficients.

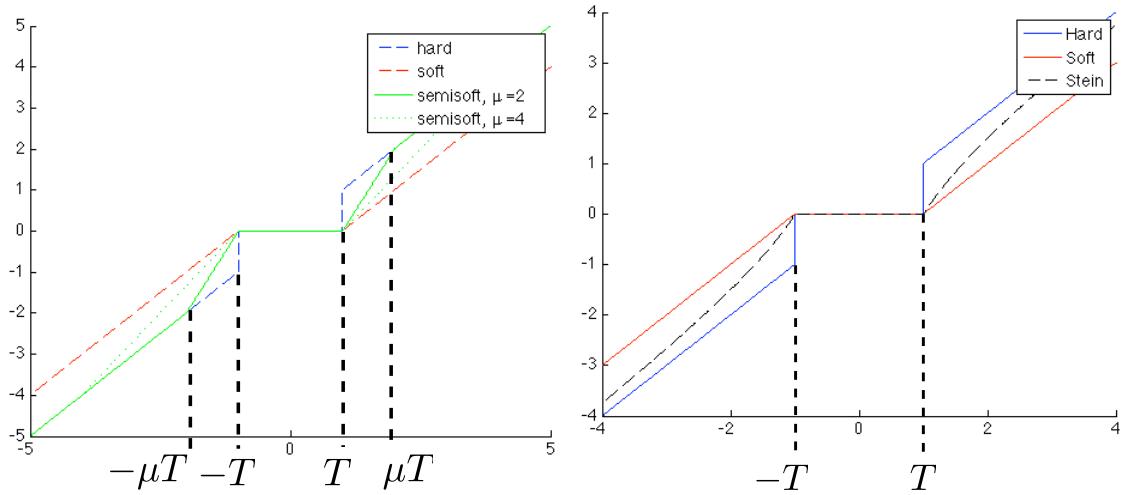


Figure 8.21: Left: semi-soft thresholder, right: Stein threshold器.

Figure 8.25, left, compares the three block thresholding obtained for $q \in \{0, 1, \text{stein}\}$. Numerically, on natural images, Stein block thresholding gives the best results. Figure 8.25, right, compares the block size for the Stein block thresholder. Numerically, for a broad range of images, a value of $s = 4$ works well.

Figure 8.26 shows a visual comparison of the denoising results. Block stein thresholding of orthogonal wavelet coefficients gives a result nearly as good as a translation invariant wavelet hard thresholding, with a faster algorithm. The block thresholding strategy can also be applied to wavelet coefficients in translation invariant tight frame, which produces the best results among all denoisers detailed in this book.

Code ?? implement this block thresholding.

One should be aware that more advanced denoisers use complicated statistical models that improves over the methods proposed in this book, see for instance [32].

8.4 Data-dependant Noises

For many imaging devices, the variance of the noise that perturbs $f_{0,n}$ depends on the value of $f_{0,n}$. This is a major departure from the additive noise formation model considered so far. We present here two popular examples of such non-additive models.

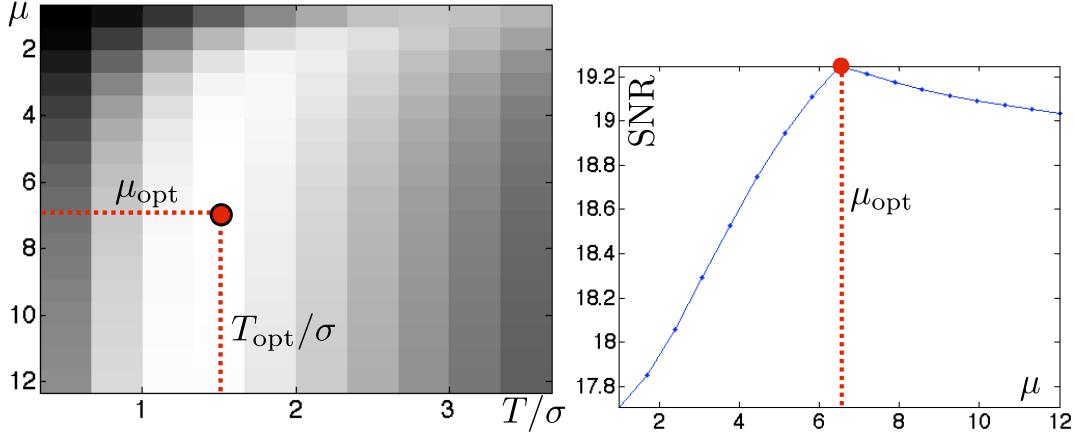


Figure 8.22: Left: image of SNR with respect to the parameters μ and T/σ , right: curve of SNR with respect to μ using the best T/σ for each μ .

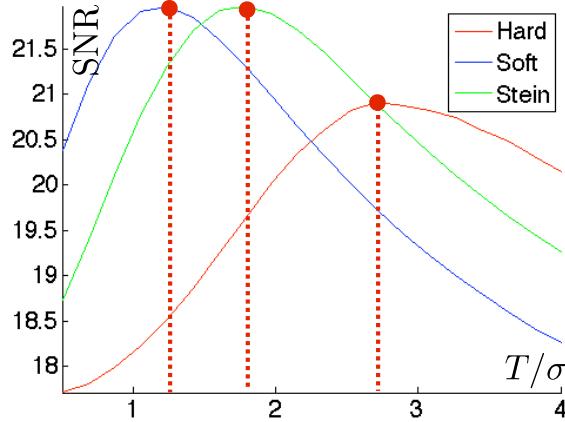


Figure 8.23: SNR curves with respect to T/σ for Stein threhsolding.

8.4.1 Poisson Noise

Many imaging devices sample an image through a photons counting operation. This is for instance the case in digital camera, confocal microscopy, TEP and SPECT tomography.

Poisson model. The uncertainty of the measurements for a quantized unknown image $f_{0,n} \in \mathbb{N}$ is then modeled using a Poisson noise distribution

$$f_n \sim \mathcal{P}(\lambda) \quad \text{where} \quad \lambda = f_{0,n} \in \mathbb{N},$$

and where the Poisson distribution $\mathcal{P}(\lambda)$ is defined as

$$\mathbb{P}(f_n = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

and thus varies from pixel to pixel. Figure 8.27 shows examples of Poisson distributions.

One has

$$\mathbb{E}(f_n) = \lambda = f_{0,n} \quad \text{and} \quad \text{Var}(f_n) = \lambda = f_{0,n}$$

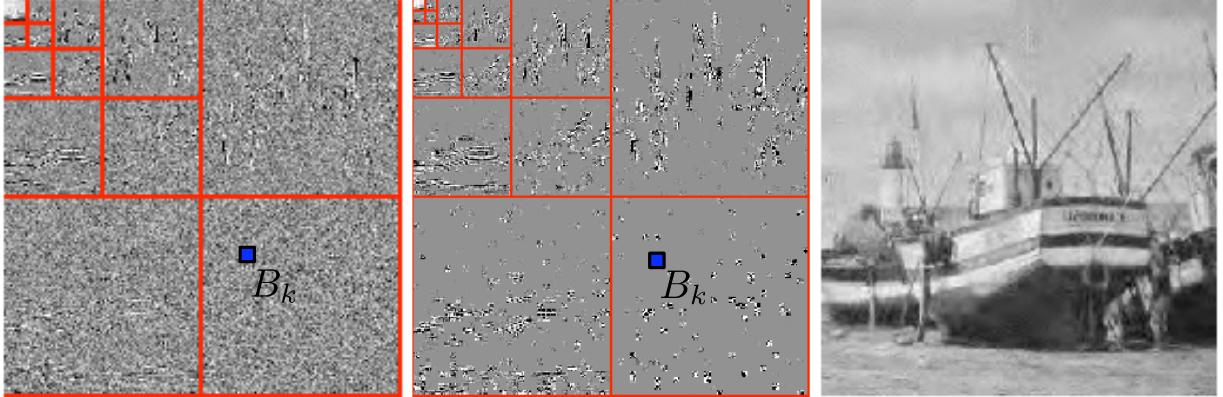


Figure 8.24: Left: wavelet coefficients, center: block thresholded coefficients, right: denoised image.

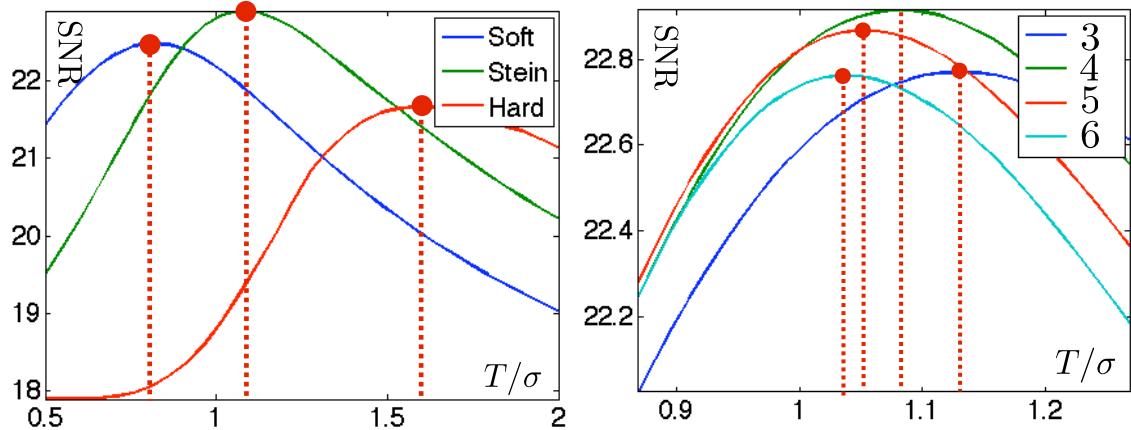


Figure 8.25: Curve of SNR with respect to T/σ (left) and comparison of SNR for different block size (right).

so that the denoising corresponds to estimating the mean of a random vector from a single observation, but the variance now depends on the pixel intensity. This shows that the noise level increase with the intensity of the pixel (more photons are coming to the sensor) but the relative variation $(f_n - f_{0,n})/f_{0,n}$ tends to zero in expectation when $f_{0,n}$ increases.

Figure 8.28 shows examples of a clean image f_0 quantized using different values of λ_{\max} and perturbed with the Poisson noise model.

Variance stabilization. Applying thresholding estimator

$$\mathcal{D}(f) = \sum_m S_T^q(\langle f, \psi_m \rangle) \psi_m$$

to f might give poor results since the noise level fluctuates from point to point, and thus a single threshold T might not be able to capture these variations. A simple way to improve the thresholding results is to first apply a variance stabilization non-linearity $\varphi : \mathbb{R} \rightarrow \mathbb{R}$ to the image, so that $\varphi(f)$ is as close as possible to an additive Gaussian white noise model

$$\varphi(f) \approx \varphi(f_0) + w \quad (8.12)$$



Figure 8.26: Left: translation invariant wavelet hard thresholding, center: block orthogonal Stein thresholding, right: block translation invariant Stein thresholding.

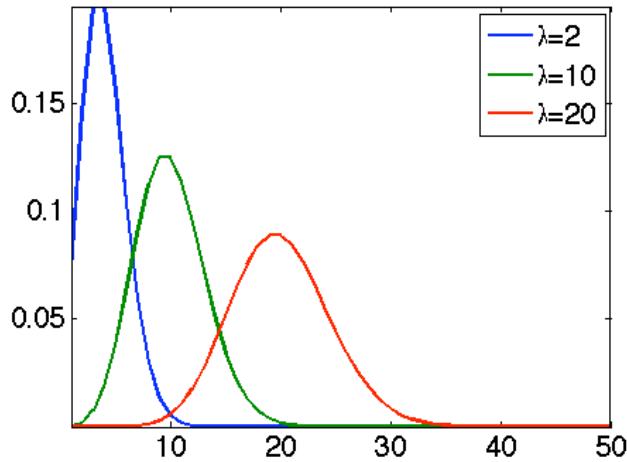


Figure 8.27: Poisson distributions for various λ .

where $w_n \sim \mathcal{N}(0, \sigma)$ is a Gaussian white noise of fixed variance σ^2 .

Perfect stabilization is impossible, so that (8.12) only approximately holds for a limited intensity range of $f_{0,n}$. Two popular variation stabilization functions for Poisson noise are the Anscombe mapping

$$\varphi(x) = 2\sqrt{x + 3/8}$$

and the mapping of Freeman and Tukey

$$\varphi(x) = \sqrt{x + 1} + \sqrt{x}.$$

Figure 8.29 shows the effect of these variance stabilizations on the variance of $\varphi(f)$.

A variance stabilized denoiser is defined as

$$\Delta^{\text{stab},q}(f) = \varphi^{-1} \left(\sum_m S_T^q(\langle \varphi(f), \psi_m \rangle) \psi_m \right)$$

where φ^{-1} is the inverse mapping of φ .

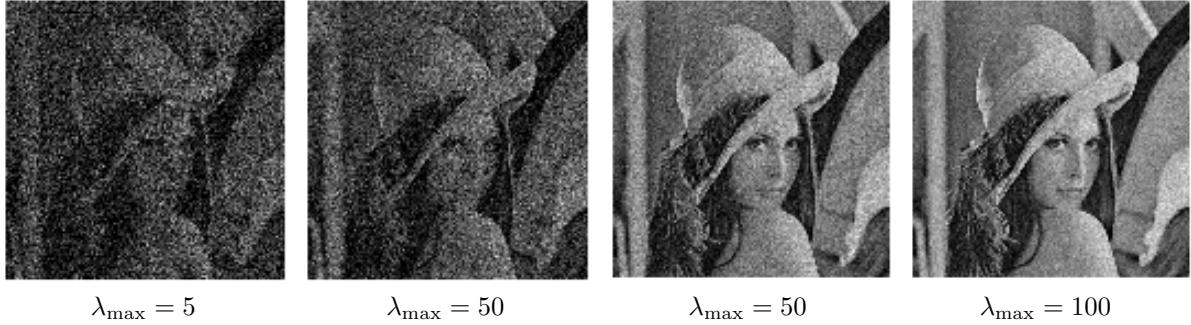


Figure 8.28: Noisy image with Poisson noise model, for various $\lambda_{\max} = \max_n f_{0,n}$.

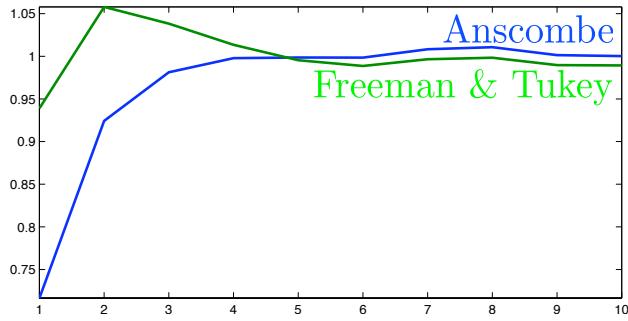


Figure 8.29: Comparison of variance stabilization: display of $\text{Var}(\varphi(f_n))$ as a function of $f_{0,n}$.

Figure 8.30 shows that for moderate intensity range, variance stabilization improves over non-stabilized denoising.

8.4.2 Multiplicative Noise

Multiplicative image formation. A multiplicative noise model assumes that

$$f_n = f_{0,n} w_n$$

where w is a realization of a random vector with $\mathbb{E}(w) = 1$. Once again, the noise level depends on the pixel value

$$\mathbb{E}(f_n) = f_{0,n} \quad \text{and} \quad \text{Var}(f_n) = f_{0,n}^2 \sigma^2 \quad \text{where} \quad \sigma^2 = \text{Var}(w).$$

Such a mutiplicative noise is a good model for SAR satellite imaging, where f is obtained by averaging S images

$$\forall 0 \leq s < K, \quad f_n^{(s)} = f_{0,n} w_n^{(s)} + r_n^{(s)}$$

where $r^{(s)}$ is a Gaussian white noise, and $w_n^{(s)}$ is distributed according to a one-sided exponential distribution

$$\mathcal{P}(w_n^{(s)} = x) \propto e^{-x} \mathbb{I}_{x>0}.$$

For K large enough, averaging the images cancels the additive noise and one obtains

$$f_n = \frac{1}{K} \sum_{s=1}^K f_n^{(s)} \approx f_{0,n} w_n$$



Figure 8.30: Left: noisy image, center: denoising without variance stabilization, right: denoising after variance stabilization.

where w is distributed according to a Gamma distribution

$$w \sim \Gamma(\sigma = K^{-\frac{1}{2}}, \mu = 1) \quad \text{where} \quad \mathbb{P}(w = x) \propto x^{K-1} e^{-Kx},$$

One should note that increasing the value of K reduces the overall noise level.

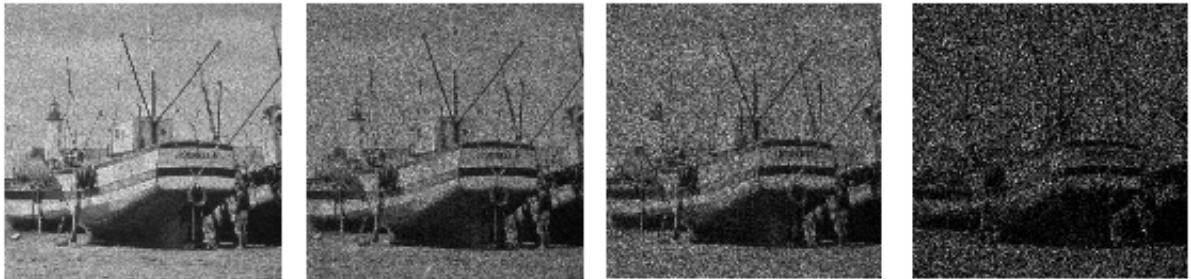


Figure 8.31: Noisy images with multiplicative noise, with varying σ .

Figure ?? shows an example of such image formation for a varying number $K = 1/\sigma^2$ of averaged images. A simple variance stabilization transform is

$$\varphi(x) = \log(x) - c$$

where

$$c = \mathbb{E}(\log(w)) = \psi(K) - \log(K) \quad \text{where} \quad \psi(x) = \Gamma'(x)/\Gamma(x)$$

and where Γ is the Gamma function that generalizes the factorial function to non-integer. One thus has

$$\varphi(f)_n = \varphi(f_0)_n + z_n,$$

where $z_n = \log(w) - c$ is a zero-mean additive noise.

Figure 8.32 shows the effect of this variance stabilization on the repartition of w and z .

Figure 8.33 shows that for moderate noise level σ , variance stabilization improves over non-stabilized denoising.

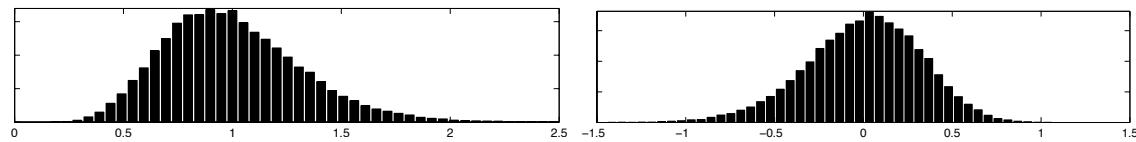


Figure 8.32: Histogram of multiplicative noise before (left) and after (right) stabilization.



Figure 8.33: Left: noisy image, center: denoising without variance stabilization, right: denoising after variance stabilization.

Chapter 9

Variational Priors and Regularization

9.1 Sobolev and Total Variation Priors

The simplest prior are obtained by integrating local differential quantity over the image. They corresponds to norms in functional spaces that imposes some smoothness on the signal of the image. We detail here the Sobolev and the total variation priors, that are the most popular in image processing.

9.1.1 Continuous Priors

In the following, we consider either continuous functions $f \in L^2([0, 1]^2)$ or discrete vectors $f \in \mathbb{R}^N$, and consider continuous priors and there discrete counterparts in Section 9.1.2.

Sobolev prior. The prior energy $J(f) \in \mathbb{R}$ is intended to be low for images in a class $f \in \Theta$. The class of uniformly smooth functions detailed in Section 6.2.1 corresponds to functions in Sobolev spaces. A simple prior derived from this Sobolev class is thus

$$J_{\text{Sob}}(f) = \frac{1}{2} \|f\|_{\text{Sob}}^2 = \frac{1}{2} \int \|\nabla f(x)\|^2 dx, \quad (9.1)$$

where ∇f is the gradient in the sense of distributions.

Total variation prior. To take into account discontinuities in images, one considers a total variation energy, introduced in Section 6.2.3. It was introduced for image denoising by Rudin, Osher and Fatemi [34]

The total variation of a smooth image f is defined as

$$J_{\text{TV}}(f) = \|f\|_{\text{TV}} = \int \|\nabla_x f\| dx. \quad (9.2)$$

This energy extends to non-smooth functions of bounded variations $f \in \text{BV}([0, 1]^2)$. This class contains indicators functions $f = 1_\Omega$ of sets Ω with a bounded perimeter $|\partial\Omega|$.

The total variation norm can be computed alternatively using the co-area formula (6.12), which shows in particular that $\|1_\Omega\|_{\text{TV}} = |\partial\Omega|$.

9.1.2 Discrete Priors

An analog image $f \in L^2([0, 1]^2)$ is discretized through an acquisition device to obtain a discrete image $f \in \mathbb{R}^N$. Image processing algorithms work on these discrete data, and we thus need to define discrete priors for finite dimensional images.

Discrete gradient. Discrete Sobolev and total variation priors are obtained by computing finite differences approximations of derivatives, using for instance forward differences

$$\begin{aligned}\delta_1 f_{n_1, n_2} &= f_{n_1+1, n_2} - f_{n_1, n_2} \\ \delta_2 f_{n_1, n_2} &= f_{n_1, n_2+1} - f_{n_1, n_2},\end{aligned}$$

and one can use higher order schemes to process more precisely smooth functions. One should be careful with boundary conditions, and we consider here for simplicity periodic boundary conditions, which correspond to computing the indexes $n_i + 1$ modulo N . More advanced symmetric boundary conditions can be used as well to avoid boundary artifacts.

A discrete gradient is defined as

$$\nabla f_n = (\delta_1 f_n, \delta_2 f_n) \in \mathbb{R}^2$$

which corresponds to a mapping from images to vector fields

$$\nabla : \mathbb{R}^N \longrightarrow \mathbb{R}^{N \times 2}.$$

Figure 9.1 shows examples of gradient vectors. They point in the direction of the largest slope of the function discretized by f . Figure 9.2 shows gradients and their norms displayed as an image. Regions of high gradients correspond to large intensity variations, and thus typically to edges or textures.

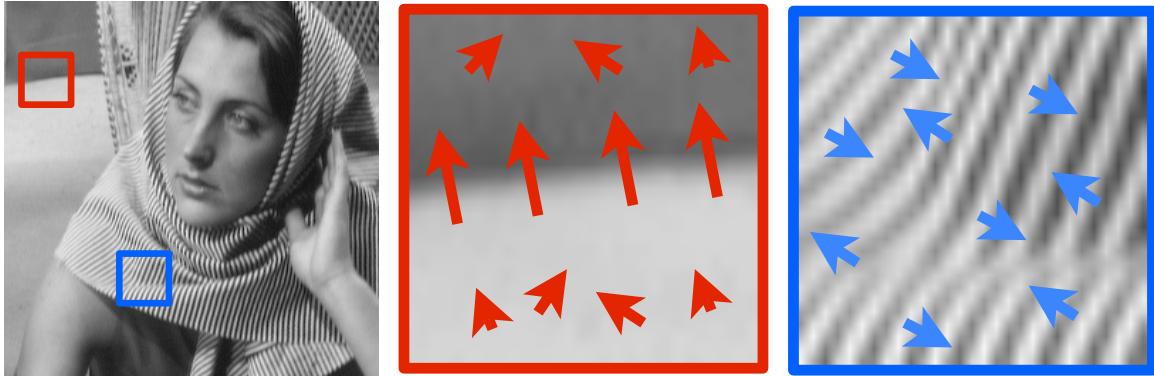


Figure 9.1: Discrete gradient vectors.

Discrete divergence. One can also use backward differences,

$$\begin{aligned}\tilde{\delta}_1 f_{n_1, n_2} &= f_{n_1, n_2} - f_{n_1-1, n_2} \\ \tilde{\delta}_2 f_{n_1, n_2} &= f_{n_1, n_2} - f_{n_1, n_2-1}.\end{aligned}$$

They are dual to the forward differences, so that

$$\delta_i^* = -\tilde{\delta}_i,$$

which means that

$$\forall f, g \in \mathbb{R}^N, \quad \langle \delta_i f, g \rangle = -\langle f, \tilde{\delta}_i g \rangle,$$

which is a discrete version of the integration by part formula

$$\int_0^1 f' g = - \int_0^1 f g'$$

for smooth periodic functions on $[0, 1]$.

A divergence operator is defined using backward differences,

$$\operatorname{div}(v)_n = \tilde{\delta}_1 v_{1,n} + \tilde{\delta}_2 v_{2,n},$$

and corresponds to a mapping from vector fields to images

$$\operatorname{div} : \mathbb{R}^{N \times 2} \longrightarrow \mathbb{R}^N.$$

It is related to the dual of the gradient

$$\operatorname{div} = -\nabla^*$$

which means that

$$\forall f \in \mathbb{R}^N, \forall v \in \mathbb{R}^{N \times 2}, \quad \langle \nabla f, v \rangle_{\mathbb{R}^{N \times 2}} = -\langle f, \operatorname{div}(v) \rangle_{\mathbb{R}^N}$$

which corresponds to a discrete version of the divergence theorem.

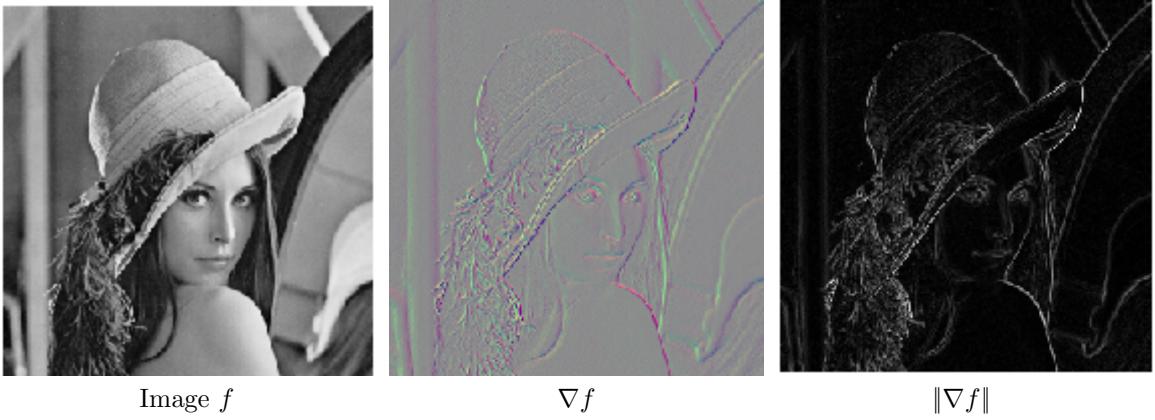


Figure 9.2: Discrete operators.

Discrete laplacian. A general definition of a Laplacian is

$$\Delta f = \operatorname{div}(\nabla f),$$

which corresponds to a semi-definite negative operator.

For discrete images, and using the previously defined gradient and divergence, it is a local high pass filter

$$\Delta f_n = \sum_{p \in V_4(n)} f_p - 4f_n, \tag{9.3}$$

that approximates the continuous second order derivative

$$\frac{\partial^2 f}{\partial x_1^2}(x) + \frac{\partial^2 f}{\partial x_2^2} \approx N^2 \Delta f_n \quad \text{for } x = n/N.$$

Laplacian operators thus correspond to filterings. A continuous Laplacian is equivalently defined over the Fourier domain in diagonal form as

$$g = \Delta f \implies \hat{g}(\omega) = \|\omega\|^2 \hat{f}(\omega)$$

and the discrete Laplacian (9.3) as

$$g = \Delta f \implies \hat{g}_\omega = \rho_\omega^2 \hat{f}(\omega) \quad \text{where} \quad \rho_\omega^2 = \sin\left(\frac{\pi}{N}\omega_1\right)^2 + \sin\left(\frac{\pi}{N}\omega_2\right)^2. \tag{9.4}$$

Discrete energies. A discrete Sobolev energy is obtained by using the ℓ^2 norm of the discrete gradient vector field

$$J_{\text{Sob}}(f) = \frac{1}{2} \sum_n (\delta_1 f_n)^2 + (\delta_2 f_n)^2 = \frac{1}{2} \|\nabla f\|^2. \quad (9.5)$$

Similarly, a discrete TV energy is defined as the ℓ^1 norm of the gradient field

$$J_{\text{TV}}(f) = \sum_n \sqrt{(\delta_1 f_n)^2 + (\delta_2 f_n)^2} = \|\nabla f\|_1 \quad (9.6)$$

where the ℓ^1 norm of a vector field $v \in \mathbb{R}^{N \times 2}$ is

$$\|v\|_1 = \sum_n \|v_n\| \quad (9.7)$$

where $v_n \in \mathbb{R}^2$.

9.2 PDE and Energy Minimization

Image smoothing is obtained by minimizing the prior using a gradient descent.

9.2.1 General Flows

The gradient of the prior $J : \mathbb{R}^N \rightarrow \mathbb{R}$ is a vector $\text{grad } J(f)$. It describes locally up to the first order the variation of the prior

$$J(f + \varepsilon) = J(f) + \langle \varepsilon, \text{grad } J(f) \rangle + o(\|\varepsilon\|).$$

If J is a smooth function of the image f , a discrete energy minimization is obtained through a gradient descent

$$f^{(k+1)} = f^{(k)} - \tau \text{grad } J(f^{(k)}), \quad (9.8)$$

where the step size τ must be small enough to guarantee convergence.

For infinitesimal step size τ , one replaces the discrete parameter k by a continuous time, and the flow

$$t > 0 \longmapsto f_t \in \mathbb{R}^N$$

solves the following partial differential equation

$$\frac{\partial f_t}{\partial t} = -\text{grad } J(f_t) \quad \text{and} \quad f_0 = f. \quad (9.9)$$

The gradient descent can be seen as an explicit discretization in time of this PDE at times $t_k = k\tau$.

9.2.2 Heat Flow

The heat flow corresponds to the instantiation of the generic PDE (9.9) to the case of the Sobolev energies $J_{\text{Sob}}(f)$ defined for continuous function in (9.1) and for discrete images in (9.5).

Since it is a quadratic energy, its gradient is easily computed

$$J(f + \varepsilon) = \frac{1}{2} \|\nabla f + \nabla \varepsilon\|^2 = J(f) - \langle \Delta f, \varepsilon \rangle + o(\|\varepsilon\|^2),$$

so that

$$\text{grad } J_{\text{Sob}}(f) = -\Delta f.$$

Figure 9.4, left, shows an example of Laplacian. It is typically large (positive or negative) near edges.

The heat flow is thus

$$\frac{\partial f_t}{\partial t}(x) = -(\text{grad } J(f_t))(x) = \Delta f_t(x) \quad \text{and} \quad f_0 = f. \quad (9.10)$$



Figure 9.3: Display of f_t for increasing time t for heat flow (top row) and TV flow (bottom row).

Continuous in space. For continuous images and an unbounded domain \mathbb{R}^2 , the PDE (9.10) has an explicit solution as a convolution with a Gaussian kernel of increasing variance as time increases

$$f_t = f \star h_t \quad \text{where} \quad h_t(x) = \frac{1}{4\pi t} e^{-\frac{\|x\|^2}{4t}}. \quad (9.11)$$

This shows the regularizing property of the heat flow, that operates a blurring to make the image more regular as time evolves.

Discrete in space. The discrete Sobolev energy (9.5) minimization defined a PDE flow that is discrete in space

$$\frac{\partial f_{n,t}}{\partial t} = -(\text{grad } J(f_t))_n = (\Delta f_t)_n.$$

It can be further discretized in time as done in (9.8) and leads to a fully discrete flow

$$f_n^{(k+1)} = f_n^{(k)} + \tau \left(\sum_{p \in V_4(n)} f_p - 4f_n \right) = (f \star h)_n$$

where $V_4(n)$ are the four neighbor to a pixel n . The flow thus corresponds to iterative convolutions

$$f^{(k)} = f \star h \star \dots \star h = f \star^k h.$$

where h is a discrete filter.

It can be shown to be stable and convergent if $\tau < 1/4$.

9.2.3 Total Variation Flows

Total variation gradient. The total variation energy J_{TV} , both continuous (9.2) and discrete (9.6) is not a smooth function of the image. For instance, the discrete J_{TV} is non-differentiable at an image f such that there exists a pixel n where $\nabla f_n = 0$.

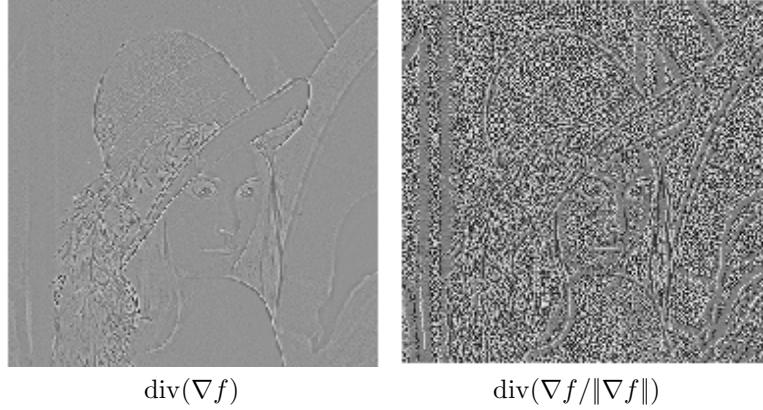


Figure 9.4: Discrete Laplacian and discrete TV gradient.

If $\nabla f_n \neq 0$, one can compute the gradient of the TV energy specialized at pixel n as

$$(\text{grad } J(f))_n = -\text{div} \left(\frac{\nabla f}{\|\nabla f\|} \right)_n$$

which exhibits a division by zero singularity for a point with vanishing gradient. Figure 9.4 shows an example of TV gradient, which appears noisy in smooth areas, because $\|\nabla f_n\|$ is small in such regions.

This non-differentiability makes impossible the definition of a gradient descent and a TV flow.

Regularized total variation. To avoid this issue, one can modify the TV energy, and define a smoothed TV prior

$$J_{\text{TV}}^\varepsilon(f) = \sum_n \sqrt{\varepsilon^2 + \|\nabla f_n\|^2} \quad (9.12)$$

where $\varepsilon > 0$ is a small regularization parameter. Figure 9.5 shows this effect of this regularization on the absolute value function.

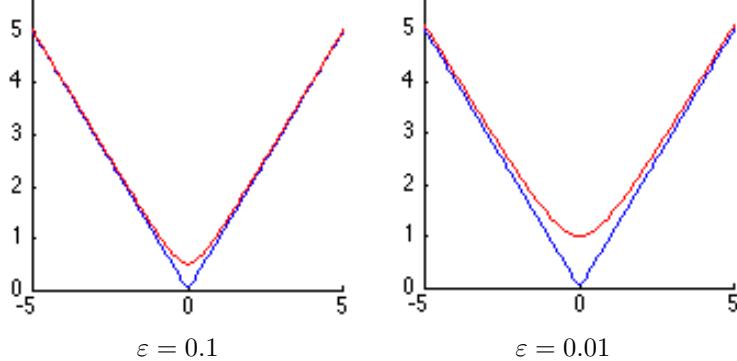


Figure 9.5: Regularized absolute value $x \mapsto \sqrt{x^2 + \varepsilon^2}$.

This smoothed TV energy is a differentiable function of the image, and its gradient is

$$\text{grad } J_{\text{TV}}^\varepsilon(f) = -\text{div} \left(\frac{\nabla f}{\sqrt{\varepsilon^2 + \|\nabla f\|^2}} \right). \quad (9.13)$$

One can see that this smoothing interpolate between TV and Sobolev, as

$$\text{grad}_f^\varepsilon \sim -\Delta/\varepsilon \quad \text{when } \varepsilon \rightarrow +\infty.$$

Figure 9.6 shows the evolution of this gradient for several value of the smoothing parameter.

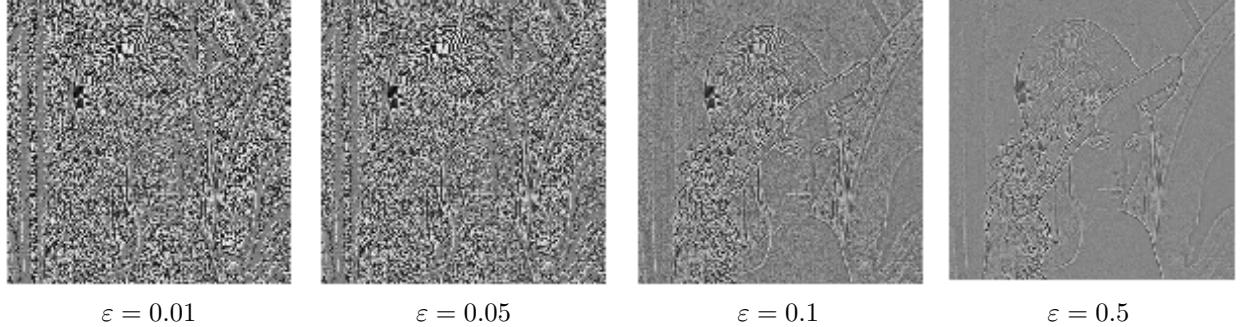


Figure 9.6: Regularized gradient norm $\sqrt{\|\nabla f(x)\|^2 + \varepsilon^2}$.

Regularized total variation flow. The smoothed total variation flow is then defined as

$$\frac{\partial f_t}{\partial t} = \text{div} \left(\frac{\nabla f_t}{\sqrt{\varepsilon^2 + \|\nabla f_t\|^2}} \right). \quad (9.14)$$

Choosing a small ε makes the flow closer to a minimization of the total variation, but makes the computation unstable.

In practice, the flow is computed with a discrete gradient descent (9.8). For the smoothed total variation flow to converge, one needs to impose that $\tau < \varepsilon/4$, which shows that being more faithful to the TV energy requires smaller time steps and thus slower algorithms.

Figure 9.3 shows a comparison between the heat flow and the total variation flow for a small value of ε . This shows that the TV flow smooth less the edges than heat diffusion, which is consistent with the ability of the TV energy to better characterize sharp edges.

9.2.4 PDE Flows for Denoising

PDE flows can be used to remove noise from an observation $f = f_0 + w$. As detailed in Section 8.1.2 a simple noise model assumes that each pixel is corrupted with a Gaussian noise $w_n \sim \mathcal{N}(0, \sigma)$, and that these perturbations are independent (white noise).

The denoising is obtained using the PDE flow within initial image f at time $t = 0$

$$\frac{\partial f_t}{\partial t} = -\text{grad}_{f_t} J \quad \text{and} \quad f_{t=0} = f.$$

An estimator $\tilde{f} = f_{t_0}$ is obtained for a well chose $t = t_0$. Figure 9.7 shows examples of Sobolev and TV flows for denoising.

Since f_t converges to a constant image when $t \rightarrow +\infty$, the choice of t_0 corresponds to a tradeoff between removing enough noise and not smoothing too much the edges in the image. This is usually a difficult task. During simulation, if one has access to the clean image f_0 , one can monitor the denoising error $\|f_0 - f_t\|$ and choose the $t = t_0$ that minimizes this error. Figure 9.8, top row, shows an example of this oracle estimation of the best stopping time.



Figure 9.7: Denoising using f_t displayed for various time t for Sobolev (top) and TV (bottom) flows.

9.3 Regularization for Denoising

Instead of performing a gradient descent flow for denoising as detailed in Section 9.2.4 and select a stopping time, one can formulate an optimization problem. The estimator is then defined as a solution of this optimization. This setup has the advantage as being well defined mathematically even for non-smooth priors such as the TV prior J_{TV} or the sparsity prior J_1 . Furthermore, this regularization framework is also useful to solve general inverse problems as detailed in Chapter ??.

9.3.1 Regularization

Given some noisy image $f = f_0 + w$ of N pixels and a prior J , we consider the convex optimization problem

$$f_\lambda^* \in \operatorname{argmin}_{g \in \mathbb{R}^N} \frac{1}{2} \|f - g\|^2 + \lambda J(g), \quad (9.15)$$

where $\lambda > 0$ is a Lagrange multiplier parameter that weights the influence of the data fitting term $\|f - g\|^2$ and the regularization term $J(g)$.

If one has at his disposal a clean original image f_0 , one can minimize the denoising error $\|f_\lambda^* - f_0\|$, but it is rarely the case. In practice, this parameter should be adapted to the noise level and to the regularity of the unknown image f_0 , which might be a non trivial task.

We note that since we did not impose J to be convex, the problem (9.15) might have several optimal solutions.

An estimator is thus defined as

$$\tilde{f} = f_\lambda^*$$

for a well chosen λ .

If J is differentiable and convex, one can compute the solution of (9.15) through a gradient descent

$$f^{(k+1)} = f^{(k)} - \tau \left(f^{(k)} - \lambda \operatorname{grad} J(f^{(k)}) \right) \quad (9.16)$$

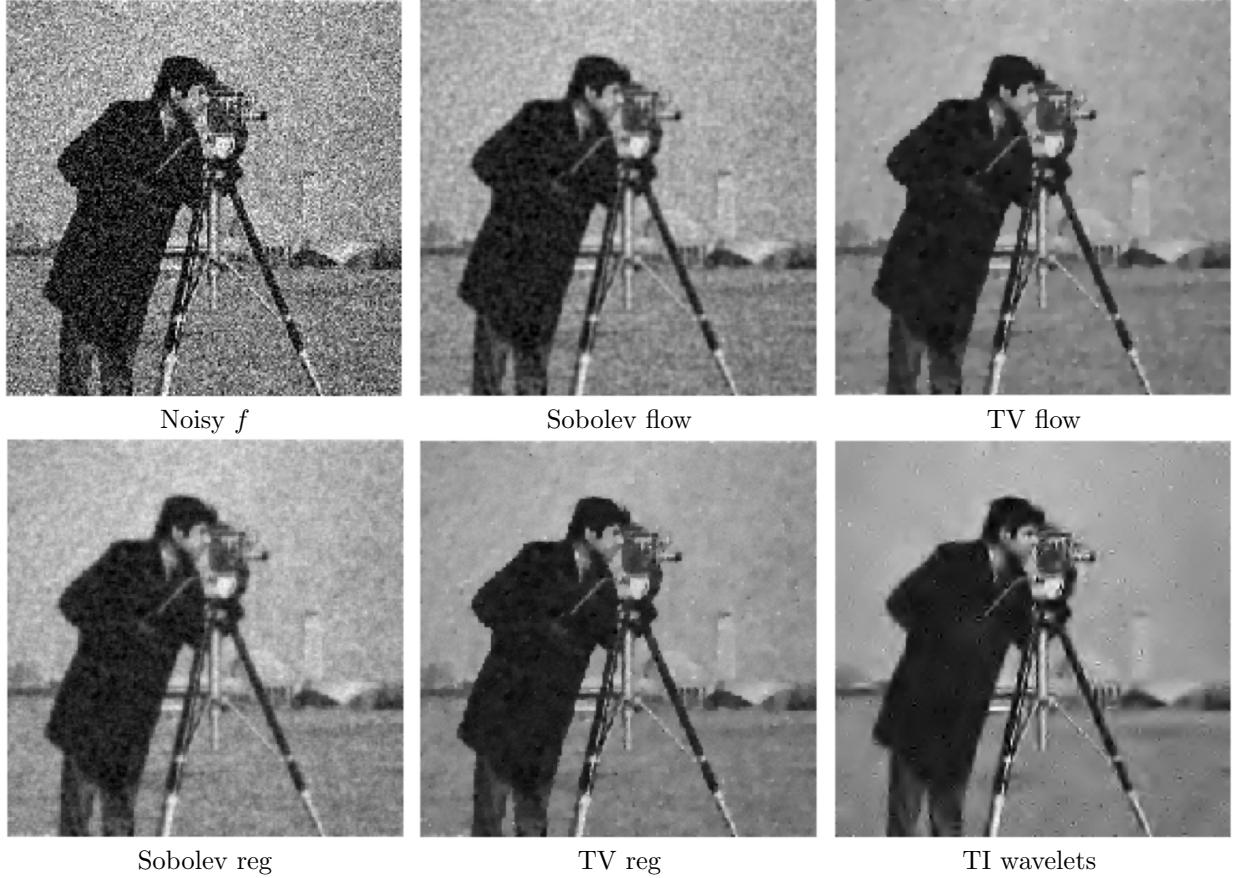


Figure 9.8: Denoising using PDE flows and regularization.

where the descent step size $\tau > 0$ should be small enough. This gradient descent is similar to the time-discretized minimization flow (9.8), excepted that the data fitting term prevent the flow to converge to a constant image.

Unfortunately, priors such as the total variation J_{TV} or the sparsity J_1 are non-differentiable. Some priors such as the ideal sparsity J_0 might even be non-convex. This makes the simple gradient descent not usable to solve for (9.15). In the following Section we show how to compute f_λ^* for several priors.

9.3.2 Sobolev Regularization

The discrete Sobolev prior defined in (9.5) is differentiable, and the gradient descent (9.16) reads

$$f^{(k+1)} = (1 - \tau)f^{(k)} + \tau f - \tau\lambda\Delta J(f^{(k)}).$$

Since $J(f) = \|\nabla f\|^2$ is quadratic, one can use a conjugate gradient descent, which converges faster.

The solution f_λ^* can be computed in closed form as the solution of a linear system

$$f_\lambda^* = (\text{Id}_N - \lambda\Delta)^{-1}f,$$

which shows that the regularization (9.15) is computing an estimator that depends linearly on the observations f .

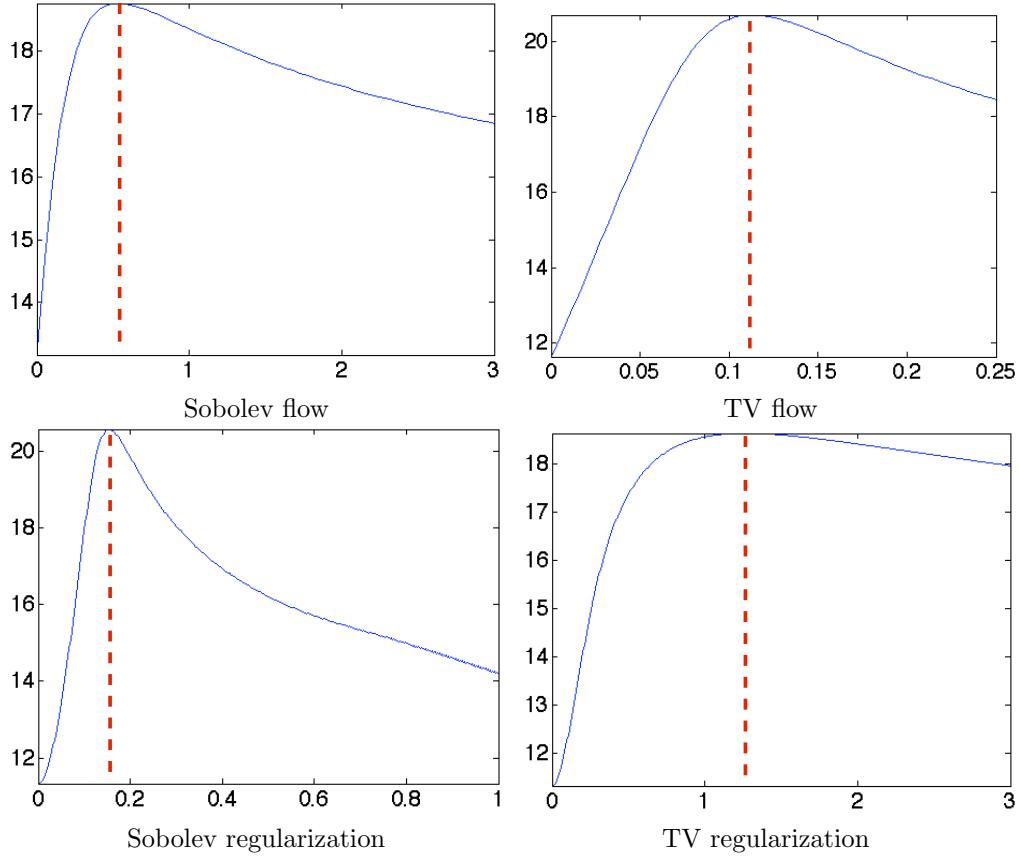


Figure 9.9: SNR as a function of time t for flows (top) and λ for regularization (bottom).

If the differential operators are computed with periodic boundary conditions, this linear system can be solved exactly over the Fourier domain

$$(\hat{f}_\lambda^*)_\omega = \frac{1}{1 + \lambda \rho_\omega^2} \hat{f}_\omega \quad (9.17)$$

where ρ_ω depends on the discretization of the Laplacian, see for instance (9.4).

Equation (9.17) shows that denoising using Sobolev regularization corresponds to a low pass filtering, whose strength is controlled by λ . This should be related to the solution (9.11) of the heat equation, which also corresponds to a filtering, but using a Gaussian low pass kernel parameterized by its variance t^2 .

This Sobolev regularization denoising is a particular case of the linear estimator considered in Section 8.2. The selection of the parameter λ is related to the selection of an optimal filter as considered in Section 8.2.2, but with the restriction that the filter is computed in a parametric family.

9.3.3 TV Regularization

The total variation prior J_{TV} defined in (9.6) is non-differentiable. One can either use a smoothed approximation of the prior, or use an optimization algorithm not based on a gradient descent.

The TV prior can be approximated to obtain the prior $J_{\text{TV}}^\varepsilon(g)$ defined in (9.12), which is differentiable with respect to g . Using the gradient of this prior defined in (9.13), one obtains the following instantiation

of the gradient descent (9.16)

$$f^{(k+1)} = (1 - \tau)f^{(k)} + \tau f + \lambda\tau \operatorname{div} \left(\frac{\nabla f_t}{\sqrt{\varepsilon^2 + \|\nabla f_t\|^2}} \right). \quad (9.18)$$

which converge to the unique minimizer f_λ^* of (9.15).

Section 13.4.1 details a better alternative which does not require introducing this ε smoothing.

Chapter 10

Inverse Problems

The main references for this chapter are [28, 35, 19].

10.1 Inverse Problems Regularization

Increasing the resolution of signals and images requires to solve an ill posed inverse problem. This corresponds to inverting a linear measurement operator that reduces the resolution of the image. This chapter makes use of convex regularization introduced in Chapter ?? to stabilize this inverse problem.

We consider a (usually) continuous linear map $\Phi : \mathcal{S} \rightarrow \mathcal{H}$ where \mathcal{S} can be an Hilbert or a more general Banach space. This operator is intended to capture the hardware acquisition process, which maps a high resolution unknown signal $f_0 \in \mathcal{S}$ to a noisy low-resolution observation

$$y = \Phi f_0 + w \in \mathcal{H}$$

where $w \in \mathcal{H}$ models the acquisition noise. In this section, we do not use a random noise model, and simply assume that $\|w\|_{\mathcal{H}}$ is bounded.

In most applications, $\mathcal{H} = \mathbb{R}^P$ is finite dimensional, because the hardware involved in the acquisition can only record a finite (and often small) number P of observations. Furthermore, in order to implement numerically a recovery process on a computer, it also makes sense to restrict the attention to $\mathcal{S} = \mathbb{R}^N$, where N is number of point on the discretization grid, and is usually very large, $N \gg P$. However, in order to perform a mathematical analysis of the recovery process, and to be able to introduce meaningful models on the unknown f_0 , it still makes sense to consider infinite dimensional functional space (especially for the data space \mathcal{S}).

The difficulty of this problem is that the direct inversion of Φ is in general impossible or not advisable because Φ^{-1} have a large norm or is even discontinuous. This is further increased by the addition of some measurement noise w , so that the relation $\Phi^{-1}y = f_0 + \Phi^{-1}w$ would leads to an explosion of the noise $\Phi^{-1}w$.

We now gives a few representative examples of forward operators Φ .

Denoising. The case of the identity operator $\Phi = \text{Id}_{\mathcal{S}}$, $\mathcal{S} = \mathcal{H}$ corresponds to the classical denoising problem, already treated in Chapters ?? and ??.

De-blurring and super-resolution. For a general operator Φ , the recovery of f_0 is more challenging, and this requires to perform both an inversion and a denoising. For many problem, this two goals are in contradiction, since usually inverting the operator increases the noise level. This is for instance the case for the deblurring problem, where Φ is a translation invariant operator, that corresponds to a low pass filtering with some kernel h

$$\Phi f = f \star h. \tag{10.1}$$

One can for instance consider this convolution over $\mathcal{S} = \mathcal{H} = L^2(\mathbb{T}^d)$, see Proposition 3. In practice, this convolution is followed by a sampling on a grid $\Phi f = \{(f \star h)(x_k) ; 0 \leq k < P\}$, see Figure 10.1, middle, for an example of a low resolution image Φf_0 . Inverting such operator has important industrial application to upsample the content of digital photos and to compute high definition videos from low definition videos.

Interpolation and inpainting. Inpainting corresponds to interpolating missing pixels in an image. This is modelled by a diagonal operator over the spacial domain

$$(\Phi f)(x) = \begin{cases} 0 & \text{if } x \in \Omega, \\ f(x) & \text{if } x \notin \Omega. \end{cases} \quad (10.2)$$

where $\Omega \subset [0, 1]^d$ (continuous model) or $\{0, \dots, N - 1\}$ which is then a set of missing pixels. Figure 10.1, right, shows an example of damaged image Φf_0 .

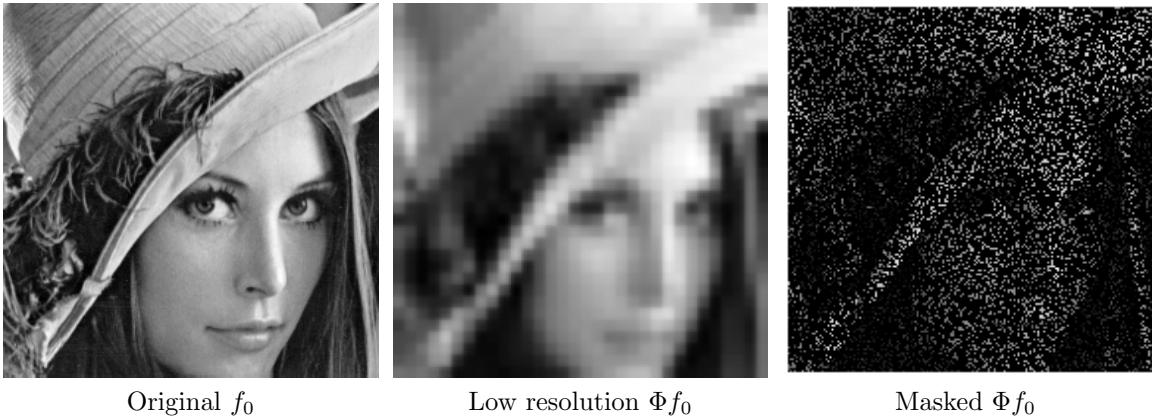


Figure 10.1: Example of inverse problem operators.

Medical imaging. Most medical imaging acquisition device only gives indirect access to the signal of interest, and is usually well approximated by such a linear operator Φ . In scanners, the acquisition operator is the Radon transform, which, thanks to the Fourier slice theorem, is equivalent to partial Fourier measurements along radial lines. Medical resonance imaging (MRI) is also equivalent to partial Fourier measures

$$\Phi f = \{\hat{f}(x) ; x \in \Omega\}. \quad (10.3)$$

Here, Ω is a set of radial line for a scanner, and smooth curves (e.g. spirals) for MRI.

Other indirect applications are obtained by electric or magnetic fields measurements of the brain activity (corresponding to MEG/EEG). Denoting $\Omega \subset \mathbb{R}^3$ the region around which measurements are performed (e.g. the head), in a crude approximation of these measurements, one can assume $\Phi f = \{(\psi \star f)(x) ; x \in \partial\Omega\}$ where $\psi(x)$ is a kernel accounting for the decay of the electric or magnetic field, e.g. $\psi(x) = 1/\|x\|^2$.

10.2 Theoretical Study of Quadratic Regularization

We now give a glimpse on the typical approach to obtain theoretical guarantee on recovery quality in the case of Hilbert space. The goal is not to be exhaustive, but rather to insist on the modelling hypothesis, namely smoothness implies a so called “source condition”, and the inherent limitations of quadratic methods (namely slow rates and the impossibility to recover information in $\ker(\Phi)$, i.e. to achieve super-resolution).

10.2.1 Singular Value Decomposition

Finite dimension. Let us start by the simple finite dimensional case $\Phi \in \mathbb{R}^{P \times N}$ so that $\mathcal{S} = \mathbb{R}^N$ and $\mathcal{H} = \mathbb{R}^P$ are Hilbert spaces. In this case, the Singular Value Decomposition (SVD) is the key to analyze the operator very precisely, and to describe linear inversion process.

Proposition 20 (SVD). *There exists $(U, V) \in \mathbb{R}^{N \times R} \times \mathbb{R}^{P \times R}$, where $R = \text{rank}(\Phi) = d(\text{Im}(\Phi))$, with $U^\top U = V^\top V = \text{Id}_R$, i.e. having orthogonal columns $(u_m)_{m=1}^R \subset \mathbb{R}^N$, $(v_m)_{m=1}^R \subset \mathbb{R}^P$, and $(\sigma_m)_{m=1}^R$ with $\sigma_m > 0$, such that*

$$\Phi = U \text{diag}_m(\sigma_m) V^\top = \sum_{m=1}^R \sigma_m u_m v_m^\top. \quad (10.4)$$

Proof. We first analyze the problem, and notice that if $\Phi = U \Sigma V^\top$ with $\Sigma = \text{diag}_m(\sigma_m)$, then $\Phi \Phi^\top = U \Sigma^2 U^\top$ and then $V^\top = \Sigma^{-1} U^\top \Phi$. We can use this insight. Since $\Phi \Phi^\top$ is a positive symmetric matrix, we write its eigendecomposition as $\Phi \Phi^\top = U \Sigma^2 U^\top$ where $\Sigma = \text{diag}_{m=1}^R(\sigma_m)$ with $\sigma_m > 0$. We then define $V \stackrel{\text{def.}}{=} \Phi^\top U \Sigma^{-1}$. One then verifies that

$$V^\top V = (\Sigma^{-1} U^\top \Phi)(\Phi^\top U \Sigma^{-1}) = \Sigma^{-1} U^\top (U \Sigma^2 U^\top) U \Sigma^{-1} = \text{Id}_P \quad \text{and} \quad U \Sigma V^\top = U \Sigma \Sigma^{-1} U^\top \Phi = \Phi.$$

□

This theorem is still valid with complex matrice, replacing $^\top$ by $*$. Expression (10.4) describes Φ as a sum of rank-1 matrices $u_m v_m^\top$. One usually order the singular values $(\sigma_m)_m$ in decaying order $\sigma_1 \geq \dots \geq \sigma_R$. If these values are different, then the SVD is unique up to ± 1 sign change on the singular vectors.

The left singular vectors U is an orthonormal basis of $\text{Im}(\Phi)$, while the right singular values is an orthonormal basis of $\text{Im}(\Phi^\top) = \ker(\Phi)^\perp$. The decomposition (10.4) is often call the “reduced” SVD because one has only kept the R non-zero singular values. The “full” SVD is obtained by completing U and V to define orthonormal bases of the full spaces \mathbb{R}^P and \mathbb{R}^N . Then Σ becomes a rectangular matrix of size $P \times N$.

A typical example is for $\Phi f = f \star h$ over $\mathbb{R}^P = \mathbb{R}^N$, in which case the Fourier transform diagonalizes the convolution, i.e.

$$\Phi = (u_m)_m^* \text{diag}(\hat{h}_m)(u_m)_m \quad (10.5)$$

where $(u_m)_n \stackrel{\text{def.}}{=} \frac{1}{\sqrt{N}} e^{\frac{2i\pi}{N} nm}$ so that the singular values are $\sigma_m = |\hat{h}_m|$ (removing the zero values) and the singular vectors are $(u_m)_n$ and $(v_m \theta_m)_n$ where $\theta_m \stackrel{\text{def.}}{=} |\hat{h}_m|/\hat{h}_m$ is a unit complex number.

Computing the SVD of a full matrix $\Phi \in \mathbb{R}^{N \times N}$ has complexity N^3 .

Compact operators. One can extend the decomposition to compact operators $\Phi : \mathcal{S} \rightarrow \mathcal{H}$ between separable Hilbert space. A compact operator is such that ΦB_1 is pre-compact where $B_1 = \{s \in \mathcal{S} ; \|s\| \leq 1\}$ is the unit-ball. This means that for any sequence $(\Phi s_k)_k$ where $s_k \in B_1$ one can extract a converging sub-sequence. Note that in infinite dimension, the identity operator $\Phi : \mathcal{S} \rightarrow \mathcal{S}$ is never compact.

Compact operators Φ can be shown to be equivalently defined as those for which an expansion of the form (10.4) holds

$$\Phi = \sum_{m=1}^{+\infty} \sigma_m u_m v_m^\top \quad (10.6)$$

where $(\sigma_m)_m$ is a decaying sequence onverging to 0, $\sigma_m \rightarrow 0$. Here in (10.6) convergence holds in the operator norm, which is the algebra norm on linear operator inherited from those of \mathcal{S} and \mathcal{H}

$$\|\Phi\|_{\mathcal{L}(\mathcal{S}, \mathcal{H})} \stackrel{\text{def.}}{=} \min_{\|\Phi u\|_{\mathcal{H}}} \|u\|_{\mathcal{S}} \leq 1.$$

For Φ having an SVD decomposition (10.6), $\|\Phi\|_{\mathcal{L}(\mathcal{S}, \mathcal{H})} = \sigma_1$.

When $\sigma_m = 0$ for $m > R$, Φ has a finite rank $R = d(\text{Im}(\Phi))$. As we explain in the sections bellow, when using *linear* recovery methods (such as quadratic regularization), the inverse problem is equivalent to a finite

dimensional problem, since one can restrict its attention to functions in $\ker(\Phi)^\perp$ which has dimension R . Of course, this is not true anymore when one can retrieve function inside $\ker(\Phi)$, which is often referred to as a “super-resolution” effect of non-linear methods. Another definition of compact operator is that they are the limit of finite rank operator. They are thus in some sense the extension of finite dimensional matrices, and are the correct setting to model ill-posed inverse problems. This definition can be extended to linear operator between Banach spaces, but this conclusion does not hold.

Typical example of compact operator are matrix-like operator with a continuous kernel $k(x, y)$ for $(x, y) \in \Omega$ where Ω is a compact sub-set of \mathbb{R}^d (or the torus \mathbb{T}^d), i.e.

$$(\Phi f)(x) = \int_{\Omega} k(x, y) f(y) dy$$

where dy is the Lebesgue measure. An example of such a setting which generalizes (10.5) is when $\Phi f = f \star h$ on $\mathbb{T}^d = (\mathbb{R}/2\pi\mathbb{Z})^d$, which corresponds to a translation invariant kernel $k(x, y) = h(x - y)$, in which case $u_m(x) = (2\pi)^{-d/2} e^{i\omega x}$, $\sigma_m = |\hat{f}_m|$. Another example on $\Omega = [0, 1]$ is the integration, $(\Phi f)(x) = \int_0^x f(y) dy$, which corresponds to k being the indicator of the “triangle”, $k(x, y) = 1_{x \leq y}$.

Pseudo inverse. In the case where $w = 0$, it makes to try to directly solve $\Phi f = y$. The two obstruction for this is that one not necessarily has $y \in \text{Im}(\Phi)$ and even so, there are an infinite number of solutions if $\ker(\Phi) \neq \{0\}$. The usual workaround is to solve this equation in the least square sense

$$f^+ \stackrel{\text{def.}}{=} \underset{\Phi f = y^+}{\operatorname{argmin}} \|f\|_{\mathcal{S}} \quad \text{where} \quad y^+ = \operatorname{Proj}_{\text{Im}(\Phi)}(y) = \underset{z \in \text{Im}(\Phi)}{\operatorname{argmin}} \|y - z\|_{\mathcal{H}}.$$

The following proposition shows how to compute this least square solution using the SVD and by solving linear systems involving either $\Phi\Phi^*$ or $\Phi^*\Phi$.

Proposition 21. *One has*

$$f^+ = \Phi^+ y \quad \text{where} \quad \Phi^+ = V \operatorname{diag}_m(1/\sigma_m) U^*. \quad (10.7)$$

In case that $\text{Im}(\Phi) = \mathcal{H}$, one has $\Phi^+ = \Phi^*(\Phi\Phi^*)^{-1}$. In case that $\ker(\Phi) = \{0\}$, one has $\Phi^+ = (\Phi^*\Phi)^{-1}\Phi^*$.

Proof. Since U is an ortho-basis of $\text{Im}(\Phi)$, $y^+ = UU^*y$, and thus $\Phi f = y^+$ reads $U\Sigma V^*f = UU^*y$ and hence $V^*f = \Sigma^{-1}U^*y$. Decomposition orthogonally $f = f_0 + r$ where $f_0 \in \ker(\Phi)^\perp$ and $r \in \ker(\Phi)$, one has $f_0 = VV^*f = V\Sigma^{-1}U^*y = \Phi^+y$ is a constant. Minimizing $\|f\|^2 = \|f_0\|^2 + \|r\|^2$ is thus equivalent to minimizing $\|r\|$ and hence $r = 0$ which is the desired result. If $\text{Im}(\Phi) = \mathcal{H}$, then $R = N$ so that $\Phi\Phi^* = U\Sigma^2U^*$ is the eigen-decomposition of an invertible and $(\Phi\Phi^*)^{-1} = U\Sigma^{-2}U^*$. One then verifies $\Phi^*(\Phi\Phi^*)^{-1} = V\Sigma U^*U\Sigma^{-2}U^*$ which is the desired result. One deals similarly with the second case. \square

For convolution operators $\Phi f = f \star h$, then

$$\Phi^+y = y \star h^+ \quad \text{where} \quad \hat{h}_m^+ = \begin{cases} \hat{h}_m^{-1} & \text{if } \hat{h}_m \neq 0 \\ 0 & \text{if } \hat{h}_m = 0. \end{cases}.$$

10.2.2 Tikonov Regularization

Regularized inverse. When there is noise, using formula (10.7) is not acceptable, because then

$$\Phi^+y = \Phi^+\Phi f_0 + \Phi^+w = f_0^+ + \Phi^+w \quad \text{where} \quad f_0^+ \stackrel{\text{def.}}{=} \operatorname{Proj}_{\ker(\Phi)^\perp},$$

so that the recovery error is $\|\Phi^+y - f_0^+\| = \|\Phi^+w\| \geq \|w\|/\sigma_R$. The noise is thus amplified by the inverse $1/\sigma_R$ of the smallest amplitude non-zero singular values, which can be very large. In infinite dimension, one typically has $R = +\infty$, so that the inverse is actually not bounded (discontinuous). It is thus mandatory to replace Φ^+ by a regularized approximate inverse, which should have the form

$$\Phi_\lambda^+ = V \operatorname{diag}_m(\mu_\lambda(\sigma_m)) U^* \quad (10.8)$$

where μ_λ , indexed by some parameter $\lambda > 0$, is a regularization of the inverse, that should typically satisfies

$$\mu_\lambda(\sigma) \leq C_\lambda < +\infty \quad \text{and} \quad \lim_{\lambda \rightarrow 0} \mu_\lambda(\sigma) = \frac{1}{\sigma}$$

Variational regularization. A typical example of such regularized inverse is obtained by considering a penalized least square involving a regularization functional

$$f_\lambda \stackrel{\text{def.}}{=} \operatorname{argmin}_{f \in \mathcal{S}} \|y - \Phi f\|_{\mathcal{H}}^2 + \lambda J(f) \quad (10.9)$$

where J is some regularization functional which should at least be continuous on \mathcal{S} . The simplest example is the quadratic norm $J = \|\cdot\|_{\mathcal{S}}^2$,

$$f_\lambda \stackrel{\text{def.}}{=} \operatorname{argmin}_{f \in \mathcal{S}} \|y - \Phi f\|_{\mathcal{H}}^2 + \lambda \|f\|^2 \quad (10.10)$$

which is indeed a special case of (10.8) as we now prove.

Proposition 22. *The solution of (10.10) has the form $f_\lambda = \Phi_\lambda^+ y$ as defined in (10.8) for the specific choice of function*

$$\forall \sigma \in \mathbb{R}, \quad \mu_\lambda(\sigma) = \frac{\sigma}{\sigma^2 + \lambda}.$$

Proof. Problem (10.10) can be conveniently rewritten in the basis of singular vectors as

$$f_\lambda = \operatorname{argmin}_{f \in \operatorname{Im}(\Phi^*)} \sum_m (\sigma_m \langle f, v_m \rangle - \langle y, u_m \rangle)^2 + \lambda \langle f, v_m \rangle^2 + \lambda \|\operatorname{Proj}_{\ker(\Phi)^\perp} f\|^2. \quad (10.11)$$

This shows that necessarily $\operatorname{Proj}_{\ker(\Phi)^\perp} f_\lambda = 0$, i.e. $f_\lambda \in \operatorname{Im}(\Phi^*)$. The minimization (10.11) thus boils down to independent scalar minimization over each coordinate $f_m \stackrel{\text{def.}}{=} \langle f, v_m \rangle$ and the first order condition reads

$$\sigma_m (\sigma_m f_m - y_m) + \lambda f_m = 0 \quad \text{where} \quad y_m \stackrel{\text{def.}}{=} \langle y, u_m \rangle$$

which shows the desired formula. \square

The question is to understand how to choose λ as a function of the noise level $\|w\|_{\mathcal{H}}$ in order to guarantees that $f_\lambda \rightarrow f_0$ and furthermore establish convergence speed. One first needs to ensure at least $f_0 = f_0^+$, which in turns requires that $f_0 \in \operatorname{Im}(\Phi^*) = \ker(\Phi)^\perp$. Indeed, an important drawback of linear recovery methods (such as quadratic regularization) is that necessarily $f_\lambda \in \operatorname{Im}(\Phi^*) = \ker(\Phi)^\perp$ so that no information can be recovered inside $\ker(\Phi)$. Non-linear methods must be used to achieve a “super-resoltution” effect and recover this missing information.

Source condition. In order to ensure convergence speed, one quantify this condition and impose a so-called source condition of order β , which reads

$$f_0 \in \operatorname{Im}((\Phi^* \Phi)^\beta) = \operatorname{Im}(V \operatorname{diag}(\sigma_m^{2\beta}) V^*). \quad (10.12)$$

This condition means that there should exists $z \in \mathcal{S}$ such that $f_0 = V \operatorname{diag}(\sigma_m^{2\beta}) V^* z$, i.e. $z = V \operatorname{diag}(\sigma_m^{-2\beta}) V^* f_0$. Denoting $\rho \stackrel{\text{def.}}{=} \|z\|$, we thus in fact impose the following constraint

$$\sum_m \sigma_m^{-2\beta} \langle f, v_m \rangle^2 \leq \rho^2 < +\infty. \quad (S_{\beta, \rho})$$

This is a Sobolev-type constraint, similar to those imposed in 8.4. A prototypical example is for a low-pass filter $\Phi f = f \star h$ where h has a slow polynomial-like decay of frequency, i.e. $|\hat{h}_m| \sim 1/m^\alpha$ for large m . In this case, since v_m is the Fourier basis, the source condition $(S_{\beta, \rho})$ reads

$$\sum_m m^{2\alpha\beta} |\hat{f}_m|^2 \leq \rho^2 < +\infty,$$

which is a Sobolev ball of radius ρ and differential order $\alpha\beta$.

Sublinear convergence speed. The following theorem shows that this source condition leads to a convergence speed of the regularization.

Theorem 31. *Assuming the source condition $(S_{\beta,\rho})$ with $0 < \beta \leq 2$, then the solution of (10.10) for $\|w\| \leq \delta$ satisfies*

$$\|f_\lambda - f_0\| \leq C\rho^{\frac{1}{\beta+1}}\delta^{\frac{\beta}{\beta+1}}$$

for a constant C which depends only on β , and for a choice

$$\lambda \sim \delta^{\frac{2}{\beta+1}}\rho^{-\frac{2}{\beta+1}}.$$

Proof. Because of the source condition, $f_0 \in \text{Im}(\Phi^*)$. Denoting

$$f_\lambda^0 \stackrel{\text{def.}}{=} \Phi_\lambda^+(\Phi f_0)$$

so that $f_\lambda = f_\lambda^0 + \Phi_\lambda^+ w$, one has for any regularized inverse of the form (10.8)

$$\|f_\lambda - f_0\| \leq \|f_\lambda - f_\lambda^0\| + \|f_\lambda^0 - f_0\|. \quad (10.13)$$

The term $\|f_\lambda - f_\lambda^0\|$ is a variance term which account for residual noise, and thus decays when λ increases (more regularization). The term $\|f_\lambda^0 - f_0\|$ is independent of the noise, it is a bias term coming from the approximartion (smoothing) of f_0 , and thus increases when λ increases. The choice of an optimal λ thus results in a bias-variance tradeoff between these two terms. Assuming

$$\forall \sigma \geq 0, \quad \mu_\lambda(\sigma) \leq C_\lambda$$

the variance term is bounded as

$$\|f_\lambda - f_\lambda^0\|^2 = \|\Phi_\lambda^+ w\|^2 = \sum_m \mu_\lambda(\sigma_m)^2 w_m^2 \leq C_\lambda^2 \|w\|_{\mathcal{H}}^2.$$

The bias term is bounded as

$$\|f_\lambda^0 - f_0\|^2 = \sum_m (1 - \mu_\lambda(\sigma_m)\sigma_m)^2 f_{0,m}^2 = \sum_m ((1 - \mu_\lambda(\sigma_m)\sigma_m)\sigma_m^\beta)^2 \frac{f_{0,m}^2}{\sigma_m^{2\beta}} \leq D_{\lambda,\beta}^2 \rho^2 \quad (10.14)$$

where we assumed

$$\forall \sigma \geq 0, \quad |(1 - \mu_\lambda(\sigma)\sigma)\sigma^\beta| \leq D_{\lambda,\beta}. \quad (10.15)$$

Putting (10.14) and (10.15) together, one obtains

$$\|f_\lambda - f_0\| \leq C_\lambda \delta + D_{\lambda,\beta} \rho. \quad (10.16)$$

In the case of the regularization (10.10), one has $\mu_\lambda(\sigma) = \frac{\sigma}{\sigma^2 + \lambda}$, and thus $(1 - \mu_\lambda(\sigma)\sigma)\sigma^\beta = \frac{\lambda\sigma^\beta}{\sigma^2 + \lambda}$. For $\beta \leq 2$, one verifies (see Figure 16.8) that

$$C_\lambda = \frac{1}{2\sqrt{\lambda}} \quad \text{and} \quad D_{\lambda,\beta} = C_\beta \lambda^{\frac{\beta}{2}},$$

for some constant C_β . Equalizing the contributions of the two terms in (10.16) (a better constant would be reached by finding the best λ) leads to selecting $\frac{\delta}{\sqrt{\lambda}} = \lambda^{\frac{\beta}{2}}\rho$ i.e. $\lambda = (\delta/\rho)^{\frac{2}{\beta+1}}$. With this choice,

$$\|f_\lambda - f_0\| = O(\delta/\sqrt{\lambda}) = O(\delta(\delta/\rho)^{-\frac{1}{\beta+1}}) = O(\delta^{\frac{\beta}{\beta+1}}\rho^{\frac{1}{\beta+1}}).$$

□

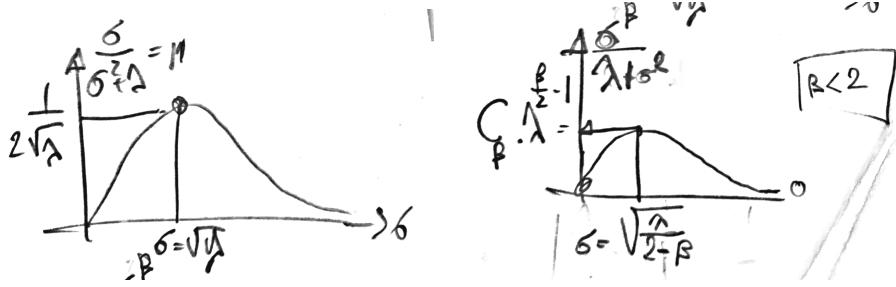


Figure 10.2: Bounding $\mu_\lambda(\sigma) \leq C_\lambda = \frac{1}{2\sqrt{\lambda}}$ (left) and $\frac{\sigma^\beta}{\lambda + \sigma^2} \leq D_{\lambda,\beta}/\lambda$ (right).

This theorem shows that using larger $\beta \leq 2$ leads to faster convergence rates as $\|w\|$ drops to zero. The rate (10.13) however suffers from a “saturation” effect, indeed, choosing $\beta > 2$ does not help (it gives the same rate as $\beta = 2$), and the best possible rate is thus

$$\|f_\lambda - f_0\| = O(\rho^{\frac{1}{3}} \delta^{\frac{2}{3}}).$$

By choosing more alternative regularization functional μ_λ and choosing β large enough, one can show that it is possible to reach rate $\|f_\lambda - f_0\| = O(\delta^{1-\kappa})$ for an arbitrary small $\kappa > 0$, but one cannot reach a linear rate $\|f_\lambda - f_0\| = O(\|w\|)$. Such rates are achievable using non-linear sparse ℓ^1 regularizations as detailed in Chapter 11.

10.3 Quadratic Regularization

After this theoretical study in infinite dimension, we now turn our attention to more practical matters, and focus only on the finite dimensional setting.

Convex regularization. Following (10.9), the ill-posed problem of recovering an approximation of the high resolution image $f_0 \in \mathbb{R}^N$ from noisy measures $y = \Phi f_0 + w \in \mathbb{R}^P$ is regularized by solving a convex optimization problem

$$f_\lambda \in \operatorname{argmin}_{f \in \mathbb{R}^N} \mathcal{E}(f) \stackrel{\text{def.}}{=} \frac{1}{2} \|y - \Phi f\|^2 + \lambda J(f) \quad (10.17)$$

where $\|y - \Phi f\|^2$ is the data fitting term (here $\|\cdot\|$ is the ℓ^2 norm on \mathbb{R}^P) and $J(f)$ is a convex functional on \mathbb{R}^N .

The Lagrange multiplier λ weights the importance of these two terms, and is in practice difficult to set. Simulation can be performed on high resolution signal f_0 to calibrate the multiplier by minimizing the super-resolution error $\|f_0 - \tilde{f}\|$, but this is usually difficult to do on real life problems.

In the case where there is no noise, $w = 0$, the Lagrange multiplier λ should be set as small as possible. In the limit where $\lambda \rightarrow 0$, the unconstrained optimization problem (10.17) becomes a constrained optimization, as the following proposition explains. Let us stress that, without loss of generality, we can assume that $y \in \operatorname{Im}(\Phi)$, because one has the orthogonal decomposition

$$\|y - \Phi f\|^2 = \|y - \operatorname{Proj}_{\operatorname{Im}(\Phi)}(y)\|^2 + \|\operatorname{Proj}_{\operatorname{Im}(\Phi)}(y) - \Phi f\|^2$$

so that one can replace y by $\operatorname{Proj}_{\operatorname{Im}(\Phi)}(y)$ in (10.17).

Let us recall that a function J is coercive if

$$\lim_{\|f\| \rightarrow +\infty} f = +\infty$$

i.e.

$$\forall K, \exists R, \|x\| \geq R \implies |J(x)| \geq K.$$

This means that its non-empty levelsets $\{f ; J(f) \leq c\}$ are bounded (and hence compact) for all c .

Proposition 23. *We assume that J is coercive and that $y \in \Phi$. Then, if for each λ , f_λ is a solution of (10.17), then $(f_\lambda)_\lambda$ is a bounded set and any accumulation point f^* is a solution of*

$$f^* = \operatorname{argmin}_{f \in \mathbb{R}^N} \{J(f) ; \Phi f = y\}. \quad (10.18)$$

Proof. Denoting h , any solution to (10.18), which in particular satisfies $\Phi h = y$, because of the optimality condition of f_λ for (10.17), one has

$$\frac{1}{2\lambda} \|\Phi f_\lambda - y\|^2 + J(f_\lambda) \leq \frac{1}{2\lambda} \|\Phi h - y\|^2 + J(h) = J(h).$$

This shows that $J(f_\lambda) \leq J(h)$ so that since J is coercive the set $(f_\lambda)_\lambda$ is bounded and thus one can consider an accumulation point $f_{\lambda_k} \rightarrow f^*$ for $k \rightarrow +\infty$. Since $\|\Phi f_{\lambda_k} - y\|^2 \leq \lambda_k J(h)$, one has in the limit $\Phi f^* = y$, so that f^* satisfies the constraints in (10.18). Furthermore, by continuity of J , passing to the limit in $J(f_{\lambda_k}) \leq J(h)$, one obtains $J(f^*) \leq J(h)$ so that f^* is a solution of (10.18). \square

Note that it is possible to extend this proposition in the case where J is not necessarily coercive on the full space (for instance the TV functionals in Section 10.4.1 bellow) but on the orthogonal to $\ker(\Phi)$. The proof is more difficult.

Quadratic Regularization. The simplest class of prior functional are quadratic, and can be written as

$$J(f) = \frac{1}{2} \|Gf\|_{\mathbb{R}^K}^2 = \frac{1}{2} \langle Lf, f \rangle_{\mathbb{R}^N} \quad (10.19)$$

where $G \in \mathbb{R}^{K \times N}$ and where $L = G^*G \in \mathbb{R}^{N \times N}$ is a positive semi-definite matrix. The special case (10.10) is recovered when setting $G = L = \text{Id}_N$.

Writing down the first order optimality conditions for (10.17) leads to

$$\nabla \mathcal{E}(f) = \Phi^*(\Phi f - y) + \lambda L f = 0,$$

hence, if

$$\ker(\Phi) \cap \ker(G) = \{0\},$$

then (10.19) has a unique minimizer f_λ , which is obtained by solving a linear system

$$f_\lambda = (\Phi^*\Phi + \lambda L)^{-1}\Phi^*y. \quad (10.20)$$

In the special case where L is diagonalized by the singular basis $(v_m)_m$ of Φ , i.e. $L = V \operatorname{diag}(\alpha_m^2) V^*$, then f_λ reads in this basis

$$\langle f_\lambda, v_m \rangle = \frac{\sigma_m}{\sigma_m^2 + \lambda \alpha_m^2} \langle y, v_m \rangle. \quad (10.21)$$

Example of convolution. A specific example is for convolution operator

$$\Phi f = h \star f, \quad (10.22)$$

and using $G = \nabla$ be a discretization of the gradient operator, such as for instance using first order finite differences (2.16). This corresponds to the discrete Sobolev prior introduced in Section 9.1.2. Such an operator compute, for a d -dimensional signal $f \in \mathbb{R}^N$ (for instance a 1-D signal for $d = 1$ or an image when $d = 2$), an approximation $\nabla f_n \in \mathbb{R}^d$ of the gradient vector at each sample location n . Thus typically, $\nabla : f \mapsto (\nabla f_n)_n \in \mathbb{R}^{N \times d}$ maps to d -dimensional vector fields. Then $-\nabla^* : \mathbb{R}^{N \times d} \rightarrow \mathbb{R}^N$ is a discretized divergence operator. In this case, $\Delta = -GG^*$ is a discretization of the Laplacian, which is itself a convolution operator. One then has

$$\hat{f}_{\lambda,m} = \frac{\hat{h}_m^* \hat{y}_m}{|\hat{h}_m|^2 - \lambda \hat{d}_{2,m}}, \quad (10.23)$$

where \hat{d}_2 is the Fourier transform of the filter d_2 corresponding to the Laplacian. For instance, in dimension 1, using first order finite differences, the expression for $\hat{d}_{2,m}$ is given in (2.18).

10.3.1 Solving Linear System

When Φ and L do not share the same singular spaces, using (10.21) is not possible, so that one needs to solve the linear system (10.20), which can be rewritten as

$$Af = b \quad \text{where} \quad A \stackrel{\text{def.}}{=} \Phi^*\Phi + \lambda L \quad \text{and} \quad b = \Phi^*y.$$

It is possible to solve exactly this linear system with direct methods for moderate N (up to a few thousands), and the numerical complexity for a generic A is $O(N^3)$. Since the involved matrix A is symmetric, the best option is to use Choleski factorization $A = BB^*$ where B is lower-triangular. In favorable cases, this factorization (possibly with some re-re-ordering of the row and columns) can take advantage of some sparsity in A .

For large N , such exact resolution is not an option, and should use approximate iterative solvers, which only access A through matrix-vector multiplication. This is especially advantageous for imaging applications, where such multiplications are in general much faster than a naive $O(N^2)$ explicit computation. If the matrix A is highly sparse, this typically necessitates $O(N)$ operations. In the case where A is symmetric and positive definite (which is the case here), the most well known method is the conjugate gradient methods, which is actually an optimization method solving

$$\min_{f \in \mathbb{R}^N} \mathcal{E}(f) \stackrel{\text{def.}}{=} \mathcal{Q}(f) \stackrel{\text{def.}}{=} \langle Af, f \rangle - \langle f, b \rangle \quad (10.24)$$

which is equivalent to the initial minimization (10.17). Instead of doing a naive gradient descent (as studied in Section 13.1 below), starting from an arbitrary $f^{(0)}$, it compute a new iterate $f^{(\ell+1)}$ from the previous iterates as

$$f^{(\ell+1)} \stackrel{\text{def.}}{=} \operatorname{argmin}_f \left\{ \mathcal{E}(f) ; f \in f^{(\ell)} + \operatorname{Span}(\nabla \mathcal{E}(f^{(0)}), \dots, \nabla \mathcal{E}(f^{(\ell)})) \right\}.$$

The crucial and remarkable fact is that this minimization can be computed in closed form at the cost of two matrix-vector product per iteration, for $k \geq 1$ (posing initially $d^{(0)} = \nabla \mathcal{E}(f^{(0)}) = Af^{(0)} - b$)

$$f^{(\ell+1)} = f^{(\ell)} - \tau_\ell d^{(\ell)} \quad \text{where} \quad d^{(\ell)} = g_\ell + \frac{\|g^{(\ell)}\|^2}{\|g^{(\ell-1)}\|^2} d^{(\ell-1)} \quad \text{and} \quad \tau_\ell = \frac{\langle g_\ell, d^{(\ell)} \rangle}{\langle Ad^{(\ell)}, d^{(\ell)} \rangle} \quad (10.25)$$

$g^{(\ell)} \stackrel{\text{def.}}{=} \nabla \mathcal{E}(f^{(\ell)}) = Af^{(\ell)} - b$. It can also be shown that the direction $d^{(\ell)}$ are orthogonal, so that after $\ell = N$ iterations, the conjugate gradient computes the unique solution $f^{(\ell)}$ of the linear system $Af = b$. It is however rarely used this way (as an exact solver), and in practice much less than N iterates are computed. It should also be noted that iterations (10.25) can be carried over for an arbitrary smooth convex function \mathcal{E} , and it typically improves over the gradient descent (although in practice quasi-Newton method are often preferred).

10.4 Non-Quadratic Regularization

10.4.1 Total Variation Regularization

A major issue with quadratic regularization such as (10.19) is that they typically leads to blurry recovered data f_λ , which is thus not a good approximation of f_0 when it contains sharp transition such as edges in images. This is can clearly be seen in the convolutive case (10.23), this the restoration operator $\Phi_\lambda^+\Phi$ is a filtering, which tends to smooth sharp part of the data.

This phenomena can also be understood because the restored data f_λ always belongs to $\operatorname{Im}(\Phi^*) = \ker(\Phi)^\perp$, and thus cannot contains “high frequency” details that are lost in the kernel of Φ . To alleviate this shortcoming, and recover missing information in the kernel, it is thus necessarily to consider non-quadratic and in fact non-smooth regularization.

Total variation. The most well known instance of such a non-quadratic and non-smooth regularization is the total variation prior. For smooth function $f : \mathbb{R}^d \mapsto \mathbb{R}$, this amounts to replacing the quadratic Sobolev energy (often called Dirichlet energy)

$$J_{\text{Sob}}(f) \stackrel{\text{def.}}{=} \frac{1}{2} \int_{\mathbb{R}^d} \|\nabla f\|_{\mathbb{R}^d}^2 dx,$$

where $\nabla f(x) = (\partial_{x_1} f(x), \dots, \partial_{x_d} f(x))^{\top}$ is the gradient, by the (vectorial) L^1 norm of the gradient

$$J_{\text{TV}}(f) \stackrel{\text{def.}}{=} \int_{\mathbb{R}^d} \|\nabla f\|_{\mathbb{R}^d} dx.$$

We refer also to Section 9.1.1 about these priors. Simply “removing” the square 2 inside the integral might seem like a small change, but in fact it is a game changer.

Indeed, while $J_{\text{Sob}}(1_{\Omega}) = +\infty$ where 1_{Ω} is the indicator of a set Ω with finite perimeter $|\Omega| < +\infty$, one can show that $J_{\text{TV}}(1_{\Omega}) = |\Omega|$, if one interpret ∇f as a distribution Df (actually a vectorial Radon measure) and $\int_{\mathbb{R}^d} \|\nabla f\|_{\mathbb{R}^d} dx$ is replaced by the total mass $|Df|(\Omega)$ of this distribution $m = Df$

$$|m|(\Omega) = \sup \left\{ \int_{\mathbb{R}^d} \langle h(x), dm(x) \rangle ; h \in \mathcal{C}(\mathbb{R}^d \mapsto \mathbb{R}^d), \forall x, \|h(x)\| \leq 1 \right\}.$$

The total variation of a function such that Df has a bounded total mass (a so-called bounded variation function) is hence defined as

$$J_{\text{TV}}(f) \stackrel{\text{def.}}{=} \sup \left\{ \int_{\mathbb{R}^d} f(x) \operatorname{div}(h)(x) dx ; h \in \mathcal{C}_c^1(\mathbb{R}^d; \mathbb{R}^d), \|h\|_{\infty} \leq 1 \right\}.$$

Generalizing the fact that $J_{\text{TV}}(1_{\Omega}) = |\Omega|$, the functional co-area formula reads

$$J_{\text{TV}}(f) = \int_{\mathbb{R}} \mathcal{H}_{d-1}(L_t(f)) dt \quad \text{where } L_t(f) = \{x ; f(x) = t\}$$

and where \mathcal{H}_{d-1} is the Hausdorff measures of dimension $d-1$, for instance, for $d=2$ if L has finite perimeter $|L|$, then $\mathcal{H}_{d-1}(L) = |L|$ is the perimeter of L .

Discretized Total variation. For discretized data $f \in \mathbb{R}^N$, one can define a discretized TV semi-norm as detailed in Section 9.1.2, and it reads, generalizing (9.6) to any dimension

$$J_{\text{TV}}(f) = \sum_n \|\nabla f_n\|_{\mathbb{R}^d}$$

where $\nabla f_n \in \mathbb{R}^d$ is a finite difference gradient at location indexed by n .

The discrete total variation prior $J_{\text{TV}}(f)$ defined in (9.6) is a convex but non differentiable function of f , since a term of the form $\|\nabla f_n\|$ is non differentiable if $\nabla f_n = 0$. We defer to chapters 12 and 13 the study of advanced non-smooth convex optimization technics that allows to handle this kind of functionals.

In order to be able to use simple gradient descent methods, one needs to smooth the TV functional. The general machinery proceeds by replacing the non-smooth ℓ^2 Euclidean norm $\|\cdot\|$ by a smoothed version, for instance

$$\forall u \in \mathbb{R}^d, \quad \|u\|_{\varepsilon} \stackrel{\text{def.}}{=} \sqrt{\varepsilon^2 + \|u\|^2}.$$

This leads to the definition of a smoothed approximate TV functional, already introduced in (9.12),

$$J_{\text{TV}}^{\varepsilon}(f) \stackrel{\text{def.}}{=} \sum_n \|\nabla f_n\|_{\varepsilon}$$

One has the following asymptotics for $\varepsilon \rightarrow \{0, +\infty\}$

$$\|u\|_\varepsilon \xrightarrow{\varepsilon \rightarrow 0} \|u\| \quad \text{and} \quad \|u\|_\varepsilon = \varepsilon + \frac{1}{2\varepsilon} \|u\|^2 + O(1/\varepsilon^2)$$

which suggest that $J_{\text{TV}}^\varepsilon$ interpolates between J_{TV} and J_{Sob} .

The resulting inverse regularization problem (10.17) thus reads

$$f_\lambda \stackrel{\text{def.}}{=} \underset{f \in \mathbb{R}^N}{\operatorname{argmin}} \mathcal{E}(f) = \frac{1}{2} \|y - \Phi f\|^2 + \lambda J_{\text{TV}}^\varepsilon(f) \quad (10.26)$$

It is a strictly convex problem (because $\|\cdot\|_\varepsilon$ is strictly convex for $\varepsilon > 0$) so that its solution f_λ is unique.

10.4.2 Gradient Descent Method

The optimization program (10.26) is a example of smooth unconstrained convex optimization of the form

$$\min_{f \in \mathbb{R}^N} \mathcal{E}(f) \quad (10.27)$$

where $\mathcal{E} : \mathbb{R}^N \rightarrow \mathbb{R}$ is a \mathcal{C}^1 function. Recall that the gradient $\nabla \mathcal{E} : \mathbb{R}^N \mapsto \mathbb{R}^N$ of this functional (not to be confound with the discretized gradient $\nabla f \in \mathbb{R}^N$ of f) is defined by the following first order relation

$$\mathcal{E}(f + r) = \mathcal{E}(f) + \langle f, r \rangle_{\mathbb{R}^N} + O(\|r\|_{\mathbb{R}^N}^2)$$

where we used $O(\|r\|_{\mathbb{R}^N}^2)$ in place of $o(\|r\|_{\mathbb{R}^N})$ (for differentiable function) because we assume here \mathcal{E} is of class \mathcal{C}^1 (i.e. the gradient is continuous).

For such a function, the gradient descent algorithm is defined as

$$f^{(\ell+1)} \stackrel{\text{def.}}{=} f^{(\ell)} - \tau_\ell \nabla \mathcal{E}(f^{(\ell)}), \quad (10.28)$$

where the step size $\tau_\ell > 0$ should be small enough to guarantee convergence, but large enough for this algorithm to be fast.

We refer to Section 13.1 for a detailed analysis of the convergence of the gradient descent, and a study of the influence of the step size τ_ℓ .

10.4.3 Examples of Gradient Computation

Note that the gradient of a quadratic function $\mathcal{Q}(f)$ of the form (10.24) reads

$$\nabla \mathcal{G}(f) = Af - b.$$

In particular, one retrieves that the first order optimality condition $\nabla \mathcal{G}(f) = 0$ is equivalent to the linear system $Af = b$.

For the quadratic fidelity term $\mathcal{G}(f) = \frac{1}{2} \|\Phi f - y\|^2$, one thus obtains

$$\nabla \mathcal{G}(f) = \Phi^*(\Phi y - y).$$

In the special case of the regularized TV problem (10.26), the gradient of \mathcal{E} reads

$$\nabla \mathcal{E}(f) = \Phi^*(\Phi y - y) + \lambda \nabla J_{\text{TV}}^\varepsilon(f).$$

Recall the chain rule for differential reads $\partial(\mathcal{G}_1 \circ \mathcal{G}_2) = \partial \mathcal{G}_1 \circ \partial \mathcal{G}_2$, but that gradient vectors are actually transposed of differentials, so that for $\mathcal{E} = \mathcal{F} \circ \mathcal{H}$ where $\mathcal{F} : \mathbb{R}^P \rightarrow \mathbb{R}$ and $\mathcal{H} : \mathbb{R}^N \rightarrow \mathbb{R}^P$, one has

$$\nabla \mathcal{E}(f) = [\partial \mathcal{H}(f)]^* (\nabla \mathcal{F}(\mathcal{H}f)).$$

Since $J_{\text{TV}}^\varepsilon = \|\cdot\|_{1,\varepsilon} \circ \nabla$, one thus has

$$\nabla J_{\text{TV}}^\varepsilon = \nabla^* \circ (\partial \|\cdot\|_{1,\varepsilon}) \quad \text{where} \quad \|u\|_{1,\varepsilon} = \sum_n \|u_n\|_\varepsilon$$

so that

$$J_{\text{TV}}^\varepsilon(f) = -\operatorname{div}(\mathcal{N}^\varepsilon(\nabla f)),$$

where $\mathcal{N}^\varepsilon(u) = (u_n/\|u_n\|_\varepsilon)_n$ is the smoothed-normalization operator of vector fields (the differential of $\|\cdot\|_{1,\varepsilon}$), and where $\operatorname{div} = -\nabla^*$ is minus the adjoint of the gradient.

Since $\operatorname{div} = -\nabla^*$, their Lipschitz constants are equal $\|\operatorname{div}\|_{\text{op}} = \|\nabla\|_{\text{op}}$, and is for instance equal to $\sqrt{2d}$ for the discretized gradient operator. Computation shows that the Hessian of $\|\cdot\|_\varepsilon$ is bounded by $1/\varepsilon$, so that for the smoothed-TV functional, the Lipschitz constant of the gradient is upper-bounded by

$$L = \frac{\|\nabla\|^2}{\varepsilon} + \|\Phi\|_{\text{op}}^2.$$

Furthermore, this functional is strongly convex because $\|\cdot\|_\varepsilon$ is ε -strongly convex, and the Hessian is lower bounded by

$$\mu = \varepsilon + \sigma_{\min}(\Phi)^2$$

where $\sigma_{\min}(\Phi)$ is the smallest singular value of Φ . For ill-posed problems, typically $\sigma_{\min}(\Phi) = 0$ or is very small, so that both L and μ degrades (tends respectively to 0 and $+\infty$) as $\varepsilon \rightarrow 0$, so that gradient descent becomes prohibitive for small ε , and it is thus required to use dedicated non-smooth optimization methods detailed in the following chapters. On the good news side, note however that in many case, using a small but non-zero value for ε often leads to better a visually more pleasing results, since it introduce a small blurring which diminishes the artifacts (and in particular the so-called “stair-casing” effect) of TV regularization.

10.5 Examples of Inverse Problems

We detail here some inverse problem in imaging that can be solved using quadratic regularization or non-linear TV.

10.5.1 Deconvolution

The blurring operator (10.1) is diagonal over Fourier, so that quadratic regularization are easily solved using Fast Fourier Transforms when considering periodic boundary conditions. We refer to (10.22) and the correspond explanations. TV regularization in contrast cannot be solved with fast Fourier technics, and is thus much slower.

10.5.2 Inpainting

For the inpainting problem, the operator defined in (10.3) is diagonal in space

$$\Phi = \operatorname{diag}_m(\delta_{\Omega^c}[m]),$$

and is an orthogonal projector $\Phi^* = \Phi$.

In the noiseless case, to constrain the solution to lie in the affine space $\{f \in \mathbb{R}^N ; y = \Phi f\}$, we use the orthogonal projector

$$\forall x, \quad P_y(f)(x) = \begin{cases} f(x) & \text{if } x \in \Omega, \\ y(x) & \text{if } x \notin \Omega. \end{cases}$$

In the noiseless case, the recovery (10.18) is solved using a projected gradient descent. For the Sobolev energy, the algorithm iterates

$$f^{(\ell+1)} = P_y(f^{(\ell)} + \tau \Delta f^{(\ell)}).$$

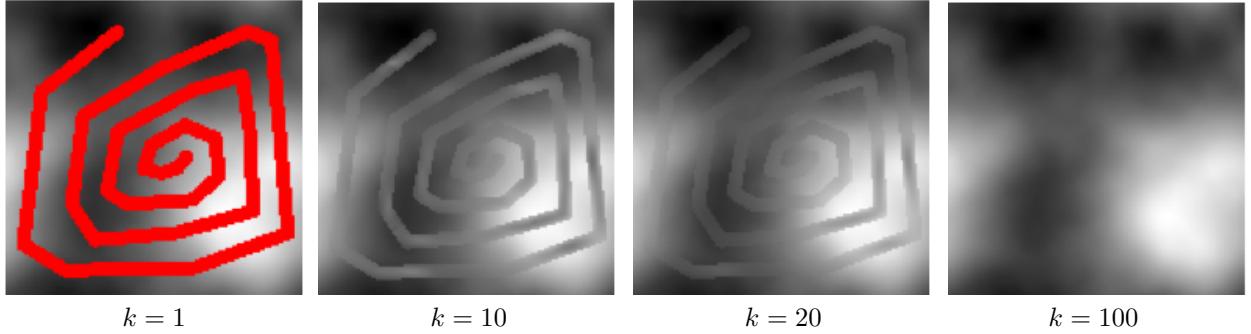


Figure 10.3: Sobolev projected gradient descent algorithm.

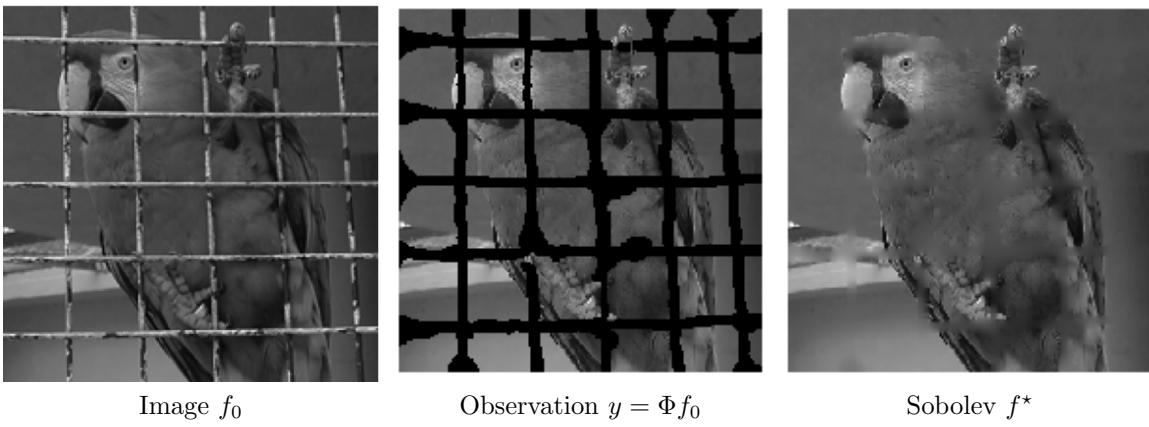


Figure 10.4: Inpainting the parrot cage.

which converges if $\tau < 2/\|\Delta\| = 1/4$. Figure 10.3 shows some iteration of this algorithm, which progressively interpolate within the missing area.

Figure 10.4 shows an example of Sobolev inpainting to achieve a special effect.

For the smoothed TV prior, the gradient descent reads

$$f^{(\ell+1)} = P_y \left(f^{(\ell)} + \tau \operatorname{div} \left(\frac{\nabla f^{(\ell)}}{\sqrt{\varepsilon^2 + \|\nabla f^{(\ell)}\|^2}} \right) \right)$$

which converges if $\tau < \varepsilon/4$.

Figure 10.5 compare the Sobolev inpainting and the TV inpainting for a small value of ε . The SNR is not improved by the total variation, but the result looks visually slightly better.

10.5.3 Tomography Inversion

In medical imaging, a scanner device compute projection of the human body along rays $\Delta_{t,\theta}$ defined

$$x \cdot \tau_\theta = x_1 \cos \theta + x_2 \sin \theta = t$$

where we restrict ourself to 2D projection to simplify the exposition.

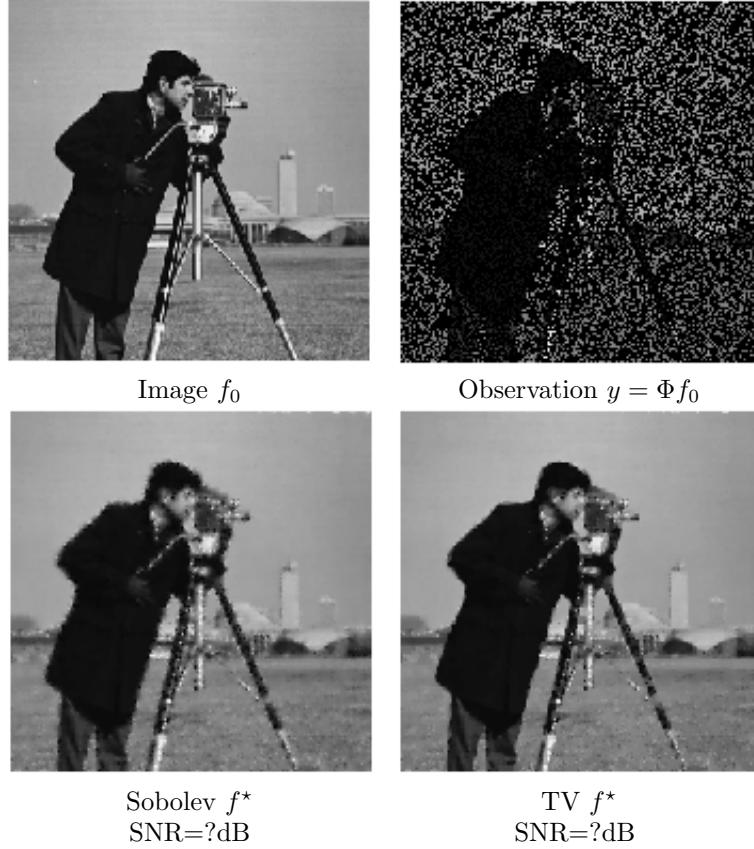


Figure 10.5: Inpainting with Sobolev and TV regularization.

The scanning process computes a Radon transform, which compute the integral of the function to acquires along rays

$$\forall \theta \in [0, \pi), \forall t \in \mathbb{R}, \quad p_\theta(t) = \int_{\Delta_{t,\theta}} f(x) ds = \iint f(x) \delta(x \cdot \tau_\theta - t) dx$$

see Figure (10.6)

The Fourier slice theorem relates the Fourier transform of the scanned data to the 1D Fourier transform of the data along rays

$$\forall \theta \in [0, \pi), \forall \xi \in \mathbb{R} \quad \hat{p}_\theta(\xi) = \hat{f}(\xi \cos \theta, \xi \sin \theta). \quad (10.29)$$

This shows that the pseudo inverse of the Radon transform is computed easily over the Fourier domain using inverse 2D Fourier transform

$$f(x) = \frac{1}{2\pi} \int_0^\pi p_\theta * h(x \cdot \tau_\theta) d\theta$$

with $\hat{h}(\xi) = |\xi|$.

Imaging devices only capture a limited number of equispaced rays at orientations $\{\theta_k = \pi/k\}_{0 \leq k < K}$. This defines a tomography operator which corresponds to a partial Radon transform

$$Rf = (p_{\theta_k})_{0 \leq k < K}.$$

Relation (10.29) shows that knowing Rf is equivalent to knowing the Fourier transform of f along rays,

$$\{\hat{f}(\xi \cos(\theta_k), \xi \sin(\theta_k))\}_k.$$

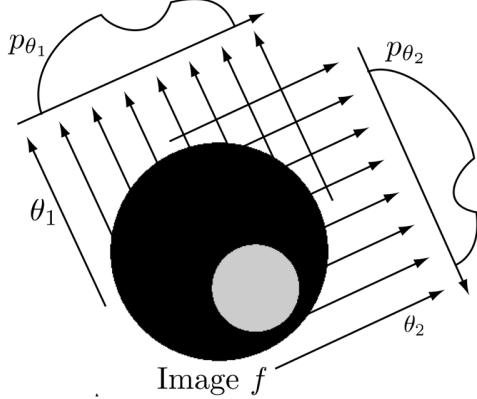


Figure 10.6: Principle of tomography acquisition.

We thus simplify the acquisition process over the discrete domain and model it as computing directly samples of the Fourier transform

$$\Phi f = (\hat{f}[\omega])_{\omega \in \Omega} \in \mathbb{R}^P$$

where Ω is a discrete set of radial lines in the Fourier plane, see Figure 10.7, right.

In this discrete setting, recovering from Tomography measures $y = Rf_0$ is equivalent in this setup to inpaint missing Fourier frequencies, and we consider partial noisy Fourier measures

$$\forall \omega \in \Omega, \quad y[\omega] = \hat{f}[\omega] + w[\omega]$$

where $w[\omega]$ is some measurement noise, assumed here to be Gaussian white noise for simplicity.

The pseudo-inverse $f^+ = R^+y$ defined in (10.7) of this partial Fourier measurements reads

$$\hat{f}^+[\omega] = \begin{cases} y[\omega] & \text{if } \omega \in \Omega, \\ 0 & \text{if } \omega \notin \Omega. \end{cases}$$

Figure 10.8 shows examples of pseudo inverse reconstruction for increasing size of Ω . This reconstruction exhibit serious artifact because of bad handling of Fourier frequencies (zero padding of missing frequencies).

The total variation regularization (??) reads

$$f^* \in \operatorname{argmin}_f \frac{1}{2} \sum_{\omega \in \Omega} |y[\omega] - \hat{f}[\omega]|^2 + \lambda \|f\|_{\text{TV}}.$$

It is especially suitable for medical imaging where organ of the body are of relatively constant gray value, thus resembling to the cartoon image model introduced in Section 6.2.4. Figure 10.9 compares this total variation recovery to the pseudo-inverse for a synthetic cartoon image. This shows the hability of the total variation to recover sharp features when inpainting Fourier measures. This should be contrasted with the difficulties that faces TV regularization to inpaint over the spacial domain, as shown in Figure 11.10.

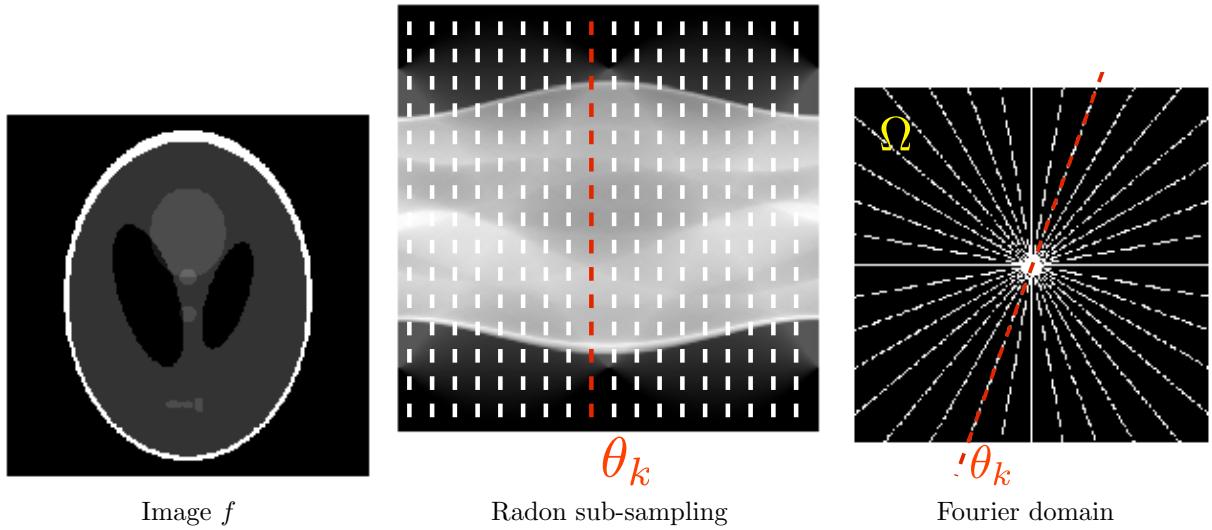


Figure 10.7: Partial Fourier measures.

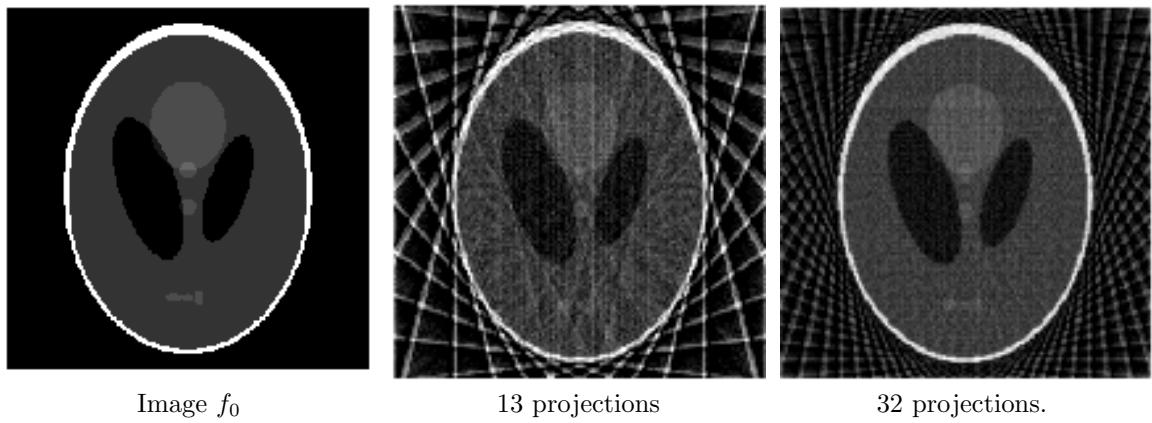


Figure 10.8: Pseudo inverse reconstruction from partial Radon projections.

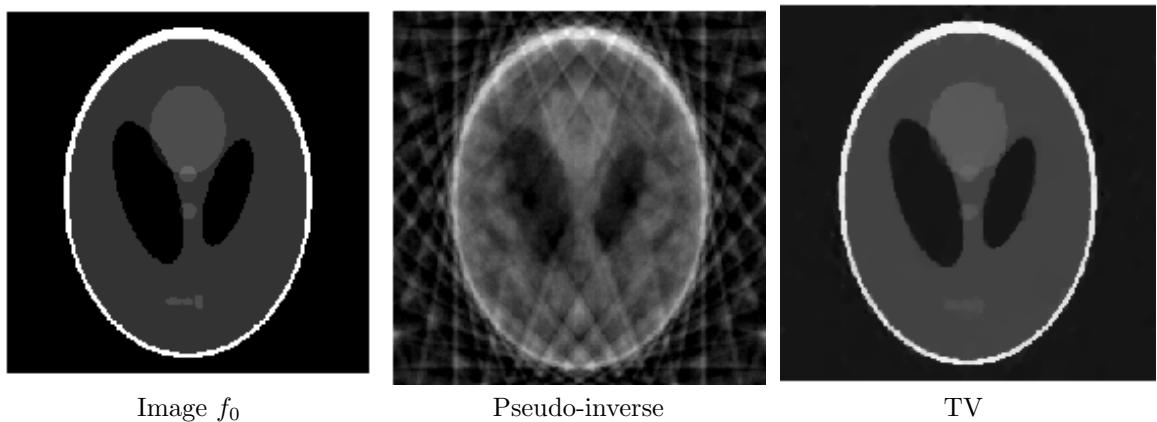


Figure 10.9: Total variation tomography inversion.

Chapter 11

Sparse Regularization

Ref [28, 40, 35]

11.1 Sparsity Priors

11.1.1 Ideal sparsity prior.

As detailed in Chapter ??, it is possible to use an orthogonal basis $\mathcal{B} = \{\psi_m\}_m$ to efficiently approximate an image f in a given class $f \in \Theta$ with a few atoms from \mathcal{B} .

To measure the complexity of an approximation with \mathcal{B} , we consider the ℓ^0 prior, which counts the number of non-zero coefficients in \mathcal{B}

$$J_0(f) \stackrel{\text{def.}}{=} \#\{m ; \langle f, \psi_m \rangle \neq 0\} \quad \text{where } x_m = \langle f, \psi_m \rangle.$$

One often also denote it as the ℓ^0 “pseudo-norm”

$$\|x\|_0 \stackrel{\text{def.}}{=} J_0(f).$$

which we treat here as an ideal sparsity measure for the coefficients x of f in \mathcal{B} .

Natural images are not exactly composed of a few atoms, but they can be well approximated by a function f_M with a small ideal sparsity $M = J_0(f)$. In particular, the best M -term approximation defined in (6.3) is defined by

$$f_M = \sum_{|\langle f, \psi_m \rangle| > T} \langle f, \psi_m \rangle \psi_m \quad \text{where } M = \#\{m ; |\langle f, \psi_m \rangle| > T\}.$$

As detailed in Section 6.2, discontinuous images with bounded variation have a fast decay of the approximation error $\|f - f_M\|$. Natural images f are well approximated by images with a small value of the ideal sparsity prior J_0 .

Figure 11.1 shows an examples of decomposition of a natural image in a wavelet basis, $\psi_m = \psi_{j,n}^\omega$, $m = (j, n, \omega)$. This shows that most $\langle f, \psi_m \rangle$ are small, and hence the decomposition is quite sparse.

11.1.2 Convex relaxation

Unfortunately, the ideal sparsity prior J_0 is difficult to handle numerically because $J_0(f)$ is not a convex function of f . For instance, if f and g have non-intersecting supports of there coefficients in \mathcal{B} , then $J_0((f + g)/2) = J_0(f) + J_0(g)$, which shows the highly non-convex behavior of J_0 .

This ideal sparsity J_0 is thus not amenable to minimization, which is an issue to solve general inverse problems considered in Section ??.

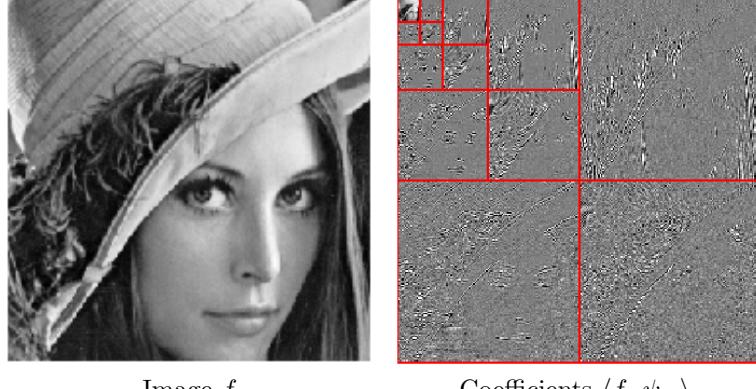


Figure 11.1: Wavelet coefficients of natural images are relatively sparse.

We consider a family of ℓ^q priors for $q > 0$, intended to approximate the ideal prior J_0

$$J_q(f) = \sum_m |\langle f, \psi_m \rangle|^q.$$

As shown in Figure 11.2, the unit balls in \mathbb{R}^2 associated to these priors are shrinking toward the axes, which corresponds to the unit ball for the ℓ^0 pseudo norm. In some sense, the J_q priors are becoming closer to J_0 as q tends to zero, and thus J_q favors sparsity for small q .

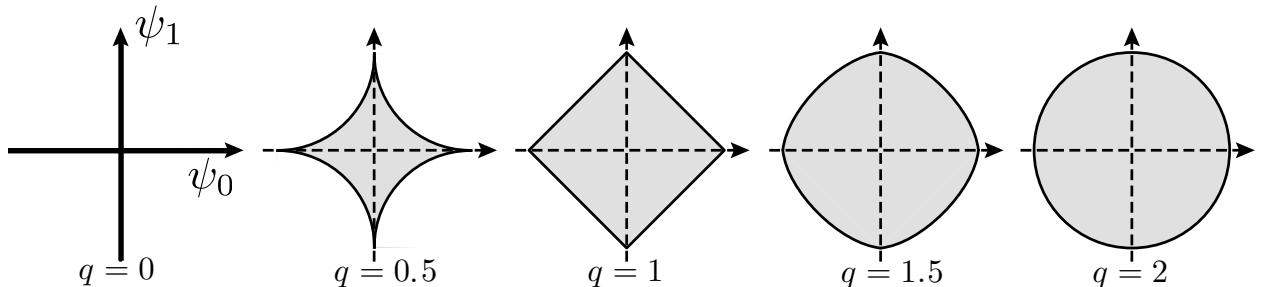


Figure 11.2: ℓ^q balls $\{x ; J_q(x) \leq 1\}$ for varying q .

The prior J_q is convex if and only if $q \geq 1$. To reach the highest degree of sparsity while using a convex prior, we consider the ℓ^1 sparsity prior J_1 , which is thus defined as

$$J_1(f) = \|(\langle f, \psi_m \rangle)\|_1 = \sum_m |\langle f, \psi_m \rangle|. \quad (11.1)$$

In the following, we consider discrete orthogonal bases $\mathcal{B} = \{\psi_m\}_{m=0}^{N-1}$ of \mathbb{R}^N .

11.1.3 Sparse Regularization and Thresholding

Given some orthogonal basis $\{\psi_m\}_m$ of \mathbb{R}^N , the denoising by regularization (9.15) is written using the sparsity J_0 and J_1 as

$$f^* = \operatorname{argmin}_{g \in \mathbb{R}^N} \frac{1}{2} \|f - g\|^2 + \lambda J_q(f)$$

for $q = 0$ or $q = 1$. It can be re-written in the orthogonal basis as

$$f^* = \sum_m x_m^* \psi_m$$

$$\text{where } x_m^* = \operatorname{argmin}_{y \in \mathbb{R}^N} \sum_m \frac{1}{2} |x_m - y_m|^2 + \lambda |y_m|^q$$

where $x_m \stackrel{\text{def}}{=} \langle f, \psi_m \rangle$, $y_m \stackrel{\text{def}}{=} \langle g, \psi_m \rangle$, and where we use the following slight abuse of notation for $q = 0$

$$\forall u \in \mathbb{R}, \quad |u|^0 = \begin{cases} 0 & \text{if } u = 0, \\ 1 & \text{otherwise.} \end{cases}$$

Each coefficients of the denoised image is the solution of a 1-D optimization problem

$$x_m^* = \operatorname{argmin}_{u \in \mathbb{R}} \frac{1}{2} |x_m - u|^2 + \lambda |u|^q \quad (11.2)$$

and the following proposition this optimization is solved exactly in closed form using thresholding.

Proposition 24. *One has*

$$x_m^* = S_T^q(x_m) \quad \text{where} \quad \begin{cases} T = \sqrt{2\lambda} & \text{for } q = 0, \\ T = \lambda & \text{for } q = 1, \end{cases} \quad (11.3)$$

where

$$\forall u \in \mathbb{R}, \quad S_T^0(u) \stackrel{\text{def}}{=} \begin{cases} 0 & \text{if } |u| < T, \\ u & \text{otherwise} \end{cases} \quad (11.4)$$

is the hard thresholding introduced in (8.6), and

$$\forall u \in \mathbb{R}, \quad S_T^1(u) \stackrel{\text{def}}{=} \operatorname{sign}(u)(|u| - T)_+ \quad (11.5)$$

is the soft thresholding introduced in (8.7).

Proof. One needs to solve (11.2). Figure 11.3, left shows the function $\|x - y\|^2 + T^2 \|x\|_0$, and the minimum is clearly at $x = 0$ when $T \geq y$, and at $x = y$ otherwise. This is thus a hard thresholding with threshold $T^2 = 2\lambda$. Figure 11.3, right, shows the evolution with λ of the function $\frac{1}{2}\|x - y\|^2 + \lambda|x|$. For $x > 0$, one has $F'(x) = x - y + \lambda$ which is 0 at $x = y - \lambda$. The minimum is at $x = y - \lambda$ for $\lambda \leq y$, and stays at 0 for all $\lambda > y$. \square

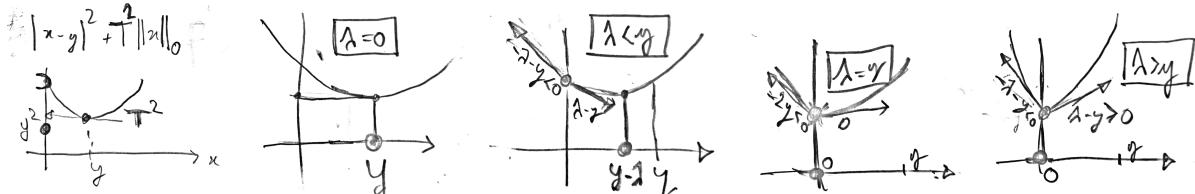


Figure 11.3: Leftmost: function $\|\cdot - y\|^2 + T^2 \|\cdot\|_0$. Others: evolution with λ of the function $F(x) \stackrel{\text{def}}{=} \frac{1}{2} \|\cdot - y\|^2 + \lambda |\cdot|$.

One thus has

$$f_{\lambda,q} = \sum_m S_T^q(\langle f, \psi_m \rangle) \psi_m.$$

As detailed in Section 8.3, these denoising methods has the advantage that the threshold is simple to set for Gaussian white noise w of variance σ^2 . Theoretical values indicated that $T = \sqrt{2 \log(N)} \sigma$ is asymptotically optimal, see Section 8.3.3. In practice, one should choose $T \approx 3\sigma$ for hard thresholding (ℓ^0 regularization), and $T \approx 3\sigma/2$ for soft thresholding (ℓ^1 regularization), see Figure 8.14.

11.2 Sparse Regularization of Inverse Problems

Sparse ℓ^1 regularization in an orthogonal basis $\{\psi_m\}_m$ of \mathbb{R}^N makes use of the J_1 prior defined in (11.1), so that the inversion is obtained by solving the following convex program

$$f_\lambda \in \operatorname{argmin}_{f \in \mathbb{R}^N} \frac{1}{2} \|y - \Phi f\|^2 + \lambda \sum_m |\langle f, \psi_m \rangle|. \quad (11.6)$$

This corresponds to the basis pursuit denoising for sparse approximation introduced by Chen, Donoho and Saunders in [11]. The resolution of (11.6) can be performed using an iterative thresholding algorithm as detailed in Section 11.3.

Analysis vs. synthesis priors. When the set of atoms $\Psi = \{\psi_m\}_{m=1}^Q$ is non-orthogonal (and might even be redundant in the case $Q > N$), there are two distinct ways to generalize problem (11.6), which we formulate as in (??), by introducing a generic convex prior J

$$f_\lambda \in \operatorname{argmin}_{f \in \mathbb{R}^N} \frac{1}{2} \|y - \Phi f\|^2 + \lambda J(f). \quad (11.7)$$

In the following, with a slight abuse of notation, we denote the ‘‘analysis’’ and ‘‘synthesis’’ operator as

$$\Psi : x \in \mathbb{R}^Q \mapsto \Psi x = \sum_m x_m \psi_m \quad \text{and} \quad \Psi^* : f \in \mathbb{R}^N \mapsto (\langle f, \psi_m \rangle)_{m=1}^Q \in \mathbb{R}^Q.$$

The so-called analysis-type prior is simply measuring the sparsity of the correlations with the atoms in the dictionary

$$J_1^A(f) \stackrel{\text{def.}}{=} \sum_m |\langle f, \psi_m \rangle| = \|\Psi^* f\|_1. \quad (11.8)$$

The so-called synthesis-type prior is contrast measure the sparsity of the sparsest expansion of f in Ψ , i.e.

$$J_1^S(f) \stackrel{\text{def.}}{=} \min_{x \in \mathbb{R}^q, \Psi x = f} \|x\|_1. \quad (11.9)$$

While the analysis regularization (11.8) seems simpler to handle, it is actually the contrary. Solving (11.7) with $J = J_1^A$ is in fact quite involved, and necessitate typically primal-dual algorithm as detailed in Chapter 13. Furthermore, the theoretical study of the performance of the resulting regularization method is mostly an open problem.

We thus now focus on the synthesis regularization problem $J = J_1^S$, and we re-write (11.7) conveniently as $f_\lambda = \Psi x_\lambda$ where x_λ is any solution of the following Basis Pursuit Denoising problem

$$x_\lambda \in \operatorname{argmin}_{x \in \mathbb{R}^Q} \frac{1}{2\lambda} \|y - Ax\|^2 + \|x\|_1 \quad (11.10)$$

where we introduced the following matrix

$$A \stackrel{\text{def.}}{=} \Phi \Psi \in \mathbb{R}^{P \times Q}.$$

As $\lambda \rightarrow 0$, we consider the following limit constrained problem

$$x^* = \operatorname{argmin}_{Ax=y} \|x\|_1 \quad (11.11)$$

and the signal is recovered as $f^* = \Psi x^* \in \mathbb{R}^N$.

11.3 Iterative Soft Thresholding Algorithm

This section details an iterative algorithm that computes a solution of (11.10).

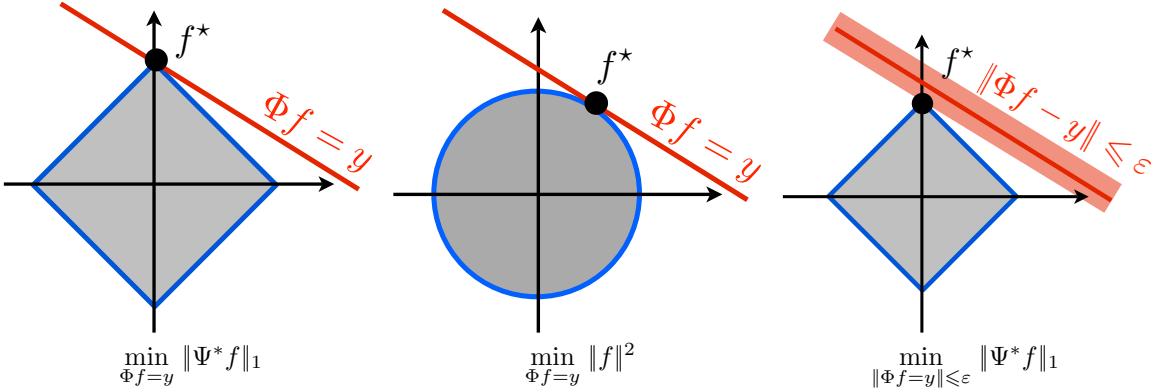


Figure 11.4: Geometry of convex optimizations.

11.3.1 Noiseless Recovery as a Linear Program

Before detailing this methods, which only deal with the case $\lambda > 0$, let us note that in the noiseless setting, $\lambda = 0$ and (11.11) is actually equivalent to a linear program. Indeed, decomposing $a = x_+ - x_-$ with $(x_+, x_-) \in (\mathbb{R}_+^Q)^2$, one has

$$x^* = \operatorname{argmin}_{(x_+, x_-) \in (\mathbb{R}_+^Q)^2} \{ \langle x_+, \mathbf{1}_Q \rangle + \langle x_-, \mathbf{1}_Q \rangle ; y = A(x_+ - x_-) \}. \quad (11.12)$$

which is a linear program. For small to medium scale problem (Q of the order of a few thousands) it can be solved using the simplex algorithm or interior point methods. For large scale problems such as those encountered in imaging or machine learning, this is not possible, and one has to resort to simpler first order schemes. A possible option is the Douglas-Rachford splitting scheme, which is detailed in Section ???. Let us however stress that the constrained problem (11.11), because of its polyhedral (linear) nature, is in fact harder to solve than the penalized problem (11.10) that we now target.

11.3.2 Projected Gradient Descent for ℓ^1 .

As a first practical example to solve (11.10), we will show how to use the projected gradient descent method, which is analyzed in detail in Section 13.1.3. Similarly to (11.12), we remap (11.10) as the resolution of a constraint minimization problem of the form (13.17) where here \mathcal{C} is a positivity constraint and

$$u = (u_+, u_-) \in (\mathbb{R}^Q)^2, \quad \mathcal{C} = (\mathbb{R}_+^Q)^2, \quad \text{and} \quad \mathcal{E}(u) = \frac{1}{2} \|\Phi(u_+ - u_-) - y\|^2 + \lambda \langle u_+, \mathbf{1}_Q \rangle + \lambda \langle u_-, \mathbf{1}_Q \rangle.$$

The projection on \mathcal{C} is here simple to compute

$$\operatorname{Proj}_{(\mathbb{R}_+^Q)^2}(u_+, u_-) = ((u_+)_{\oplus}, (u_-)_{\oplus}) \quad \text{where} \quad (r)_{\oplus} \stackrel{\text{def.}}{=} \max(r, 0),$$

and the gradient reads

$$\nabla \mathcal{E}(u_+, u_-) = (\eta + \lambda \mathbf{1}_Q, -\eta + \lambda \mathbf{1}_Q) \quad \text{where} \quad \eta = \Phi^*(\Phi(u_+ - u_-) - y)$$

Denoting $u^{(\ell)} = (u_+^{(\ell)}, u_-^{(\ell)})$ and $x^{(\ell)} \stackrel{\text{def.}}{=} u_+^{(\ell)} - u_-^{(\ell)}$, the iterate of the projected gradient descent algorithm (13.18) read

$$u_+^{(\ell+1)} \stackrel{\text{def.}}{=} \left(u_+^{(\ell)} - \tau_\ell(\eta^{(\ell)} + \lambda) \right)_{\oplus} \quad \text{and} \quad u_-^{(\ell+1)} \stackrel{\text{def.}}{=} \left(u_-^{(\ell)} - \tau_\ell(-\eta^{(\ell)} + \lambda) \right)_{\oplus}$$

where $\eta^{(\ell)} \stackrel{\text{def.}}{=} \Phi^*(\Phi x^{(\ell)} - y)$.

Theorem 37 ensures that $u^{(\ell)} \rightarrow u^*$ a solution to (13.17) if

$$\forall \ell, \quad 0 < \tau_{\min} < \tau_\ell < \tau_{\max} < \frac{2}{\|\Phi\|^2},$$

and thus $x^{(\ell)} \rightarrow x^* = u_+^* - u_-^*$ which is thus a solution to (11.10).

11.3.3 Iterative Soft Thresholding and Forward Backward

A drawback of this projected gradient descent scheme is that it necessitate to store $2Q$ coefficients. A closely related method, which comes with exactly the same convergence guarantees and rate, is the so called “iterative soft thresholding algorithm” (ISTA). This algorithm was derived by several authors, among which [20, 16], and belongs to the general family of forward-backward splitting in proximal iterations [14], which we detail in Section 13.3.2.

For the sake of simplicity, let us derive this algorithm in the special case of ℓ^1 by surrogate function minimization. We aim at minimizing (11.6)

$$\mathcal{E}(x) \stackrel{\text{def.}}{=} \frac{1}{2} \|y - Ax\|^2 + \lambda \|x\|_1$$

and we introduce for any fixed x' the function

$$\mathcal{E}_\tau(x, x') \stackrel{\text{def.}}{=} \mathcal{E}(x) - \frac{1}{2} \|Ax - Ax'\|^2 + \frac{1}{2\tau} \|x - x'\|.$$

We notice that $\mathcal{E}(x, x) = 0$ and one has

$$K(x, x') \stackrel{\text{def.}}{=} -\frac{1}{2} \|Ax - Ax'\|^2 + \frac{1}{2\tau} \|x - x'\| = \frac{1}{2} \left\langle \left(\frac{1}{\tau} \text{Id}_N - A^* A \right) (x - x'), x - x' \right\rangle.$$

This quantity $K(x, x')$ is positive if $\lambda_{\max}(A^* A) \leqslant 1/\tau$ (maximum eigenvalue), i.e. $\tau \leqslant 1/\|A\|_{\text{op}}^2$, where we recall that $\|A\|_{\text{op}} = \sigma_{\max}(A)$ is the operator (algebra) norm. This shows that $\mathcal{E}_\tau(x, x')$ is a valid surrogate functional, in the sense that

$$\mathcal{E}(x) \leqslant \mathcal{E}_\tau(x, x'), \quad \mathcal{E}_\tau(x, x') = 0, \quad \text{and} \quad \mathcal{E}(\cdot) - \mathcal{E}_\tau(\cdot, x') \text{ is smooth.}$$

This leads to define

$$x^{(\ell+1)} \stackrel{\text{def.}}{=} \underset{x}{\operatorname{argmin}} \mathcal{E}_{\tau_\ell}(x, x^{(\ell)}) \tag{11.13}$$

which by construction satisfies

$$\mathcal{E}(x^{(\ell+1)}) \leqslant \mathcal{E}(x^{(\ell)}).$$

Proposition 25. *The iterates $x^{(\ell)}$ defined by (11.13) satisfy*

$$x^{(\ell+1)} = \mathcal{S}_{\lambda\tau_\ell}^1 \left(x^{(\ell)} - \tau_\ell A^*(Ax^{(\ell)} - y) \right) \tag{11.14}$$

where $\mathcal{S}_\lambda^1(x) = (S_\lambda^1(x_m))_m$ where $S_\lambda^1(r) = \text{sign}(r)(|r| - \lambda)_+$ is the soft thresholding operator defined in (11.5).

Proof. One has

$$\begin{aligned} \mathcal{E}_\tau(x, x') &= \frac{1}{2} \|Ax - y\|^2 - \frac{1}{2} \|Ax - Ax'\|^2 + \frac{1}{2\tau} \|x - x'\| + \lambda \|x\|_1 \\ &= C + \frac{1}{2} \|Ax\|^2 - \frac{1}{2} \|Ax\|^2 + \frac{1}{2\tau} \|x\|^2 - \langle Ax, y \rangle + \langle Ax, Ax' \rangle - \frac{1}{\tau} \langle x, x' \rangle + \lambda \|x\|_1 \\ &= C + \frac{1}{2\tau} \|x\|^2 + \langle x, -A^* y + AA^* x' - \frac{1}{\tau} x' \rangle + \lambda \|x\|_1 \\ &= C' + \frac{1}{\tau} (\|x - (x' - \tau A^*(Ax' - y))\|^2 + \tau \lambda \|x\|_1). \end{aligned}$$

Proposition (24) shows that the minimizer of $\mathcal{E}_\tau(x, x')$ is thus indeed $\mathcal{S}_{\lambda\tau}^1(x' - \tau A^*(Ax' - y))$. \square

Of course, these iterations (11.14) are the same as the FB iterates (13.28), when, for the special case (11.6), one can consider a splitting of the form (13.28) defining

$$\mathcal{F} = \frac{1}{2} \|A \cdot -y\|^2 \quad \text{and} \quad \mathcal{G} = \lambda \|\cdot\|_1. \quad (11.15)$$

In the case (11.15), Proposition (24) shows that $\text{Prox}_{\rho J}$ is the soft thresholding.

11.4 Example: Sparse Deconvolution

11.4.1 Sparse Spikes Deconvolution

Sparse spikes deconvolution makes use of sparsity in the spacial domain, which corresponds to the orthogonal basis of Diracs $\psi_m[n] = \delta[n - m]$. This sparsity was first introduced in the seismic imaging community [], where the signal f_0 represent the change of density in the underground and is assumed to be composed of a few Diracs impulse.

In a simplified linearized 1D set-up, ignoring multiple reflexions, the acquisition of underground data f_0 is modeled as a convolution $y = h \star f_0 + w$, where h is a so-called “wavelet” signal sent in the ground. This should not be confounded with the construction of orthogonal wavelet bases detailed in Chapter ??, although the term “wavelet” originally comes from seismic imaging.

The wavelet filter h is typically a band pass signal that perform a tradeoff between space and frequency concentration especially tailored for seismic exploration. Figure (11.5) shows a typical wavelet that is a second derivative of a Gaussian, together with its Fourier transform. This shows the large amount of information removed from f during the imaging process.

The sparse ℓ^1 regularization in the Dirac basis reads

$$f^* = \underset{f \in \mathbb{R}^N}{\operatorname{argmin}} \frac{1}{2} \|f \star h - y\|^2 + \lambda \sum_m |f_m|.$$

Figure 11.5 shows the result of ℓ^1 minimization for a well chosen λ parameter, that was optimized in an oracle manner to minimize the error $\|f^* - f_0\|$.

The iterative soft thresholding for sparse spikes inversion iterates

$$\tilde{f}^{(k)} = f^{(k)} - \tau h \star (h \star f^{(k)} - y)$$

and

$$f_m^{(k+1)} = S_{\lambda\tau}^1(\tilde{f}_m^{(k)})$$

where the step size should obeys

$$\tau < 2/\|\Phi^* \Phi\| = 2/\max_{\omega} |\hat{h}(\omega)|^2$$

to guarantee convergence. Figure 11.6 shows the progressive convergence of the algorithm, both in term of energy minimization and iterates. Since the energy is not strictly convex, we note that convergence in energy is not enough to guarantee convergence of the algorithm.

11.4.2 Sparse Wavelets Deconvolution

Signal and image acquired by camera always contain some amount of blur because of objects being out of focus, movements in the scene during exposure, and diffraction. A simplifying assumption assumes a spatially invariant blur, so that Φ is a convolution

$$y = f_0 \star h + w.$$

In the following, we consider h to be a Gaussian filter of width $\mu > 0$. The number of effective measurements can thus be considered to be $P \sim 1/\mu$, since Φ nearly set to 0 large enough Fourier frequencies. Table ?? details the implementation of the sparse deconvolution algorithm.

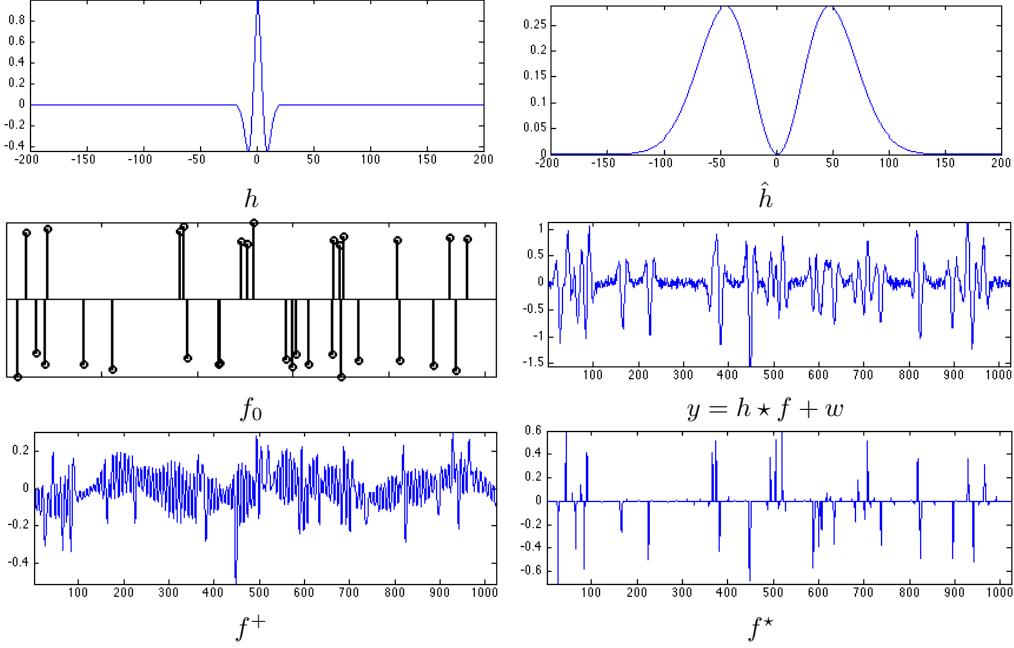


Figure 11.5: Pseudo-inverse and ℓ^1 sparse spikes deconvolution.

Figures 11.7 and 11.8 shows examples of signal and image acquisition with Gaussian blur.

Sobolev regularization (9.17) improves over ℓ^2 regularization (??) because it introduces an uniform smoothing that reduces noise artifact. It however fail to recover sharp edge and thus does a poor job in inverting the operator. To recover sharper transition and edges, one can use either a TV regularization or a sparsity in an orthogonal wavelet basis.

Figure 11.7 shows the improvement obtained in 1D with wavelets with respect to Sobolev. Figure 11.8 shows that this improvement is also visible for image deblurring. To obtain a better result with fewer artifact, one can replace the soft thresholding in orthogonal wavelets in during the iteration (??) by a thresholding in a translation invariant tight frame as defined in (8.10).

Figure 11.9 shows the decay of the SNR as a function of the regularization parameter λ . This SNR is computed in an oracle manner since it requires the knowledge of f_0 . The optimal value of λ was used in the reported experiments.

11.4.3 Sparse Inpainting

This section is a follow-up of Section 10.5.2.

To inpaint using a sparsity prior without noise, we use a small value for λ . The iterative thresholding algorithm (??) is written as follow for $\tau = 1$,

$$f^{(k+1)} = \sum_m S_\lambda^1(\langle P_y(f^{(k)}), \psi_m \rangle) \psi_m$$

Figure 11.10 shows the improvement obtained by the sparse prior over the Sobolev prior if one uses soft thresholding in a translation invariant wavelet frame.

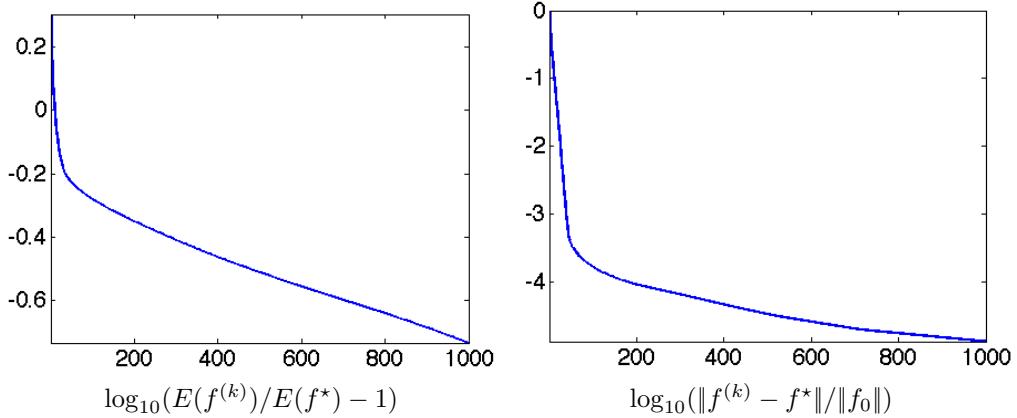


Figure 11.6: Decay of the energy and convergence through the iterative thresholding iterations.

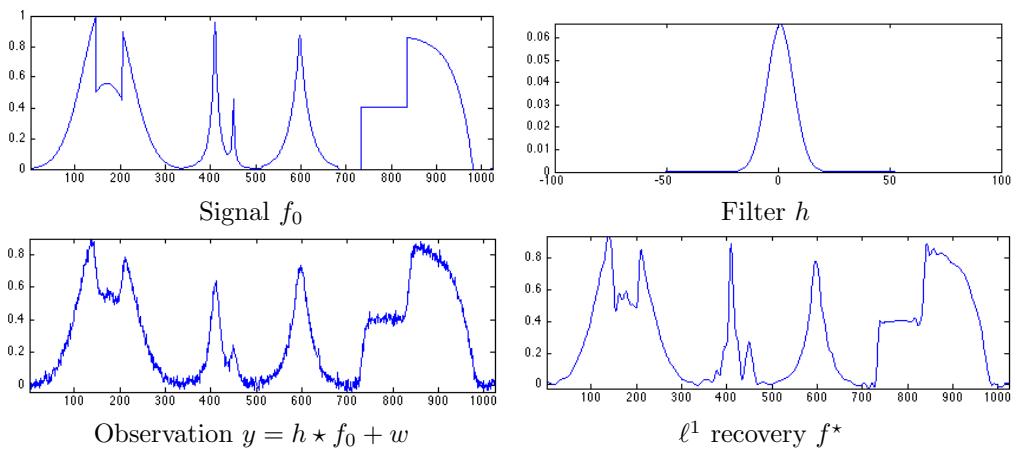


Figure 11.7: Sparse 1D deconvolution using orthogonal wavelets.

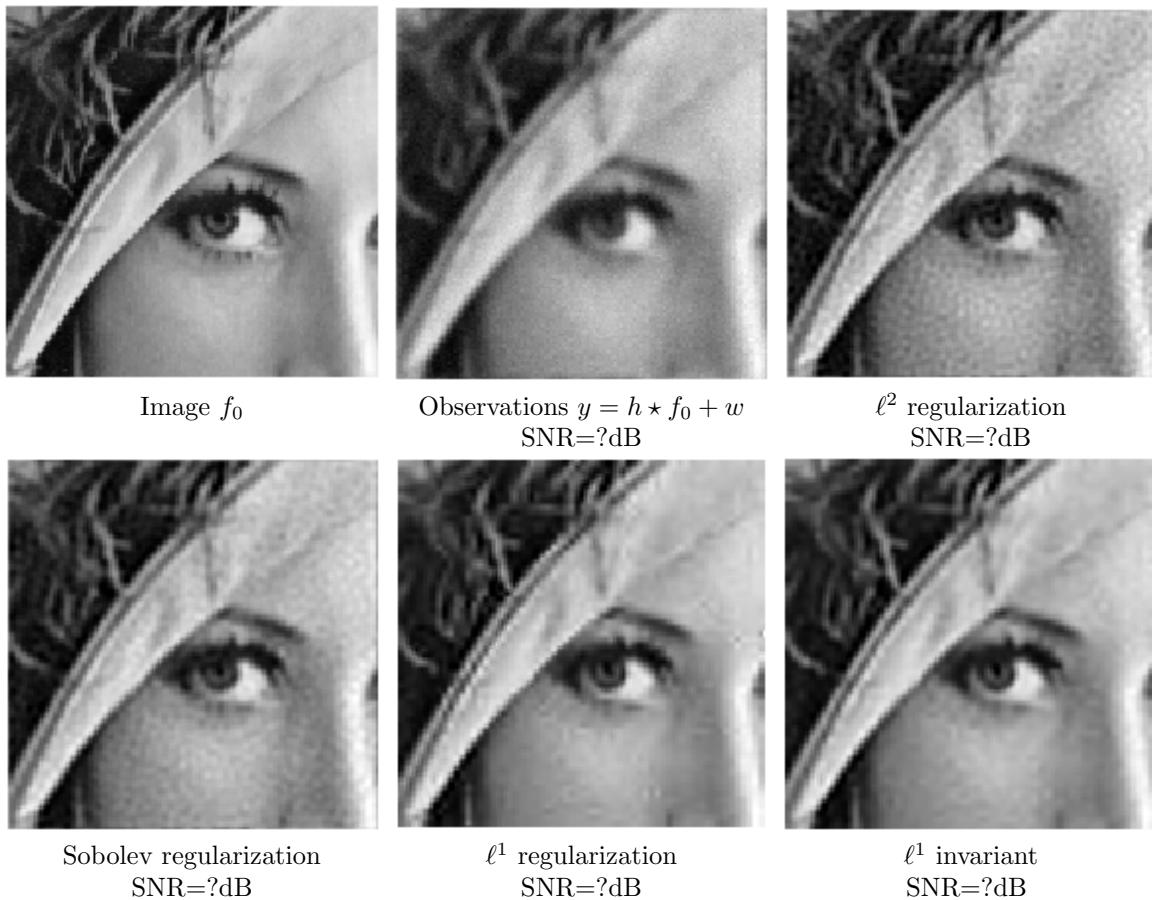


Figure 11.8: Image deconvolution.

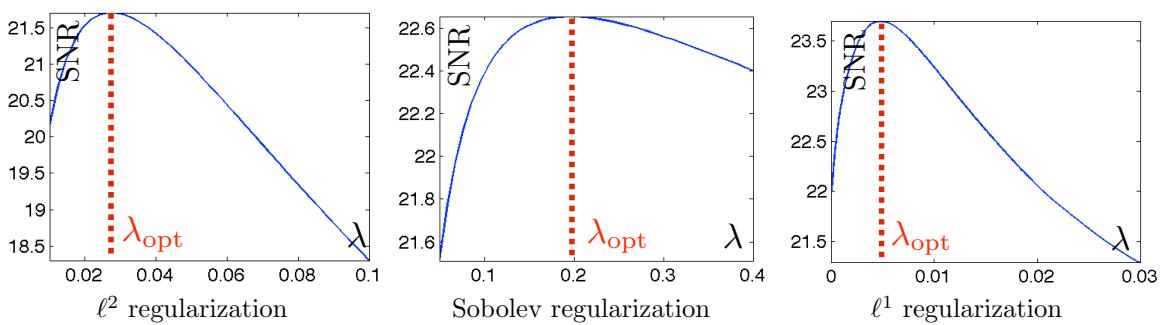


Figure 11.9: SNR as a function of λ .

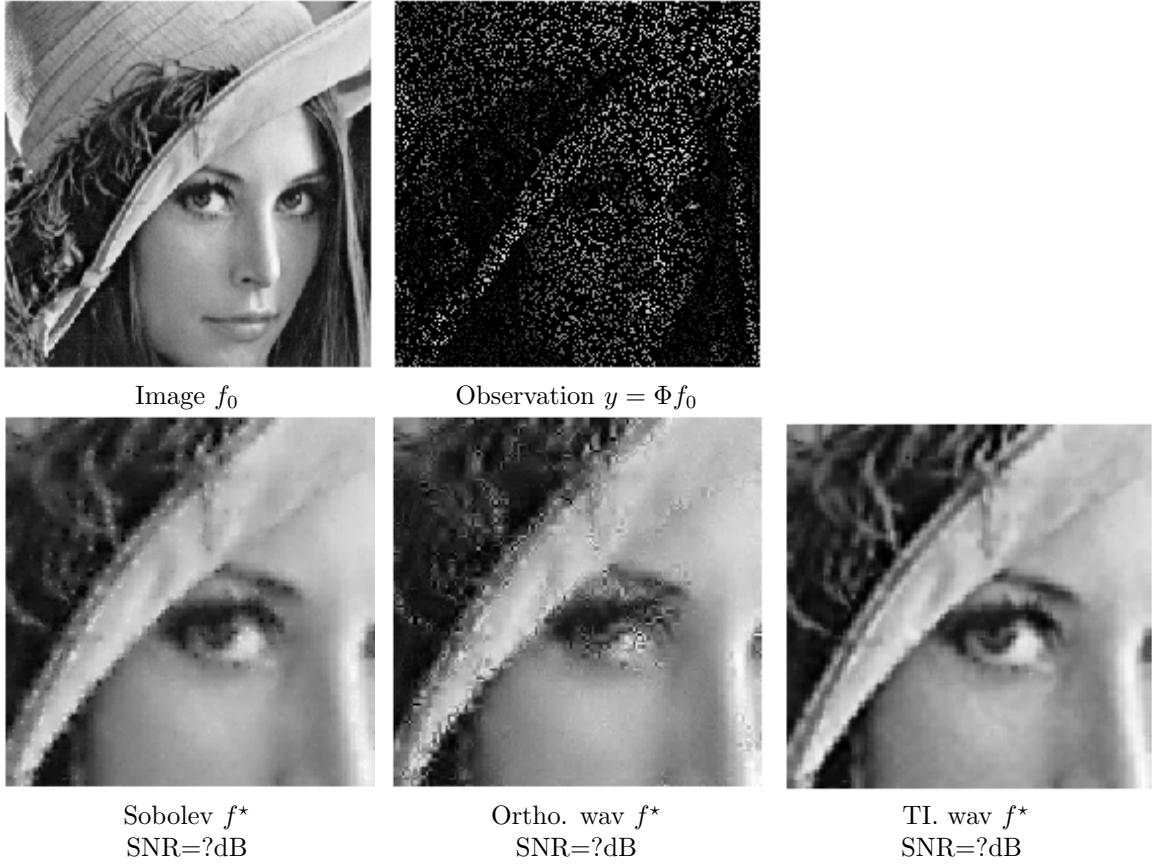


Figure 11.10: Inpainting with Sobolev and sparsity.

Chapter 12

Convex Analysis

The main references for this chapter are [13, 5]. This chapter uses different notations than the previous one, and we denote $f(x)$ a typical function to be minimized with respect to the variable x . We discuss here some important concepts from convex analysis and use them to study theoretically the performance of ℓ^1 -type methods.

12.1 Basics of Convex Analysis

We consider minimization problems of the form

$$\min_{x \in \mathcal{H}} f(x) \quad (12.1)$$

over the finite dimension (Hilbertian) space $\mathcal{H} \stackrel{\text{def.}}{=} \mathbb{R}^N$, with the canonical inner product $\langle \cdot, \cdot \rangle$. Most of the results of this chapter extends to possibly infinite dimensional Hilbert space.

Here $f : \mathcal{H} \rightarrow \bar{\mathbb{R}} \stackrel{\text{def.}}{=} \mathbb{R} \cup \{+\infty\}$ is a convex function. Note that we allow here f to take the value $+\infty$ to integrate constraints in the objective, and the constraint set is thus the “domain” of the function

$$\text{dom}(f) \stackrel{\text{def.}}{=} \{x ; f(x) < +\infty\}.$$

A useful notation is the indicator function of a set $\mathcal{C} \subset \mathcal{H}$

$$\iota_{\mathcal{C}}(x) \stackrel{\text{def.}}{=} \begin{cases} 0 & \text{if } x \in \mathcal{C}, \\ +\infty & \text{otherwise.} \end{cases}$$

12.1.1 Convex Sets and Functions

A convex set $\Omega \subset \mathcal{H}$ is such that

$$\forall (x, y, t) \in \mathcal{H}^2 \times [0, 1], \quad (1-t)x + ty \in \Omega.$$

A convex function is such that

$$\forall (x, y, t) \in \mathcal{H}^2 \times [0, 1], \quad f((1-t)x + ty) \leq (1-t)f(x) + tf(y) \quad (12.2)$$

and this is equivalent to its epigraph $\{(x, r) \in \mathcal{H} \times \mathbb{R} ; r \geq f(x)\}$ being a convex set. Note that here we use \leq as a comparison over $\bar{\mathbb{R}}$. The function f is strictly convex if equality in (12.2) only holds for $t \in \{0, 1\}$. A set Ω being convex is equivalent to $\iota_{\mathcal{C}}$ being a convex function.

In the remaining part of this chapter, we consider convex functions f which are proper, i.e. such that $\text{dom}(f) \neq \emptyset$, and that should be lower-semi-continuous (lsc), i.e. such that for all $x \in \mathcal{H}$,

$$\liminf_{y \rightarrow x} f(y) \geq f(x).$$

It is equivalent to $\text{epi}(f)$ being a closed convex set. We denote $\Gamma_0(\mathcal{H})$ the set of proper convex lsc functions.

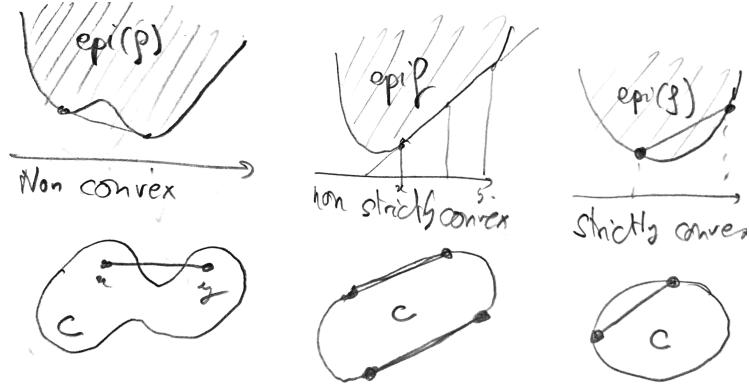


Figure 12.1: Convexity and strict convexity for function and sets.

12.1.2 First Order Conditions

Existence of minimizers. Before looking at first optimality conditions, one has to check that there exists minimizers, which is implied by the l.s.c. property and coercivity.

Proposition 26. *If f is l.s.c. and coercive (i.e. $f(x) \rightarrow +\infty$ as $x \rightarrow +\infty$), then there exists a minimizer x^* of f .*

Proof. Since f is coercive, it is bounded from below, one can consider a minimizing sequence $(x_n)_n$ such that $f(x_n) \rightarrow \min f$. Since f is l.s.c., this implies that the sub-level sets of f are closed, and coercivity imply they are bounded, hence compact. One can thus extract from $(x_n)_n$ a converging sub-sequence $(x_{n(p)})_p$, $x_{n(p)} \rightarrow x^*$. Lower semi-continuity implies that $\min f = \lim_p f(x_{n(p)}) \geq f(x^*)$, and hence x^* is a minimizer. \square

This existence proof is often called the “direct method of calculus of variation”. Note that if the function f is in $\Gamma_0(\mathcal{H})$, then the set of minimizer $\operatorname{argmin} f$ is a closed convex set, and all local minimizers (i.e. minimizer of the function restricted to an open ball) are global one. If it is furthermore strictly convex, then there is a single minimizer.

Sub-differential. The sub-differential at x of such a f is defined as

$$\partial f(x) \stackrel{\text{def.}}{=} \{u \in \mathcal{H}^* ; \forall y, f(y) \geq f(x) + \langle u, y - x \rangle\}.$$

We denote here $\mathcal{H}^* = \mathbb{R}^N$ the set of “dual” vector. Although in finite dimensional Euclidean space, this distinction is not needed, it helps to distinguish primal from dual vectors, and recall that the duality pairing implicitly used depends on the choice of an inner product. The sub-differential $\partial f(x)$ is thus the set of “slopes” u of tangent affine planes $f(x) + \langle u, z - x \rangle$ that fits bellow the graph of f .

Note that f being differentiable at x is equivalent to the sub-differential being reduced to a singleton (equal to the gradient vector)

$$\partial f(x) = \{\nabla f(x)\}.$$

Informally, the “size” of $\partial f(x)$ controls how smooth f is at x .

Note that one can have $\partial f(x) = \emptyset$, for instance if $x \notin \operatorname{dom}(f)$. Note also that one can still have $x \in \operatorname{dom}(f)$ and $\partial f(x) = \emptyset$, for instance take $f(x) = -\sqrt{1-x^2} + \iota_{[-1,1]}(x)$ at $x = \pm 1$.

Since $\partial f(x) \subset \mathcal{H}^*$ is an intersection of half space, it is a closed convex set. The operator $\partial f : \mathcal{H} \mapsto 2^{\mathcal{H}^*}$ is thus “set-valued”, and we often denote this as $\partial f : \mathcal{H} \hookrightarrow \mathcal{H}^*$.

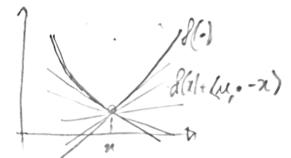


Figure 12.2: The subdifferential

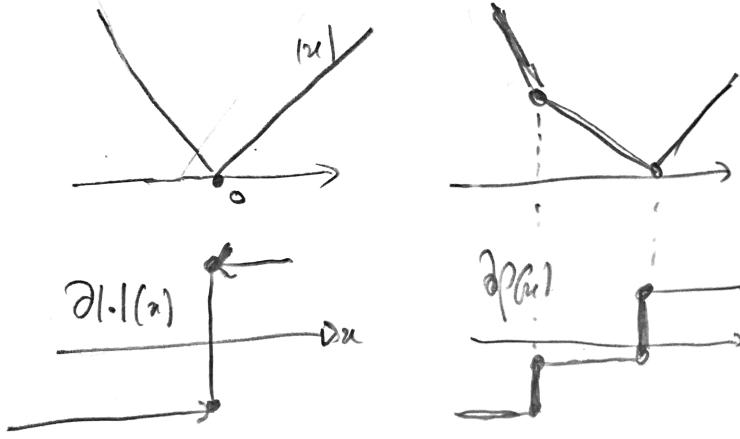


Figure 12.3: Subdifferential of the absolute value and a piecewise affine convex function.

Remark 6 (Maximally monotone operator). The operator ∂f is particular instance of so-called monotone operator, since one can check that $U = \partial f$ satisfies

$$\forall (u, v) \in U(x) \times U(y), \quad \langle y - x, v - u \rangle \geq 0.$$

In the 1-D setting, being monotone is the same as being an increasing map. Sub-differential can also be shown to be maximally monotone, in the sense that such an operator is not strictly included in the graph of another monotone operator. Note that there exists monotone maps which are not subdifferential, for instance $(x, y) \mapsto (-y, x)$. Much of the theory of convex analysis and convex optimization can be extended to deal with arbitrary maximally monotone-maps in place of subdifferential, but we will not pursue this here.

A prototypical example is the absolute value $f(x) = |\cdot|$, and writing conveniently $\partial f(x) = \partial|\cdot|(x)$, one verifies that

$$\partial|\cdot|(x) = \begin{cases} -1 & \text{if } x < 0, \\ +1 & \text{if } x > 0, \\ [-1, 1] & \text{if } x = 0. \end{cases}$$

First Order Conditions. The subdifferential is crucial for this simple but extremely important proposition.

Proposition 27. x^* is a minimizer of f if and only if $0 \in \partial f(x^*)$.

Proof. One has

$$x^* \in \operatorname{argmin} f \iff (\forall y, f(x^*) \leq f(y) + \langle 0, x^* - y \rangle) \iff 0 \in \partial f(x^*).$$

□

Sub-differential calculus. There is a large set of calculus rules that allows to simplify the computation of sub-differentials. For decomposable function $f(x_1, \dots, x_K) = \sum_{k=1}^K f_k(x_k)$, the sub-differential is the product of the sub-differentials

$$\partial f(x_1, \dots, x_K) = \partial f_1(x_1) \times \dots \times \partial f_K(x_K).$$

This can be used to compute the sub-differential of the ℓ^1 norm $\|x\|_1 = \sum_{k=1}^N |x_k|$

$$\partial\|\cdot\|_1(x) = \prod_{k=1}^N \partial|\cdot|(x_k)$$

which is thus an hyper rectangle. This means that, denoting $I = \text{supp}(x)$, one has $u \in \partial\|\cdot\|_1(x)$ is equivalent to

$$u_I = \text{sign}(x_I) \quad \text{and} \quad \|u_{I^c}\|_\infty \leq 1.$$

A tricky problem is to compute the sub-differential of the sum of two functions. If one of the two function is continuous at x (i.e. it has a finite value), then

$$\partial(f + g)(x) = \partial f(x) \oplus \partial g(x) = \{u + v ; (u, v) \in \partial f(x) \times \partial g(x)\}$$

where \oplus thus denotes the Minkowski sum. For instance, if f is differentiable at x , then

$$\partial(f + g)(x) = \nabla f(x) + \partial g(x) = \{\nabla f(x) + v ; v \in \partial g(x)\}.$$

Positive linear scaling is simple to handle

$$\forall \lambda \in \mathbb{R}_+, \quad \partial(\lambda f)(x) = \lambda(\partial f(x)).$$

The chain rule for sub-differential is difficult since in general composition does not work so-well with convexity. The only simple case is composition with linear functions, which preserves convexity. Denoting $A \in \mathbb{R}^{P \times N}$ and $f \in \Gamma_0(\mathbb{R}^P)$, one has that $f \circ A \in \Gamma_0(\mathbb{R}^N)$ and

$$\partial(f \circ A)(x) = A^*(\partial f)(Ax) \stackrel{\text{def.}}{=} \{A^*u ; u \in \partial f(Ax)\}.$$

Normal cone. The sub-differential of an indicator function is a convex cone, the so-called normal cone to the constraint

$$\forall x \in \mathcal{C}, \quad \partial \iota_{\mathcal{C}}(x) = \mathcal{N}_{\mathcal{C}}(x) \stackrel{\text{def.}}{=} \{v ; \forall z \in \mathcal{C}, \langle z - x, v \rangle \leq 0\}.$$

Note that for $x \notin \mathcal{C}$, $\partial \iota_{\mathcal{C}}(x) = \emptyset$. For an affine space $\mathcal{C} = a + \mathcal{V}$ where $\mathcal{V} \subset \mathcal{H}$ is a linear space, then $\mathcal{N}_{\mathcal{C}}(x) = \mathcal{V}^\perp$ is the usual orthogonal for linear spaces. If $x \in \text{int}(\mathcal{C})$ is in the interior of \mathcal{C} , then $\mathcal{N}_{\mathcal{C}}(x) = \{0\}$. In some sense, the more non-regular the boundary of \mathcal{C} is at x , the larger is the normal cone.

The normal cone is a way to express first order condition for constrained problem

$$\min_{x \in \mathcal{C}} f(x)$$

which reads, if f is continuous

$$0 \in \partial f(x) + \partial \iota_{\mathcal{C}}(x) \Leftrightarrow \exists \xi \in \partial f(x), -\xi \in \mathcal{N}_{\mathcal{C}}(x) \Leftrightarrow \partial f(x) \cap (-\mathcal{N}_{\mathcal{C}}(x)) \neq \emptyset.$$

If f is differentiable, it reads $-\nabla f(x) \in \mathcal{N}_{\mathcal{C}}(x)$.

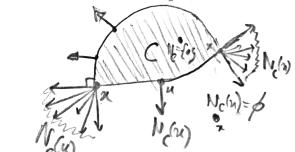


Figure 12.4: Normal cones

12.2 Convex Duality

Duality is associated to a particular formulation of the optimization problem, so that for instance making change of variables results in a different duality.

12.2.1 Lagrange Duality

We consider a minimization of the form

$$p^* = \min_{x \in \mathbb{R}^N} \{f(x) ; Ax = y \quad \text{and} \quad g(x) \leq 0\} \quad (12.3)$$

for a continuous convex functions $f : \mathcal{H} \rightarrow \mathbb{R}$, a matrix $A \in \mathbb{R}^{P \times N}$ and a function $g : \mathcal{H} \rightarrow \mathbb{R}^Q$ such that each of its coordinates $g_i : \mathcal{H} \rightarrow \mathbb{R}$ are continuous and convex. Note that it is always the case that equality in convex program corresponds to affine ones. One can always write a convex minimization problem with

positivity constraints in the form (12.3), although there exists infinite way of doing so (each one giving a different duality formula).

Here we have assumed for simplicity that f is continuous, i.e. $\text{dom}(f) = \mathbb{R}^N$. The following exposition can be generalized to $\text{dom}(f)$ being arbitrary, but this is more technical. For the sake of simplicity, we thus assume all the constraint defining the domain are encoded in $Ax = y$ and $g(x) \leq 0$

Note that it is possible to generalize the previous Lagrange duality results by replacing “ $x \geq 0$ ” by “ $X \succeq 0$ ” where X is a matrix (and in fact even more generally using convex cones).

We use the following fact

$$\sup_{u \in \mathbb{R}^P} \langle r, u \rangle = \begin{cases} 0 & \text{if } r = 0, \\ +\infty & \text{if } r \neq 0, \end{cases} \quad \text{and} \quad \sup_{v \in \mathbb{R}_+^Q} \langle s, v \rangle = \begin{cases} 0 & \text{if } s \leq 0, \\ +\infty & \text{otherwise,} \end{cases}$$

to encode the constraints $r = Ax - y = 0$ and $s = g(x) \leq 0$.

One can represent the constraints appearing in (12.3) conveniently using a maximization over so-called Lagrange multipliers

$$p^* = \inf_x \max_{u \in \mathbb{R}^P, v \in \mathbb{R}_+^Q} \mathcal{L}(x, u, v) \stackrel{\text{def.}}{=} f(x) + \langle Ax - y, u \rangle + \langle g(x), v \rangle.$$

It is tempting to inverse the inf and the sup, and study

$$d^* = \sup_{(u, v) \in \mathbb{R}^P \times \mathbb{R}_+^Q} F(u, v) \stackrel{\text{def.}}{=} \inf_x f(x) + \langle Ax - y, u \rangle + \langle g(x), v \rangle. \quad (12.4)$$

One remarks that F is a concave function (as being the minimum of linear forms), and this “dual” problem is thus a maximization of a concave function.

The following proposition is the so-called weak duality, which assert that values of the dual problems always lower bounds values of the primal one

Proposition 28. *One always has, for all $(u, v) \in \mathbb{R}^P \times \mathbb{R}_+^Q$,*

$$F(u, v) \leq p^* \implies d^* \leq p^*.$$

Proof. Since $g(x) \leq 0$ and $v \geq 0$, one has $\langle g(x), v \rangle \leq 0$, and since $Ax = y$, one has $\langle Ax - y, u \rangle = 0$, so that

$$\mathcal{L}(x, u, v) \leq f(x) \implies F(u, v) = \inf_x \mathcal{L}(x, u, v) \leq \inf_x f(x) = p^*.$$

□

The following fundamental theorem, more difficult to prove, gives a sufficient condition (so-called qualification of the constraints) such that one actually has equality.

Theorem 32. *If*

$$\exists x_0 \in \mathbb{R}^N, \quad Ax_0 = y \quad \text{and} \quad g(x_0) < 0, \quad (12.5)$$

then $p^ = d^*$. Furthermore, x^* and (u^*, v^*) are solutions of respectively (12.3) and (12.4) if and only if*

$$Ax^* = y, \quad g(x^*) \leq 0, \quad u^* \geq 0 \quad (12.6)$$

$$0 \in \partial f(x^*) + A^* u^* + \sum_i v_i^* \partial g_i(x^*) \quad (12.7)$$

$$\forall i, \quad u_i^* g_i(x^*) = 0 \quad (12.8)$$

The existence of such an x_0 is called “constraint qualification”, and as written here, this corresponds to the so-called “Slater” qualification condition (many other weaker sufficient conditions exist).

Condition (12.6) is simply the primal and dual constraints. Condition (12.7) is the first order condition for the minimization of $\mathcal{L}(x, u, v)$ over x . Condition (12.8) is the first order condition for the maximization of $\mathcal{L}(x, u, v)$ over (u, v) . These three conditions are often referred to as “Karush-Kuhn-Tucker” (KKT) conditions, and under a constraint qualification condition, they are necessary and sufficient condition for optimality.

The last condition $u_i^* g_i(x^*) = 0$ (so called “complementary slackness”) states that if $g_i(x^*) < 0$ (the constraints is not saturated) then $u_i = 0$, and also that if $u_i > 0$ then $g_i(x^*) = 0$.

Note that it is possible to weaken the hypotheses of this theorem, for the linear constraints of the form $g_i(x) = \langle x, h_i \rangle - c_i \leq 0$, by replacing the $g_i(x_0) < 0$ by the weaker condition $\langle x_0, h_i \rangle \leq c_i$.

One can generalize this theorem to the setting where $\text{dom}(f)$ is not equal to \mathbb{R}^N (i.e. it is not continuous, and thus integrates extra constraint beside the \leq). In this case, one has to add the extra constraint $x_0 \in \text{relint}(\text{dom}(f))$.

Theorem 32 generalizes the necessary conditions provided by Lagrange multipliers for equality constrained optimization. The setting is both more complex because one can deal with inequalities that might be saturated (so this introduce positivity constraints on the multipliers v) but also simpler because of convexity (which thus gives also necessary conditions).

As a simple example, we now derive the dual for a simple linear projection problem. A more complicated computation is carried over in Section 14.1.5 for the Lasso. We consider

$$p^* = \min_{Ax=y} \frac{1}{2} \|x - z\|^2 = \min_x \max_u \frac{1}{2} \|x - z\|^2 + \langle Ax - y, u \rangle = \max_u F(u) = \min_x \frac{1}{2} \|x - z\|^2 + \langle Ax - y, u \rangle,$$

where we used the fact that strong duality holds because only linear constraints are involved. For each u , the optimal x satisfies $x - z + A^*u$, i.e. $x = z - A^*u$, so that

$$F(u) = \frac{1}{2} \|A^*u\|^2 + \langle A(z - A^*u) - y, u \rangle = -\frac{1}{2} \|A^*u\|^2 + \langle u, Az - y \rangle.$$

Weak duality states $p^* \geq F(u)$ for any u , and $p^* = F(u^*)$ where the optimal u^* satisfies $AA^*u = Az - y$. If $y \in \text{Im}(A)$, then such a u^* exists and can be chosen as $u^* = u = (AA^*)^{-1}(Az - y)$, and the (unique) primal solution reads

$$x^* = \text{Proj}_{A^*=y}(z)(\text{Id} - A^+A)z - A^+y. \quad (12.9)$$

12.2.2 Legendre-Fenchel Transform

In order to simplify and accelerate computation involving Lagrange duality, it is very convenient to introduce a particular transformation of convex functions the Legendre-Fenchel transform. In some sense, it is the canonical “isomorphisms” (pairing) between convex functions. In spirit, it plays a similar role for convex function as the Fourier transform for signal or images.

For $f \in \Gamma_0(\mathcal{H})$, we define its Legendre-Fenchel transform as

$$f^*(u) \stackrel{\text{def.}}{=} \sup_x \langle x, u \rangle - f(x). \quad (12.10)$$

Being the maximum of affine functional, one obtains that f^* is itself a convex function, and that in fact $f^* \in \Gamma_0(\mathcal{H}^*)$. One can prove the following fundamental bi-duality result.

Theorem 33. *One has*

$$\forall f \in \Gamma_0(\mathcal{H}), \quad (f^*)^* = f.$$

In fact, f^* is convex even in the case where f is not, and f^{**} is the convex envelop of f (i.e. the largest convex function smaller than f). [ToDo: drawing]

One has the following basic property relating the sub-differentials of f and f^* .

Proposition 29. *One has $\partial f^* = (\partial f)^{-1}$, where the inverse of a set valued map is defined in (13.23), and*

$$\forall (x, y), \quad \langle x, y \rangle \leq f(x) + f^*(y) \quad \text{and} \quad \langle x, y \rangle = f(x) + f^*(y) \quad \Leftrightarrow \quad x \in \partial f^*(y) \quad \Leftrightarrow \quad y \in \partial f(x).$$

Proposition 30. For $1/p + 1/q = 1$,

$$(\iota_{\|\cdot\|_p \leq 1})^* = \|\cdot\|_q \quad \text{and} \quad (\|\cdot\|_q)^* = \iota_{\|\cdot\|_p \leq 1}$$

Let us now give some example of Legendre transform.

Proposition 31. For $f(x) = \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle$ with A invertible, then $f^*(u) = \frac{1}{2}\langle A^{-1}u, u\rangle - \frac{1}{2}\langle A^{-1}b, b\rangle$. In particular, for $f = \|\cdot\|^2/2$, then $f^* = f$. One has [ToDo: check]

$$f(\cdot - z)^* = f + \langle z, \cdot \rangle, \quad (f + \langle z, \cdot \rangle)^* = f(\cdot - z), \quad (\lambda f)^* = \lambda f^*(\cdot/\lambda).$$

Proof. One has $f^*(u) = \langle Ax^*, x^*\rangle - \langle b, x^*\rangle$ where x^* solves

$$u = Ax^* - b \implies x^* = A^{-1}u + A^{-1}b.$$

Hence

$$f^*(u) = \frac{1}{2}\langle AA^{-1}(u + b), A^{-1}(u + b)\rangle - \langle b, A^{-1}(u + b)\rangle = \frac{1}{2}\langle A^{-1}u, u\rangle - \frac{1}{2}\langle A^{-1}b, b\rangle$$

□

Legendre transform and smoothness. While the Fourier transform is a pairing between smoothness and decay (see Section ??), the Legendre-Fenchel is really a pairing between smoothness and strong convexity. This can be intuitively seen by the fact that the Legendre-Fenchel inverts the sub-differentials (??) and hence when the functions involved are \mathcal{C}^2 , it inverse the Hessians

$$\partial^2 f(x) = (\partial^2 f^*(y))^{-1} \quad \text{at} \quad y = \nabla f(x).$$

This relation between Hessian can be seen as implying the exchange of strong convexity and uniform bound on the Hessian, as detailed in Proposition 33.

Proposition 32. One has

$$\nabla f \text{ is } L\text{-Lipschitz} \iff \nabla f^* \text{ is } \mu\text{-strongly convex}.$$

This results suggests a way to smooth any function f . Instead of doing a convolution, one can use the infimal convolution

$$(f \otimes g)(x) \stackrel{\text{def.}}{=} \sup_{y+y'=x} f(y) + g(y').$$

One can check that if (f, g) are convex, so is $f \otimes g$, and that the Legendre transform actually exchanges sum and inf-convolution

$$(f + g)^* = f \otimes g \quad \text{and} \quad (f \otimes g)^* = f + g.$$

The Moreau-Yosida regularization of f is corresponds to a μ -strict-convexification of f^* , i.e.

$$f_\mu \stackrel{\text{def.}}{=} f \otimes \left(\frac{1}{2\mu} \|\cdot\|^2 \right) = (f^* + \frac{\mu}{2} \|\cdot\|^2)^*. \quad (12.11)$$

Since $f^* + \frac{\mu}{2} \|\cdot\|^2$ is at least μ -strongly convex, then f_μ as a $1/\mu$ -Lipchitz gradient.

As an example, the Moreau-Yosida regularization of the absolute value reads

$$(|\cdot|_\mu)(x) = \begin{cases} \frac{1}{2\mu}x^2 & \text{if } |x| \leq \mu, \\ |x| - \frac{\mu}{2} & \text{if } |x| > \mu. \end{cases}$$

This should be compared with the regularization $\sqrt{x^2 + \mu^2}$ (which is above the curve) that we used previously. [ToDo: add drawing]

12.2.3 Fenchel-Rockafellar Duality

Very often the Lagrange dual can be expressed using the conjugate of the function f . We give here a particularly important example, which is often called Fenchel-Rockafellar Duality.

We consider the following structured minimization problem

$$p^* = \inf_x f(x) + g(Ax). \quad (12.12)$$

Re-writing it as

$$\inf_{y=Ax} f(x) + g(y),$$

we can form the primal-dual problem

$$\inf_{(x,y)} \sup_u f(x) + g(y) + \langle Ax - y, u \rangle.$$

If sufficient condition on the domain of (f, g) holds (such as those stated in Theorem ??), one can exchange the min and the max and obtains the dual problem

$$d^* = \sup_u \min_{(x,y)} f(x) + g(y) + \langle Ax - y, u \rangle \quad (12.13)$$

$$= \sup_u \left(\min_x \langle x, A^*u \rangle + f(x) \right) + \left(\min_y -\langle y, u \rangle + g(y) \right) \quad (12.14)$$

which leads to the celebrated Fenchel-Rockafellar, which we summarize together with qualification sufficient condition ensuring strong duality.

Theorem 34 (Fenchel-Rockafellar). *If*

$$0 \in \text{relint}(\text{dom}(g)) - A \text{relint}(\text{dom}(f)) \quad (12.15)$$

the one has the following strong duality

$$\inf_x f(x) + g(Ax) = \inf_x \sup_u \mathcal{L}(x, u) = \sup_u \inf_x \mathcal{L}(x, u) = \sup_u -f^*(-A^*u) - g^*(u) \quad (12.16)$$

$$\text{where } \mathcal{L}(x, u) \stackrel{\text{def.}}{=} f(x) + \langle Ax, u \rangle - g^*(u).$$

Furthermore one has that (x^, u^*) is a pair of optimal primal-dual solutions if and only if*

$$-A^*u^* \in \partial f(x^*) \quad \text{and} \quad Ax^* \in \partial g^*(u^*). \quad (12.17)$$

Condition (12.15) is the constraint qualification ensuring that one can inverse the inf and the sup in (12.16). It can be recovered from Slater's qualification condition (12.5) when deriving the dual problem as in (12.13). The primal-dual relations (12.17) are the first order condition along the x and the u variables in minimization and maximization of \mathcal{L} . They are sometimes summarised in "matrix" form

$$0 \in \begin{pmatrix} \partial f & A^* \\ -A & \partial g^* \end{pmatrix} \begin{pmatrix} x^* \\ u^* \end{pmatrix}.$$

Chapter 13

Convex Optimization

The main references for this chapter are [9, 10, 5], see also [30, 4, 3].

We consider a general convex optimization problem

$$\min_{x \in \mathcal{H}} f(x) \quad (13.1)$$

where $\mathcal{H} = \mathbb{R}^N$ is a finite dimensional Hilbertian (i.e. Euclidean) space, and try to devise “cheap” algorithms with a low computational cost per iterations. The class of algorithms considered are first order, i.e. they make use of gradient information.

13.1 Gradient Descent Methods

We have already encountered the gradient descent method informally in Section ?? for the regularization of inverse problem. We now give a detailed analysis of the method.

13.1.1 Gradient Descent

The optimization program (10.26) is a example of unconstrained convex optimization of the form (13.1) where $f : \mathcal{H} \rightarrow \mathbb{R}$ is a \mathcal{C}^1 function with Lipschitz gradient (so-called “smooth” function). Recall that the gradient $\nabla f : \mathcal{H} \mapsto \mathcal{H}$ of this functional (not to be confound with the discretized gradient $\nabla x \in \mathcal{H}$ of f) is defined by the following first order relation

$$f(x + r) = f(x) + \langle f, r \rangle_{\mathcal{H}} + O(\|r\|_{\mathcal{H}}^2)$$

where we used $O(\|r\|_{\mathcal{H}}^2)$ in place of $o(\|r\|_{\mathcal{H}})$ (for differentiable function) because we assume here f is of class \mathcal{C}^1 (i.e. the gradient is continuous). Section 10.4.3 shows typical examples of gradient computation.

For such a function, the gradient descent algorithm is defined as

$$x^{(\ell+1)} \stackrel{\text{def.}}{=} x^{(\ell)} - \tau_{\ell} \nabla f(x^{(\ell)}), \quad (13.2)$$

where the step size $\tau_{\ell} > 0$ should be small enough to guarantee convergence, but large enough for this algorithm to be fast.

One also needs to quantify the smoothness of f . This is enforced by requiring that the gradient is L -Lipschitz, i.e.

$$\forall (x, x') \in \mathcal{H}^2, \quad \|\nabla f(x) - \nabla f(x')\| \leq L \|x - x'\|. \quad (\mathcal{R}_L)$$

In order to obtain fast convergence of the iterates themselves, it is needed that the function has enough “curvature” (i.e. is not too flat), which corresponds to imposing that f is μ -strongly convex

$$\forall (x, x') \in \mathcal{H}^2, \quad \langle \nabla f(x) - \nabla f(x'), x - x' \rangle \geq \mu \|x - x'\|^2. \quad (\mathcal{S}_{\mu})$$

The following proposition express these conditions as constraints on the hessian for \mathcal{C}^2 functions.

Proposition 33. *Conditions (\mathcal{R}_L) and (\mathcal{S}_μ) imply*

$$\forall (x, x'), \quad f(x') + \langle \nabla f(x), x' - x \rangle + \frac{\mu}{2} \|x - x'\|^2 \leq f(x) \leq f(x') + \langle \nabla f(x'), x' - x \rangle + \frac{L}{2} \|x - x'\|^2. \quad (13.3)$$

If f is of class \mathcal{C}^2 , conditions (\mathcal{R}_L) and (\mathcal{S}_μ) are equivalent to

$$\forall x, \quad \mu \text{Id}_{N \times N} \preceq \partial^2 f(x) \preceq L \text{Id}_{N \times N} \quad (13.4)$$

where $\partial^2 f(x) \in \mathbb{R}^{N \times N}$ is the Hessian of f , and where \preceq is the natural order on symmetric matrices, i.e.

$$A \preceq B \iff \forall x \in \mathcal{H}, \quad \langle Au, u \rangle \leq \langle Bu, u \rangle.$$

Proof. We prove (13.3), using Taylor expansion with integral remain

$$f(x') - f(x) = \int_0^1 \langle \nabla f(x_t), x' - x \rangle dt = \langle \nabla f(x), x' - x \rangle + \int_0^1 \langle \nabla f(x_t) - \nabla f(x), x' - x \rangle dt$$

where $x_t \stackrel{\text{def}}{=} f + t(x' - x)$. Using Cauchy-Schwartz, and then the smoothness hypothesis (\mathcal{R}_L)

$$f(x') - f(x) \leq \langle \nabla f(x), x' - x \rangle + \int_0^1 L \|x_t - f\| \|x' - x\| dt \leq \langle \nabla f(x), x' - x \rangle + L \|x' - x\|^2 \int_0^1 t dt$$

which is the desired upper-bound. Using directly (\mathcal{S}_μ) gives

$$f(x') - f(x) = \langle \nabla f(x), x' - x \rangle + \int_0^1 \langle \nabla f(x_t) - \nabla f(x), \frac{x_t - x}{t} \rangle dt \geq \langle \nabla f(x), x' - x \rangle + \mu \int_0^1 \frac{1}{t} \|x_t - x\|^2 dt$$

which gives the desired result since $\|x_t - x\|^2/t = t\|x' - x\|^2$. \square

The relation (13.3) shows that a smooth (resp. strongly convex) functional is bellow a quadratic tangential majorant (resp. minorant).

Condition (13.4) thus reads that the singular values of $\partial^2 f(x)$ should be contained in the interval $[\mu, L]$. The upper bound is also equivalent to $\|\partial^2 f(x)\|_{\text{op}} \leq L$ where $\|\cdot\|_{\text{op}}$ is the operator norm, i.e. the largest singular value. In the special case of a quadratic function \mathcal{Q} of the form (10.24), $\partial^2 f(x) = A$ is constant, so that $[\mu, L]$ can be chosen to be the range of the singular values of A .

In order to get some insight on the convergence proof and the associated speed, we first study the simple case of a quadratic functional.

Proposition 34. *For $f(x) = \langle Ax, x \rangle - \langle b, x \rangle$ with the singular values of A upper-bounded by L , assuming there exists $(\tau_{\min}, \tau_{\max})$ such that*

$$0 < \tau_{\min} \leq \tau_\ell \leq \tilde{\tau}_{\max} < \frac{2}{L} \quad (13.5)$$

then there exists $0 \leq \tilde{\rho} < 1$ such that

$$\|x^{(\ell)} - x^*\| \leq \tilde{\rho}^\ell \|x^{(0)} - x^*\|. \quad (13.6)$$

If the singular values are lower bounded by μ , then the best rate $\tilde{\rho}$ is obtained for

$$\tau_\ell = \frac{2}{L + \mu} \implies \tilde{\rho} \stackrel{\text{def.}}{=} \frac{L - \mu}{L + \mu} = 1 - \frac{2\varepsilon}{1 + \varepsilon} \quad \text{where } \varepsilon \stackrel{\text{def.}}{=} \mu/L. \quad (13.7)$$

Proof. One iterate of gradient descent reads

$$x^{(\ell+1)} = x^{(\ell)} - \tau_\ell (Ax^{(\ell)} - b).$$

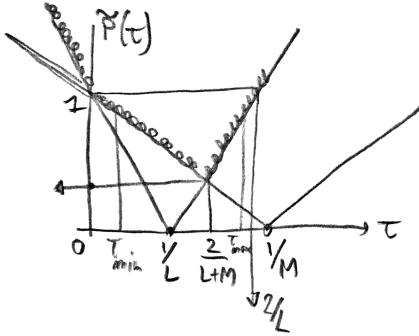


Figure 13.1: Contraction constant $h(\tau)$ for a quadratic function (right).

Since the solution x^* (which by the way is unique by strict convexity) satisfy the first order condition $Ax^* = b$, it gives

$$x^{(\ell+1)} - x^* = x^{(\ell)} - x^* - \tau_\ell A(x^{(\ell)} - x^*) = (\text{Id}_N - \tau_\ell A)(x^{(\ell)} - x^*).$$

One thus has to study the contractance ratio of the linear map $\text{Id}_N - \tau_\ell A$, i.e. its largest singular value, which reads

$$h(\tau) \stackrel{\text{def.}}{=} \|\text{Id}_N - \tau A\|_2 = \sigma_{\max}(\text{Id}_N - \tau) = \max(|1 - \tau \sigma_{\max}(A)|, |1 - \tau \sigma_{\min}(A)|).$$

For a quadratic function, one has $\sigma_{\min}(A) = \mu, \sigma_{\max}(A) = L$. Figure 13.1, right, shows a display of $h(\tau)$. One has that for $0 < \tau < 2/L$, $h(\tau) < 1$. \square

Note that when the condition number $\varepsilon \stackrel{\text{def.}}{=} \mu/L \ll 1$ is small (which is the typical setup for ill-posed problems), then the contraction constant appearing in (13.7) scales like

$$\tilde{\rho} \sim 1 - 2\varepsilon. \quad (13.8)$$

The quantity ε in some sense reflects the inverse-conditioning of the problem. For quadratic function, it indeed corresponds exactly to the inverse of the condition number (which is the ratio of the largest to smallest singular value). The condition number is minimum and equal to 1 for orthogonal matrices.

The error decay rate (13.6), although it is geometrical $O(\rho^\ell)$ is called a “linear rate” in the optimization literature. It is a “global” rate because it hold for all ℓ (and not only for large enough ℓ).

We now give convergence theorem for a general convex function. On contrast to quadratic function, if one does not assumes strong convexity, one can only show a sub-linear rate on the function values (and no rate at all on the iterates themselves!). It is only when one assume strong convexity that linear rate is obtained. Note that in this case, the solution of the minimization problem is not necessarily unique.

Theorem 35. *If f satisfy conditions (\mathcal{R}_L) , assuming there exists $(\tau_{\min}, \tau_{\max})$ such that*

$$0 < \tau_{\min} \leq \tau_\ell \leq \tau_{\max} < \frac{2}{L}, \quad (13.9)$$

then $x^{(\ell)}$ converges to a solution x^ of (13.1) and there exists $C > 0$ such that*

$$f(x^{(\ell)}) - f(x^*) \leq \frac{C}{\ell + 1}. \quad (13.10)$$

If furthermore f is μ -strongly convex, then there exists $0 \leq \rho < 1$ such that $\|x^{(\ell)} - x^\| \leq \rho^\ell \|x^{(0)} - x^*\|$.*

Proof. In the case where f is not strongly convex, we only prove (13.10) since the proof that $x^{(\ell)}$ converges is more technical. Note indeed that if the minimizer x^* is non-unique, then it might be the case that the iterate $x^{(\ell)}$ “cycle” while approaching the set of minimizer, but actually convexity of f prevents this kind of pathological behavior. For simplicity, we do the proof in the case $\tau_\ell = 1/L$, but it extends to the general case. The L -smoothness property imply (13.3), which reads

$$f(x^{(\ell+1)}) \leq f(x^{(\ell)}) + \langle \nabla f(x^{(\ell)}), x^{(\ell+1)} - x^{(\ell)} \rangle + \frac{L}{2} \|x^{(\ell+1)} - x^{(\ell)}\|^2.$$

Using the fact that $x^{(\ell+1)} - x^{(\ell)} = -\frac{1}{L} \nabla f(x^{(\ell)})$, one obtains

$$f(x^{(\ell+1)}) \leq f(x^{(\ell)}) - \frac{1}{L} \|\nabla f(x^{(\ell)})\|^2 + \frac{1}{2L} \|\nabla f(x^{(\ell)})\|^2 \leq f(x^{(\ell)}) - \frac{1}{2L} \|\nabla f(x^{(\ell)})\|^2 \quad (13.11)$$

This shows that $(f(x^{(\ell)}))_\ell$ is a decaying sequence. By convexity

$$f(x^{(\ell)}) + \langle \nabla f(x^{(\ell)}), x^* - x^{(\ell)} \rangle \leq f(x^*)$$

and plugging this in (13.11) shows

$$f(x^{(\ell+1)}) \leq f(x^*) - \langle \nabla f(x^{(\ell)}), x^* - x^{(\ell)} \rangle - \frac{1}{2L} \|\nabla f(x^{(\ell)})\|^2 \quad (13.12)$$

$$= f(x^*) + \frac{L}{2} \left(\|x^{(\ell)} - x^*\|^2 - \|x^{(\ell)} - x^* - \frac{1}{L} \nabla f(x^{(\ell)})\|^2 \right) \quad (13.13)$$

$$= f(x^*) + \frac{L}{2} \left(\|x^{(\ell)} - x^*\|^2 - \|x^* - x^{(\ell+1)}\|^2 \right). \quad (13.14)$$

Summing these inequalities for $\ell = 0, \dots, k$, one obtains

$$\sum_{\ell=0}^k f(x^{(\ell+1)}) - (k+1)f(x^*) \leq \frac{L}{2} \left(\|x^{(0)} - x^*\|^2 - \|x^{(k+1)} - x^*\|^2 \right)$$

and since $f(x^{(\ell+1)})$ is decaying $\sum_{\ell=0}^k f(x^{(\ell+1)}) \geq (k+1)f(x^{(k+1)})$, thus

$$f(x^{(k+1)}) - f(x^*) \leq \frac{L\|x^{(0)} - x^*\|^2}{2(k+1)}$$

which gives (13.10) for $C \stackrel{\text{def.}}{=} L\|x^{(0)} - x^*\|^2/2$.

If we now assume f is μ -strongly convex, then, using $\nabla f(x^*) = 0$, one has $\frac{\mu}{2} \|x^* - x\|^2 \leq f(x) - f(x^*)$ for all x . Re-manipulating (13.14) gives

$$\frac{\mu}{2} \|x^{(\ell+1)} - x^*\|^2 \leq f(x^{(\ell+1)}) - f(x^*) \leq \frac{L}{2} \left(\|x^{(\ell)} - x^*\|^2 - \|x^* - x^{(\ell+1)}\|^2 \right),$$

and hence

$$\|x^{(\ell+1)} - x^*\| \leq \sqrt{\frac{L}{L+\mu}} \|x^{(\ell+1)} - x^*\|, \quad (13.15)$$

which is the desired result. \square

Note that in the low conditioning setting $\varepsilon \ll 1$, one retrieve a dependency of the rate (13.15) similar to the one of quadratic functions (13.8), indeed

$$\sqrt{\frac{L}{L+\mu}} = (1+\varepsilon)^{-\frac{1}{2}} \sim 1 - \frac{1}{2}\varepsilon.$$

13.1.2 Sub-gradient Descent

The gradient descent (13.2) cannot be applied on a non-smooth function f . One can use in place of a gradient a sub-gradient, which defines the sub-gradient descent

$$x^{(\ell+1)} \stackrel{\text{def}}{=} x^{(\ell)} - \tau_\ell g^{(\ell)} \quad \text{where} \quad g^{(\ell)} \in \partial f(x^{(\ell)}). \quad (13.16)$$

The main issue with this scheme is that to ensure convergence, the iterate should go to zero. One can easily convince oneself why by looking at the iterates on a function $f(x) = |x|$.

Theorem 36. *If $\sum_\ell \tau_\ell = +\infty$ and $\sum_\ell \tau_\ell^2 < +\infty$, then $x^{(\ell)}$ converges to a minimizer of f .*

13.1.3 Projected Gradient Descent

We consider a generic constraint optimization problem as

$$\min_{x \in \mathcal{C}} f(x) \quad (13.17)$$

where $\mathcal{C} \subset \mathbb{R}^S$ is a closed convex set and $f : \mathbb{R}^S \rightarrow \mathbb{R}$ is a smooth convex function (at least of class \mathcal{C}^1).

The gradient descent algorithm (13.2) is generalized to solve a constrained problem using the projected gradient descent

$$x^{(\ell+1)} \stackrel{\text{def}}{=} \text{Proj}_{\mathcal{C}} \left(x^{(\ell)} - \tau_\ell \nabla f(x^{(\ell)}) \right), \quad (13.18)$$

where $\text{Proj}_{\mathcal{C}}$ is the orthogonal projector on \mathcal{C}

$$\text{Proj}_{\mathcal{C}}(x) = \operatorname{argmin}_{x' \in \mathcal{C}} \|x - x'\|$$

which is always uniquely defined because \mathcal{C} is closed and convex. The following proposition shows that all the convergence properties of the classical gradient descent carries over to this projected algorithm.

Theorem 37. *Theorems ?? and 35 still holds when replacing iterations (13.2) by (13.18).*

Proof. The proof of Theorem ?? extends because the projector is contractant, $\|\text{Proj}_{\mathcal{C}}(x) - \text{Proj}_{\mathcal{C}}(x')\| \leq \|x - x'\|$ so that the strict contraction properties of the gradient descent is maintained by this projection. \square

The main bottleneck that often prevents to use (13.18) is that the projector is often complicated to compute. We are however lucky since for ℓ^1 minimization, one can apply in a straightforward manner this method.

13.2 Proximal Algorithm

For non-smooth functions f , it is not possible to perform an “explicit” gradient descent step because the gradient is not even defined. One thus needs to replace this “explicit” step by an “implicit” one, which is possible even if f is non-smooth.

13.2.1 Proximal Map

The implicit stepping of amplitude $\tau > 0$ is defined as

$$\forall x, \quad \text{Prox}_{\tau f}(x) \stackrel{\text{def}}{=} \operatorname{argmin}_{x'} \frac{1}{2} \|x - x'\|^2 + f(x'). \quad (13.19)$$

It amounts to minimize function f locally around x , in a ball of radius controlled by τ . This the involved function $\frac{1}{2} \|x - \cdot\|^2 + f$ is strongly convex, this operator $\text{Prox}_{\tau f}$ is well defined and single-valued.

When $f = \iota_{\mathcal{C}}$ is an indicator, the proximal map boils down to a projection $\text{Prox}_{\iota_{\mathcal{C}}} = \text{Proj}_{\mathcal{C}}$, it is thus in some sense a generalization of the projection to arbitrary function. And can also be interpreted as a projector on a level set of f . An interesting feature of the proximal map is that it is a contraction, thus generalizing the well-known property of projectors.

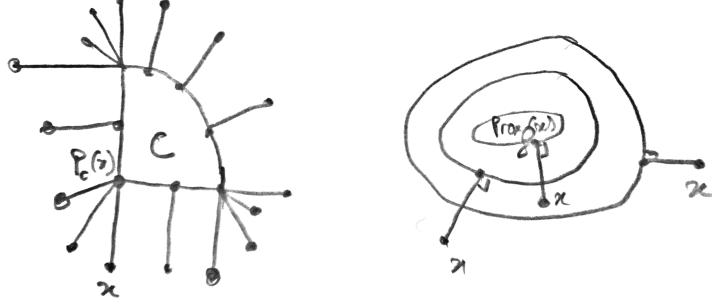


Figure 13.2: Proximal map and projection map.

Proposition 35. One has $\|\text{prox}_f(x) - \text{prox}_f(y)\| \leq \|x - y\|$.

Examples The following proposition states a few simple examples.

Proposition 36. One has

$$\text{Prox}_{\frac{\tau}{2}\|\cdot\|^2}(x) = \frac{x}{1 + \tau}, \quad \text{and} \quad \text{Prox}_{\tau\|\cdot\|_1} = \mathcal{S}_\tau^1(x), \quad (13.20)$$

where the soft-thresholding is defined as

$$\mathcal{S}_\tau^1(x) \stackrel{\text{def.}}{=} (S_\tau(x_i))_{i=1}^N \quad \text{where} \quad S_\tau(r) \stackrel{\text{def.}}{=} \text{sign}(r)(|r| - \lambda)_+,$$

(see also (11.5)). For $A \in \mathbb{R}^{P \times N}$, one has

$$\text{Prox}_{\frac{\tau}{2}\|A \cdot - y\|^2}(x) = (\text{Id}_N + \tau A^* A)^{-1}(x + \tau A^* y). \quad (13.21)$$

Proof. The proximal map of $\|\cdot\|_1$ was derived in Proposition 24. For the quadratic case

$$z = \text{Prox}_{\frac{\tau}{2}\|A \cdot - y\|^2}(x) \Leftrightarrow z - x + \tau A^*(Az - y) = 0 \Leftrightarrow (\text{Id}_N + \tau A^* A)z = x + \tau A^* y.$$

□

Note that in some cases, the proximal map of a non-convex function is well defined, for instance $\text{Prox}_{\tau\|\cdot\|_0}$ is the hard thresholding associated to the threshold $\sqrt{2\tau}$, see Proposition 24.

13.2.2 Basic Properties

We recap some useful proximal-calculus.

Proposition 37. One has

$$\text{Prox}_{f+\langle y, \cdot \rangle} = y + \text{Prox}_f, \quad \text{Prox}_{f(\cdot - y)} = y + \text{Prox}_f(\cdot - y).$$

If $f(x) = \sum_{k=1}^K f_k(x_k)$ for $x = (x_1, \dots, x_K)$ is separable, then

$$\text{Prox}_{\tau f}(x) = (\text{Prox}_{\tau f_k}(x_k))_{k=1}^K. \quad (13.22)$$

Proof. One has

$$z = \text{Prox}_{f+\langle y, \cdot \rangle}(x) \Leftrightarrow 0 \in x - z + (\partial f(x) + y) \Leftrightarrow 0 \in x - (z - y) + \partial f(x)$$

which is the optimality condition for $z - y = \text{Prox}_f(x)$.

One has

$$z = \text{Prox}_{f(\cdot-y)}(x) \Leftrightarrow 0 \in x - z + \lambda \partial f(x - y) \Leftrightarrow 0 \in x' - (z - y) + \partial f(x')$$

where we defined $x' \stackrel{\text{def.}}{=} x - y$, and this is the optimality condition for $z - y = \text{Prox}_f(x')$ \square

The following proposition is very useful.

Proposition 38. *If $A \in \mathbb{R}^{P \times N}$ is a tight frame, i.e. $AA^* = \text{Id}_P$, then*

$$\text{Prox}_{f \circ A} = A^* \circ \text{Prox}_f \circ A + \text{Id}_N - A^* A.$$

In particular, if A is orthogonal, then $\text{Prox}_{f \circ A} = A^* \circ \text{Prox}_f \circ A$.

13.2.3 Related Concepts

Link with sub-differential. For a set-valued map $U : \mathcal{H} \rightarrow \mathcal{G}$, we define the inverse set-valued map $U^{-1} : \mathcal{G} \rightarrow \mathcal{H}$ by

$$h \in U^{-1}(g) \iff g \in U(h) \quad (13.23)$$

[ToDo: add picture] The following proposition shows that the proximal map is related to a regularized inverse of the sub-differential.

Proposition 39. *One has $\text{Prox}_{\tau f} = (\text{Id} + \tau \partial f)^{-1}$.*

Proof. One has the following equivalence

$$z = \text{Prox}_{\tau f}(x) \Leftrightarrow 0 \in z - x + \tau \partial f(z) \Leftrightarrow x \in (\text{Id} + \tau \partial f)(z) \Leftrightarrow z = (\text{Id} + \tau \partial f)^{-1}(x)$$

where for the last equivalence, we have replace “ \in ” by “ $=$ ” because the proximal map is single valued. \square

The proximal operator is hence often referred to the “resolvent” $\text{Prox}_{\tau f} = (\text{Id} + \tau \partial f)^{-1}$ of the maximal monotone operator ∂f .

Link with duality. One has the following fundamental relation between the proximal operator of a function and of its Legendre-Fenchel transform

Theorem 38 (Moreau decomposition). *One has*

$$\text{Prox}_{\tau f} = \text{Id} - \tau \text{Prox}_{f^*/\tau}(\cdot/\tau).$$

This theorem shows that the proximal operator of f is simple to compute if and only the proximal operator of f^* is also simple. As a particular instantiation, since according to , one can re-write the soft thresholding as follow

$$\text{Prox}_{\tau \|\cdot\|_1}(x) = x - \tau \text{Proj}_{\|\cdot\|_\infty \leqslant 1}(x/\tau) = x - \text{Proj}_{\|\cdot\|_\infty \leqslant \tau}(x) \quad \text{where} \quad \text{Proj}_{\|\cdot\|_\infty \leqslant \tau}(x) = \min(\max(x, -\tau), \tau).$$

In the special case where $f = \iota_{\mathcal{C}}$ where \mathcal{C} is a closed convex cone, then

$$(\iota_{\mathcal{C}})^* = \iota_{\mathcal{C}^\circ} \quad \text{where} \quad \mathcal{C}^\circ \stackrel{\text{def.}}{=} \{y ; \forall x \in \mathcal{C}, \langle x, y \rangle \leqslant 0\} \quad (13.24)$$

and \mathcal{C}° is the so-called polar cone. Cones are fundament object in convex optimization because they are invariant by duality, in the sense of (13.24) (if \mathcal{C} is not a cone, its Legendre transform would not be an indicator). Using (13.24), one obtains the celebrated Moreau polar decomposition

$$x = \text{Proj}_{\mathcal{C}}(x) +^\perp \text{Proj}_{\mathcal{C}^\circ}(x)$$

where “ $+^\perp$ ” denotes an orthogonal sum (the terms in the sum are mutually orthogonal). [ToDo: add drawing] In the case where $\mathcal{C} = V$ is a linear space, this corresponds to the usual decomposition $\mathbb{R}^N = V \oplus^\perp V^\perp$.

Link with Moreau-Yosida regularization. The following proposition shows that the proximal operator can be interpreted as performing a gradient descent step on the Moreau-Yosida smoothed version f_μ of f , defined in (12.11).

Proposition 40. *One has*

$$\text{Prox}_{\mu f} = \text{Id} - \mu \nabla f_\mu.$$

13.3 Primal Algorithms

We now describe some important algorithm which assumes some structure (a so-called “splitting”) of the minimized functional to be able to apply proximal maps on sub-functions. Note that there is obviously many ways to split or structure a given initial problem, so there are many non-equivalent ways to apply a given proximal-based method to solve the problem. Finding the “best” way to split a problem is a bit like black magic, and there is no definite answer. Also all there algorithm comes with step size and related parameters, and there is no obvious way to tune these parameters automatically (although some insight might be gained by studying convergence rate).

13.3.1 Proximal Point Algorithm

One has the following equivalence

$$x^* \in \operatorname{argmin} f \Leftrightarrow 0 \in \partial f(x^*) \Leftrightarrow x^* \in (\text{Id} + \tau \partial f)(x^*) \quad (13.25)$$

$$\Leftrightarrow x^* = (\text{Id} + \tau \partial f)^{-1}(x^*) = \text{Prox}_{\tau f}(x^*). \quad (13.26)$$

This shows that being a minimizer of f is equivalent to being a fixed point of $\text{Prox}_{\tau f}$. This suggest the following fixed point iterations, which are called the proximal point algorithm

$$x^{(\ell+1)} \stackrel{\text{def.}}{=} \text{Prox}_{\tau_\ell f}(x^{(\ell)}). \quad (13.27)$$

On contrast to the gradient descent fixed point scheme, the proximal point method is converging for any sequence of steps.

Theorem 39. *If $0 < \tau_{\min} \leq \tau_\ell \leq \gamma_{\max} < +\infty$, then $x^{(\ell)} \rightarrow x^*$ a minimizer of f .*

This implicit step (13.27) should be compared with a gradient descent step (13.2)

$$x^{(\ell+1)} \stackrel{\text{def.}}{=} (\text{Id} + \tau_\ell \nabla f)(x^{(\ell)}).$$

One sees that the implicit resolvent $(\text{Id} - \tau_\ell \partial f)^{-1}$ replaces the explicit step $\text{Id} + \tau_\ell \nabla f$. For small τ_ℓ and smooth f , they are equivalent at first order. But the implicit step is well defined even for non-smooth function, and the scheme (the proximal point) is always convergent (whereas the explicit step size should be small enough for the gradient descent to converge). This is inline with the general idea the implicit stepping (e.g. implicit Euler for integrating ODE, which is very similar to the proximal point method) is more stable. Of course, the drawback is that explicit step are very easy to implement whereas in general proximal map are hard to solve (most of the time as hard as solving the initial problem).

13.3.2 Forward-Backward

It is in general impossible to compute $\text{Prox}_{\gamma f}$ so that the proximal point algorithm is not implementable. In oder to derive more practical algorithms, it is important to restrict the class of considered function, by imposing some structure on the function to be minimized. We consider functions of the form

$$\min_x \mathcal{E}(x) \stackrel{\text{def.}}{=} f(x) + g(x) \quad (13.28)$$

where $g \in \Gamma_0(\mathcal{H})$ can be an arbitrary, but f needs to be smooth.

One can modify the fixe point derivation (13.25) to account for this special structure

$$\begin{aligned} x^* \in \operatorname{argmin} f + g &\Leftrightarrow 0 \in \nabla f(x^*) + \partial g(x^*) \Leftrightarrow x^* - \tau \nabla f(x^*) \in (\operatorname{Id} + \tau \partial g)(x^*) \\ &\Leftrightarrow x^* = (\operatorname{Id} + \tau \partial g)^{-1} \circ (\operatorname{Id} - \tau \nabla f)(x^*). \end{aligned}$$

This fixed point suggests the following algorithm, with the celebrated Forward-Backward

$$x^{(\ell+1)} \stackrel{\text{def.}}{=} \operatorname{Prox}_{\tau_\ell g} \left(x^{(\ell)} - \tau_\ell \nabla f(x^{(\ell)}) \right). \quad (13.29)$$

Derivation using surrogate functionals. An intuitive way to derive this algorithm, and also a way to prove its convergence, it using the concept of surrogate functional.

To derive an iterative algorithm, we modify the energy $\mathcal{E}(x)$ to obtain a surrogate functional $\mathcal{E}(x, x^{(\ell)})$ whose minimization corresponds to a simpler optimization problem, and define the iterations as

$$x^{(\ell+1)} \stackrel{\text{def.}}{=} \operatorname{argmin}_x \mathcal{E}(x, x^{(\ell)}). \quad (13.30)$$

In order to ensure convergence, this function should satisfy the following property

$$\mathcal{E}(x) \leq \mathcal{E}(x, x') \quad \text{and} \quad \mathcal{E}(x, x) = \mathcal{E}(x) \quad (13.31)$$

and $\mathcal{E}(x) - \mathcal{E}(x, x')$ should be a smooth function. Property (13.31) guarantees that f is decaying by the iterations

$$\mathcal{E}(x^{(\ell+1)}) \leq \mathcal{E}(x^{(\ell)})$$

and it simple to check that actually all accumulation points of $(x^{(\ell)})_\ell$ are stationary points of f .

In order to derive a valid surrogate $\mathcal{E}(x, x')$ for our functional (13.28), since we assume f is L -smooth (i.e. satisfies (\mathcal{R}_L)), let us recall the quadratic majorant (13.3)

$$f(x) \leq f(x') + \langle \nabla f(x'), x' - x \rangle + \frac{L}{2} \|x - x'\|^2,$$

so that for $0 < \tau < \frac{1}{L}$, the function

$$\mathcal{E}(x, x') \stackrel{\text{def.}}{=} f(x') + \langle \nabla f(x'), x' - x \rangle + \frac{1}{2\tau} \|x - x'\|^2 + g(x) \quad (13.32)$$

satisfies the surrogate conditions (13.31). The following proposition shows that minimizing the surrogate functional corresponds to the computation of a so-called proximal operator.

Proposition 41. *The update (13.30) for the surrogate (13.32) is exactly (13.29).*

Proof. This follows from the fact that

$$\langle \nabla f(x'), x' - x \rangle + \frac{1}{2\tau} \|x - x'\|^2 = \frac{1}{2\tau} \|x - (x' - \tau \nabla f(x'))\|^2 + \text{cst.}$$

□

Convergence of FB. Although we impose $\tau < 1/L$ to ensure majorization property, one can actually show convergence under the same hypothesis as for the gradient descent, i.e. $0 < \tau < 2/L$, with the same convergence rates. This means that Theorem 37 for the projected gradient descent extend to FB.

Theorem 40. *Theorems ?? and 35 still holds when replacing iterations (13.2) by (13.29).*

Note furthermore that the projected gradient descent algorithm (13.18) is recovered as a special case of (13.29) when setting $J = \iota_C$ the indicator of the constraint set, since $\operatorname{Prox}_{\rho J} = \operatorname{Proj}_C$ in this case.

Of course the difficult point is to be able to compute in closed form $\operatorname{Prox}_{\tau g}$ in (13.29), and this is usually possible only for very simple function. We have already seen such an example in Section 11.3.3 for the resolution of ℓ^1 -regularized inverse problems (the Lasso).

13.3.3 Douglas-Rachford

We consider here the structured minimization problem

$$\min_{x \in \mathbb{R}^N} f(x) + g(x), \quad (13.33)$$

but on contrary to the Forward-Backward setting studied in Section 13.3.2, no smoothness is imposed on f . We here suppose that we can compute easily the proximal map of f and g .

Example 7 (Constrained Lasso). An example of a problem of the form (13.33) where one can apply Douglas-Rachford is the noiseless constrained Lasso problem (11.11)

$$\min_{Ax=y} \|x\|_1$$

where one can use $f = \iota_{\mathcal{C}_y}$ where $\mathcal{C}_y \stackrel{\text{def.}}{=} \{x ; Ax = y\}$ and $g = \|\cdot\|_1$. As noted in Section 11.3.1, this problem is equivalent to a linear program. The proximal operator of g is the soft thresholding as stated in (13.20), while the proximal operator of g is the orthogonal projector on the affine space \mathcal{C}_y , which can be computed by solving a linear system as stated in (12.9) (this is especially convenient for inpainting problems or deconvolution problem where this is achieved efficiently).

The Douglas-Rachford iterations read

$$\tilde{x}^{(\ell+1)} \stackrel{\text{def.}}{=} \left(1 - \frac{\mu}{2}\right) \tilde{x}^{(\ell)} + \frac{\mu}{2} \text{rProx}_{\tau g}(\text{rProx}_{\tau f}(\tilde{x}^{(\ell)})) \quad \text{and} \quad x^{(\ell+1)} \stackrel{\text{def.}}{=} \text{Prox}_{\tau f}(\tilde{x}^{(\ell+1)}), \quad (13.34)$$

where we have used the following shortcuts

$$\text{rProx}_{\tau f}(x) = 2 \text{Prox}_{\tau f}(x) - x.$$

One can show that for any value of $\tau > 0$, any $0 < \mu < 2$, and any \tilde{x}_0 , $x^{(\ell)} \rightarrow x^*$ which is a minimizer of $f + g$.

Note that it is of course possible to inter-change the roles of f and g , which defines another set of iterations.

More than two functions. Another sets of iterations can be obtained by “symetrizing” the algorithm. More generally, if we have K functions $(f_k)_k$, we re-write

$$\min_x \sum_k f_k(x) = \min_{X=(x_1, \dots, x_k)} f(X) + g(X) \quad \text{where} \quad f(X) = \sum_k f_k(x_k) \quad \text{and} \quad g(X) = \iota_{\Delta}(X)$$

where $\Delta = \{X ; x_1 = \dots = x_k\}$ is the diagonal. The proximal operator of f is

$$\text{Prox}_{\tau f}(X) = \text{Proj}_{\Delta}(X) = (\bar{x}, \dots, \bar{x}) \quad \text{where} \quad \bar{x} = \frac{1}{K} \sum_k x_k$$

while the proximal operator of f is easily computed from those of the $(f_k)_k$ using (13.22). One can thus apply DR iterations (13.34).

Handling a linear operator. One can handle a minimization of the form (13.36) by introducing extra variables

$$\inf_x f_1(x) + f_2(Ax) = \inf_{z=(x,y)} f(z) + g(z) \quad \text{where} \quad \begin{cases} f(z) = f_1(x) + f_2(y) \\ g(z) = \iota_{\mathcal{C}}(x, y), \end{cases}$$

where $\mathcal{C} = \{(x, y) ; Ax = y\}$. This problem can be handled using DR iterations (13.34), since the proximal operator of f is obtained from those of (f_1, f_2) using (13.22), while the proximal operator of g is the projector on \mathcal{C} , which can be computed in two alternative ways as the following proposition shows.

Proposition 42. One has

$$\text{Proj}_{\mathcal{C}}(x, y) = (x + A^* \tilde{y}, y - \tilde{y}) = (\tilde{x}, A\tilde{x}) \quad \text{where} \quad \begin{cases} \tilde{y} \stackrel{\text{def.}}{=} (\text{Id}_P + AA^*)^{-1}(Ax - y) \\ \tilde{x} \stackrel{\text{def.}}{=} (\text{Id}_N + A^*A)^{-1}(A^*y + x). \end{cases} \quad (13.35)$$

Proof. [ToDo: todo] □

Remark 7 (Inversion of linear operator). At many places (typically to compute some sort of projector) one has to invert matrices of the form AA^* , A^*A , $\text{Id}_P + AA^*$ or $\text{Id}_N + A^*A$ (see for instance (13.35)). There are some cases where this can be done efficiently. Typical examples where this is simple are inpainting/inverse problem where AA^* is diagonal, and deconvolution or partial Fourier measurement (e.g. fMRI) for which A^*A is diagonalized using the FFT. If this inversion is too costly, one needs to use more advanced methods, based on duality, which allows to avoid trading the inverse A by the application of A^* . They are however typically converging more slowly.

13.4 Dual and Primal-Dual Algorithms

Convex duality, detailed in Section 12.2 (either from the Lagrange or the Fenchel-Rockafellar point of view – which are essentially equivalent), is very fruitful to derive new optimization algorithm or to apply existing algorithm on a dual reformulation.

13.4.1 Forward-backward on the Dual

Let us illustrate first the idea of applying a known algorithm to the dual problem. We consider here the structured minimization problem associated to Fenchel-Rockafellar duality (12.12)

$$p^* = \inf_x f(x) + g(Ax), \quad (13.36)$$

but furthermore assume that f is μ -strongly convex, and we assume for simplicity that both (f, g) are continuous. If f were also smooth (but it needs to be!), one could think about using the Forward-Backward algorithm (13.29). But the main issue is that in general $\text{Prox}_{\tau g \circ A}$ cannot be computed easily even if one can compute $\text{Prox}_{\tau g \circ A}$. An exception to this is when A is a tight frame, as exposed in Proposition 38, but in practice it is rarely the case.

Example 8 (TV denoising). A typical example, which was the one used by Antonin Chambolle [8] to develop this class of method, is the total variation denoising

$$\min_x \frac{1}{2} \|y - x\|^2 + \lambda \|\nabla x\|_{1,2}$$

where $\nabla x \in \mathbb{R}^{N \times d}$ is the gradient (a vector field) of a signal ($d = 1$) or image ($d = 2$) x , and $\|\cdot\|_{1,2}$ is the vectorial- ℓ^1 norm (also called $\ell^1 - \ell^2$ norm), such that for a d -dimensional vector field $(v_i)_{i=1}^N$

$$\|v\|_{1,2} \stackrel{\text{def.}}{=} \sum_i \|v_i\|.$$

Here

$$f = \frac{1}{2} \|\cdot - y\|^2 \quad \text{and} \quad g = \lambda \|\cdot\|_{1,2}$$

so that f is $\mu = 1$ strongly convex, and one sets $A = \nabla$ the linear operator.

Applying Fenchel-Rockafellar Theorem 34 (since strong duality holds, all involved functions being continuous), one has that

$$p^* = \sup_u -g^*(u) - f^*(-A^*u).$$

But more importantly, since f is μ -strongly convex, one has that f^* is smooth with a $1/\mu$ -Lipschitz gradient. One can thus use the Forward-Backward algorithm (13.29) on (minus the energy of) this problem, which reads

$$u^{(\ell+1)} = \text{Prox}_{\tau_k g^*} \left(u^{(\ell)} + \tau_k A \nabla f^*(A^* u^{(\ell)}) \right).$$

To guarantee convergence, the step size τ_k should be smaller than $2/L$ where L is the Lipschitz constant of $A \circ \nabla f^* \circ A^*$, which is smaller than $\|A\|^2/\mu$.

Last but not least, one some (not necessarily unique) dual minimizer u^* is computed, the primal-dual relationships (12.17) ensures that one retrieves the unique primal minimizer x^* as

$$-A^* u^* \in \partial f(x^*) \Leftrightarrow x^* \in (\partial f)^{-1}(-A^* u^*) = \partial f^*(-A^* u^*) \Leftrightarrow x^* = \nabla f^*(-A^* u^*)$$

where we used here the crucial fact that f^* is smooth.

Example 9 (TV denoising). In the particular case of the TV denoising problem, one has

$$\begin{aligned} g^* &= \iota_{\|\cdot\|_{\infty,2} \leq \lambda} \quad \text{where} \quad \|v\|_{\infty,2} \stackrel{\text{def.}}{=} \max_i \|v_i\| \implies \text{Prox}_{\tau g^*}(u) = \left(\min(\|v_i\|, \lambda) \frac{v_i}{\|v_i\|} \right) \\ f^*(h) &= \frac{1}{2} \|h\|^2 + \langle h, y \rangle \quad \text{and} \quad \nabla f^*(h) = h + y. \end{aligned}$$

Furthermore, $\mu = 1$ and $A^* A = \Delta$ is the usual finite difference approximation of the Laplacian, so that $\|A\|^2 = \|\Delta\| = 4d$ where d is the dimension.

13.4.2 Primal-Dual Splitting

We now comeback to the more general structure problem of the form (13.36), which we consider in primal-dual form as

$$\inf_x f(x) + g(Ax) = \sup_u \inf_x f(x) + \langle Ax, u \rangle - g^*(u), \quad (13.37)$$

but we do not suppose anymore that f is strongly convex.

A typical instance of such a problem is for the TV regularization of the inverse problem $\mathcal{K}x = y$, which corresponds to solving

$$\min_x \frac{1}{2} \|y - \mathcal{K}x\|^2 + \lambda \|\nabla x\|_{1,2}.$$

where $A = \nabla$, $f(x) = \frac{1}{2} \|y - \mathcal{K}x\|^2$ and $g = \lambda \|\cdot\|_{1,2}$. Note however that with such a splitting, one will have to compute the proximal operator of f , which, following (13.21), requires inverting either $\text{Id}_P + AA^*$ or $\text{Id}_N + A^*A$, see Remark 7.

A standard primal-dual algorithm, which is detailed in [], reads

$$\begin{aligned} z^{(\ell+1)} &\stackrel{\text{def.}}{=} \text{Prox}_{\sigma g^*}(z^{(\ell)} + \sigma A(\tilde{x}^{(\ell)})) \\ x^{(\ell+1)} &\stackrel{\text{def.}}{=} \text{Prox}_{\tau f}(x^{(\ell)} - \tau A^*(z^{(\ell+1)})) \\ \tilde{x}^{(\ell)} &\stackrel{\text{def.}}{=} x^{(\ell+1)} + \theta(x^{(\ell+1)} - x^{(\ell)}) \end{aligned}$$

if $0 \leq \theta \leq 1$ and $\sigma\tau\|\mathcal{K}\|^2 < 1$, then $x^{(\ell)}$ converges to a minimizer of (13.37).

Chapter 14

Theory of Sparse Regularization

We now apply the basics elements of convex analysis from the previous chapter to perform a theoretical analysis of the properties of the Lasso, in particular its performances to recover sparse vectors.

14.1 Existence and Uniqueness

14.1.1 Existence

We consider problems (11.10) and (11.11), that we rewrite here as

$$\min_{x \in \mathbb{R}^N} f_\lambda(x) \stackrel{\text{def.}}{=} \frac{1}{2\lambda} \|y - Ax\|^2 + \lambda \|x\|_1 \quad (\mathcal{P}_\lambda(y))$$

and its limit as $\lambda \rightarrow 0$

$$\min_{Ax=y} \|x\|_1 = \min_x f_0(x) \stackrel{\text{def.}}{=} \iota_{\mathcal{L}_y}(x) + \|x\|_1. \quad (\mathcal{P}_0(y))$$

where $A \in \mathbb{R}^{P \times N}$, and $\mathcal{L}_y \stackrel{\text{def.}}{=} \{x \in \mathbb{R}^N ; Ax = y\}$.

We recall that the setup is that one observe noise measures

$$y = Ax_0 + w$$

and we would like conditions to ensure for x_0 to solution to $(\mathcal{P}_0(Ax_0))$ (i.e. when $w = 0$) and to be close (in some sense to be defined, and in some proportion to the noise level $\|w\|$) to the solutions of $(\mathcal{P}_0(y = Ax_0 + w))$ when λ is wisely chosen as a function of $\|w\|$.

First let us note that since $(\mathcal{P}_\lambda(y))$ is unconstrained and coercive (because $\|\cdot\|_1$ is), this problem always has solutions. Since A might have a kernel and $\|\cdot\|_1$ is not strongly convex, it might have non-unique solutions. If $y \in \text{Im}(A)$, the constraint set of $(\mathcal{P}_0(y))$ is non-empty, and it also has solutions, which might fail to be unique.

Figure 14.1 gives the intuition of the theory that will be developed in this chapter, regarding the exact or approximated recovery of sparse vectors x_0 , and the need for a careful selection of the λ parameter.

14.1.2 Polytope Projection for the Constraint Problem

The following proposition gives a geometric description of those vectors which are recovered by ℓ^1 minimization when there is no noise.

Proposition 43. *We denote $B \stackrel{\text{def.}}{=} \{x \in \mathbb{R}^N ; \|x\|_1 \leq 1\}$. Then, assuming $\text{Im}(A) = \mathbb{R}^P$,*

$$x_0 \text{ is a solution to } \mathcal{P}_0(Ax_0) \iff A \frac{x_0}{\|x_0\|_1} \in \partial(AB) \quad (14.1)$$

where “ ∂ ” denoted the boundary and $AB = \{Ax ; x \in B\}$.

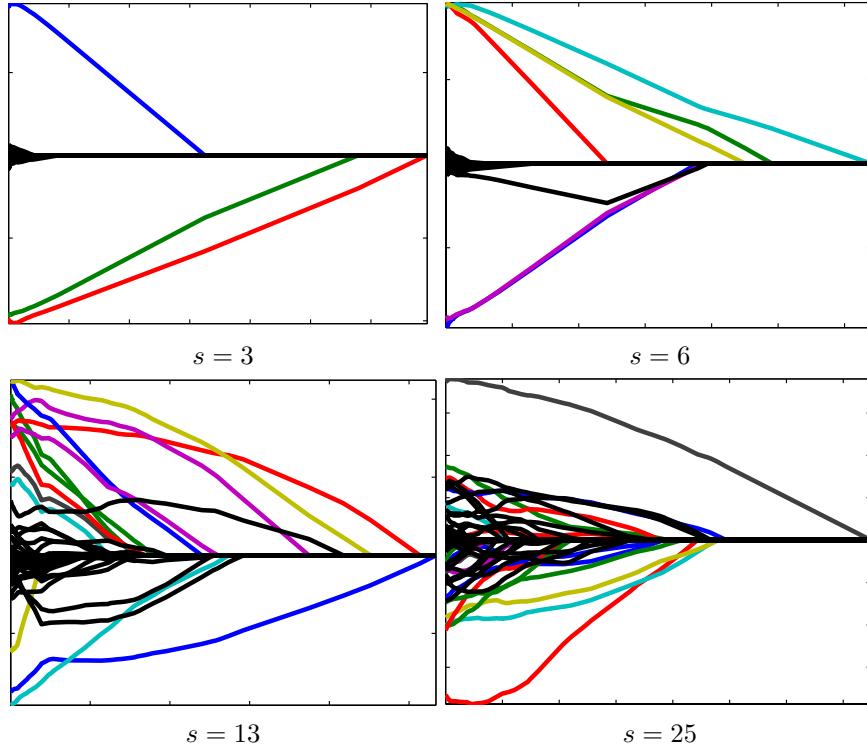


Figure 14.1: Display of the evolution $\lambda \mapsto x_\lambda$ of the solutions of $(\mathcal{P}_\lambda(y))$.

Proof. We first prove “ \Rightarrow ”. We suppose that x_0 is not a solution, and aim at showing that $A \frac{x_0}{\|x_0\|_1} \in \text{int}(AB_\rho)$. Since it is not a solution, there exists z such that $Ax_0 = Az$ and $\|z\|_1 = (1 - \delta)\|x_0\|_1$ with $\delta > 0$. Then for any displacement $h = A\varepsilon \in \text{Im}(A)$, where one can impose $\varepsilon \in \ker(A)^\perp$, i.e. $\varepsilon = A^+h$, one has $Ax_0 + h = A(z + \varepsilon)$ and

$$\|z + \varepsilon\|_1 \leq \|z\|_1 + \|\Phi^+h\| \leq (1 - \delta)\|x_0\|_1 + \|\Phi^+\|_{1,1}\|h\|_1 < \frac{\delta}{\|A^+\|_{1,1}}\|x_0\|_1.$$

This means that choosing $\|h\|_1 < \frac{\delta}{\|A^+\|_{1,1}}\|x_0\|_1$ implies that $A \frac{x_0}{\|x_0\|_1} \in \text{int}(AB)$.

We now prove “ \Leftarrow ”. We suppose that $A \frac{x_0}{\|x_0\|_1} \in \text{int}(AB)$. Then there exists z such that $Ax_0 = (1 - \delta)Az$ and $\|z\|_1 < \|x_0\|_1$. This implies $\|(1 - \delta)z\|_1 < \|x_0\|_1$ so that $(1 - \delta)z$ is better than x_0 which is thus not a solution. \square

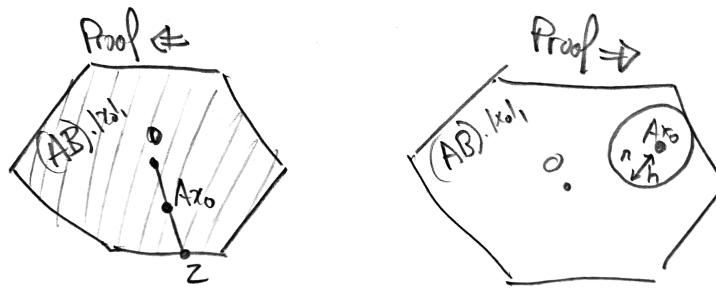


Figure 14.2: Graphical display of the proof for the polytope analysis of ℓ^1 exact recovery.

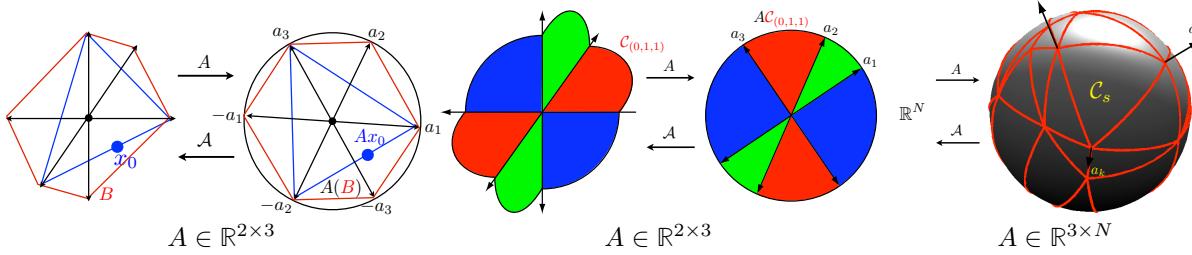


Figure 14.3: Display of the action of the linear map A on the ℓ^1 ball B , and of the inverse non-linear map \mathcal{A} defined by the solution of $(\mathcal{P}_0(y))$.

This results state that “friendly” identifiable vectors (those recovered by ℓ^1) are those who gets projected by A on the boundary of the polytope $\|x_0\|_1 AB$. Intuitively, if P is small in comparison to N , then this projected polytope is small, and most vector will failed to be reconstructed by solving ℓ^1 minimization. This also suggests why using random projections as in Chapter 15, because somehow they results in a low distortion embedding of the ℓ^1 ball from \mathbb{R}^N to \mathbb{R}^P .

Note that if x_0 is identifiable, so is λx_0 for ρx_0 for $\rho > 0$, and in fact, since the recovery condition only depends on the geometry of the faces of B , the obtained condition (14.1) only depends on $\text{sign}(x_0)$. We denote $\mathcal{A} : y \mapsto x^*$ the map from y to a solution of $(\mathcal{P}_0(y))$, which we assume is unique for simplicity of exposition. Condition (14.1) thus shows that A and \mathcal{A} are inverse bijection on a family of cones $\mathcal{C}_s = \{x ; \text{sign}(x) = s\}$ and AC_s for certain “friendly” sign patterns s . These cones AC_s form a partition of the image space \mathbb{R}^P . Assuming for simplicity that the columns $(a_j)_j$ of A have unit norm, for $P = 3$, the interaction of these AC_s with the unit sphere of \mathbb{R}^3 for a so-called Delaunay triangulation of the sphere (this construction extends to higher dimension by replacing triangle by simplexes), see also Figure 14.7. Such Delaunay triangulation is characterized by the empty spherical cap property (each circumcircle associated to a triangle should not contains any columns vector a_j of the matrix). Figure 14.3 illustrate these conclusions in \mathbb{R}^2 and \mathbb{R}^3 .

14.1.3 Optimality Conditions

In the following, given an index set $I \subset \{1, \dots, N\}$, denoting $A = (a_i)_{i=1}^N$ the columns of A , we denote $A_I \stackrel{\text{def.}}{=} (a_i)_{i \in I} \in \mathbb{R}^{P \times |I|}$ the extracted sub-matrix. Similarly, for $x \in \mathbb{R}^N$, we denote $x_I \stackrel{\text{def.}}{=} (x_i)_{i \in I} \in \mathbb{R}^{|I|}$.

The following proposition rephrases the first order optimality conditions in a handy way.

Proposition 44. x_λ is a solution to $(\mathcal{P}_\lambda(y))$ for $\lambda > 0$ if and only if

$$\eta_{\lambda, I} = \text{sign}(x_{\lambda, I}) \quad \text{and} \quad \|\eta_{\lambda, I}\| \leq \lambda$$

where we define

$$I \stackrel{\text{def.}}{=} \text{supp}(x_\lambda) \stackrel{\text{def.}}{=} \{i ; x_{\lambda, i} \neq 0\}, \quad \text{and} \quad \eta_\lambda \stackrel{\text{def.}}{=} \frac{1}{\lambda} A^*(y - Ax_\lambda). \quad (14.2)$$

Proof. Since $(\mathcal{P}_\lambda(y))$ involves a sum of a smooth and a continuous function, its sub-differential reads

$$\partial f_\lambda(x) = \frac{1}{\lambda} A^*(Ax - y) + \lambda \partial \|\cdot\|_1(x).$$

Thus x_λ is solution to $(\mathcal{P}_\lambda(y))$ if and only if $0 \in \partial f_\lambda(x_\lambda)$, which gives the desired result. \square

Note that one has in particular that $\text{supp}(x_\lambda) \subset \text{sat}(\eta_\lambda)$.

The following proposition studies the limit case $\lambda = 0$ and introduces the crucial concept of “dual certificates”, which are the Lagrange multipliers of the constraint \mathcal{L}_y .

Proposition 45. x^* being a solution to $(\mathcal{P}_0(y))$ is equivalent to having $Ax^* = y$ and that

$$\exists \eta \in \mathcal{D}_0(y, x^*) \stackrel{\text{def.}}{=} \text{Im}(A^*) \cap \partial \|\cdot\|_1(x^*). \quad (14.3)$$

Proof. Since $(\mathcal{P}_0(y))$ involves a sum with a continuous function, one can also compute its sub-differential as

$$\partial f_0(x) = \partial \iota_{\mathcal{L}_y}(x) + \partial \|\cdot\|_1(x).$$

If $x \in \mathcal{L}_y$, then $\partial \iota_{\mathcal{L}_y}(x)$ is the linear space orthogonal to \mathcal{L}_y , i.e. $\ker(A)^\perp = \text{Im}(A^*)$. \square

Note that one has in particular that $\text{supp}(x^*) \subset \text{sat}(\eta)$ for any valid vector $\eta \in \mathcal{D}_0(y, x^*)$.

Writing $I = \text{supp}(x^*)$, one thus has

$$\mathcal{D}_0(y, x^*) = \{\eta = A^*p ; \eta_I = \text{sign}(x_I^*), \|\eta\|_\infty \leq 1\}.$$

Although it looks like the definition of $\mathcal{D}_0(y, x^*)$ depends on the choice of a solution x^* , convex duality (studied in the next chapter) shows that it is not the case (it is the same set for all solutions).

14.1.4 Uniqueness

The following proposition shows that the Lasso selects a set of linearly independent regressors (columns of A). This is why this method is also often called “basis pursuit”.

Proposition 46. *For $\lambda \geq 0$, there is always a solution x^* to $(\mathcal{P}_\lambda(y))$ with $I = \text{supp}(x^*)$ such that $\ker(A_I) = \{0\}$*

Proof. Let x be a solution and denote $I = \text{supp}(x)$. If $\ker(A_I) \neq \{0\}$, one selects $h_I \in \ker(A_I)$ and define for $t \in \mathbb{R}$ the vector $x_t \stackrel{\text{def}}{=} x + th$. We denote t_0 the smallest $|t|$ such that $\text{sign}(x_t) \neq \text{sign}(x)$, i.e. $\text{supp}(x_t)$ is strictly included in I . For $t < t_0$, since $Ax_t = Ax$ and $\text{sign}(x_t) = \text{sign}(x)$, x_t still satisfies the same first order condition as x_0 , and one can apply either Proposition 45 (for $\lambda = 0$) or Proposition 44 (for $\lambda > 0$), so that x_t is a solution of $(\mathcal{P}_\lambda(y))$. Since the minimized function are lower semi continuous, $x_t \rightarrow x_{t_0}$ is still a solution. If $\ker(A_J) \neq \{0\}$ with $J = \text{supp}(x_{t_0})$, one is over, otherwise one can iterate this argument on x_{t_0} in place of x and have a sequence of supports which is strictly decaying in size, so it must terminate. \square

This results in particular that if columns of A_I are not independent, then the solution of $(\mathcal{P}_\lambda(y))$ is necessarily non-unique.

Assuming that x_λ is a solution such that $\ker(A_I) = \{0\}$, then from $(\mathcal{P}_\lambda(y))$, one obtains the following implicit expression for the solution

$$x_{\lambda, I} = A_I^+ y - \lambda (A_I^* A_I)^{-1} \text{sign}(x_{\lambda, I}). \quad (14.4)$$

This expression can be understood as a form of generalized soft thresholding (one retrieve the soft thresholding when $A = \text{Id}_N$).

The following useful lemma shows that while solutions x_λ to $(\mathcal{P}_\lambda(y))$ are not necessarily unique, the associated “predictor” (i.e. denoised version of y) Ax_λ is however always uniquely defined. Note that according to (14.5), one has

$$\Phi x_\lambda = \text{Proj}_{\text{Im}(A_I)} y - \lambda A_I (A_I^* A_I)^{-1} \text{sign}(x_{\lambda, I}). \quad (14.5)$$

so up to a $O(\lambda)$ bias, this predictor is an orthogonal projection on a low dimensional subspace indexed by I .

Lemma 3. *For $\lambda \geq 0$, if (x_1, x_2) are solution to $(\mathcal{P}_\lambda(y))$, then $Ax_1 = Ax_2$.*

Proof. For $\lambda = 0$, this is trivial because $Ax_1 = Ax_2 = y$. Otherwise, let us assume $Ax_1 \neq Ax_2$. Then for $x = (x_1 + x_2)/2$, one has

$$\|x\|_1 \leq \frac{\|x_1\|_1 + \|x_2\|_1}{2} \quad \text{and} \quad \|Ax - y\|^2 < \frac{\|Ax_1 - y\|^2 + \|Ax_2 - y\|^2}{2}$$

where the second inequality follows from the strict convexity of the square. This shows that

$$\frac{1}{2\lambda} \|Ax - y\|^2 + \|x\|_1 < \frac{1}{2\lambda} \|Ax_1 - y\|^2 + \|x_1\|_1,$$

which is a contradiction to the optimality of x_1 . \square

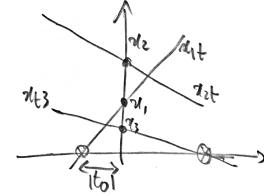


Figure 14.4: Trajectory $(x_t)_t$.

Proposition 47. For $\lambda > 0$, let x_λ be a solution to $(\mathcal{P}_\lambda(y))$ and denote $\eta_\lambda \stackrel{\text{def.}}{=} \frac{1}{\lambda} A^*(y - Ax_\lambda)$. We define the “extended support” as

$$J \stackrel{\text{def.}}{=} \text{sat}(\eta_\lambda) \stackrel{\text{def.}}{=} \{i ; |\eta_{\lambda,i}| = 1\}.$$

If $\ker(A_J) = \{0\}$ then x_λ is the unique solution of $(\mathcal{P}_\lambda(y))$.

Proof. If \tilde{x}_λ is also a minimizer, then by Lemma 3, $Ax_\lambda = A\tilde{x}_\lambda$, so that in particular they share the same dual certificate

$$\eta_\lambda = \frac{1}{\lambda} A^*(y - Ax_\lambda) = \frac{1}{\lambda} A^*(y - A\tilde{x}_\lambda).$$

The first order condition, Proposition 44, shows that necessarily $\text{supp}(x_\lambda) \subset J$ and $\text{supp}(\tilde{x}_\lambda) \subset J$. Since $A_J x_{\lambda,J} = A_J \tilde{x}_{\lambda,J}$, and since $\ker(A_J) = \{0\}$, one has $x_{\lambda,J} = \tilde{x}_{\lambda,J}$, and thus $x_\lambda = \tilde{x}_\lambda$ because of their supports are included in J . \square

Proposition 48. Let x^* be a solution to $(\mathcal{P}_0(y))$. If there exists $\eta \in \mathcal{D}_0(y, x^*)$ such that $\ker(A_J) = \{0\}$ where $J \stackrel{\text{def.}}{=} \text{sat}(\eta)$ then x^* is the unique solution of $(\mathcal{P}_0(y))$.

Proof. The proof is the same as for Proposition 47, replacing η_λ by η . \square

These propositions can be used to show that if A is drawn according to a distribution having a density over $\mathbb{R}^{p \times n}$, then with probability 1 on the matrix A , the solution to $(\mathcal{P}_\lambda(y))$ is unique. Note that this results is not true if A is non random but y is.

14.1.5 Duality

We now relate the first order conditions and “dual certificate” introduced above to the duality theory detailed in Section 12.2. This is not strictly needed to derive the theory of sparse regularization, but this offers an alternative point of view and allows to better grasp the role played by the certificates.

Theorem 41. For any $\lambda \geq 0$ (i.e. including $\lambda = 0$), one has strong duality between $(\mathcal{P}_\lambda(y))$ and

$$\sup_{p \in \mathbb{R}^p} \left\{ \langle y, p \rangle - \frac{\lambda}{2} \|p\|^2 ; \|A^*p\|_\infty \leq 1 \right\}. \quad (14.6)$$

One has for any $\lambda \geq 0$ that (x^*, p^*) are primal and dual solutions if and only if

$$A^*p^* \in \partial \|\cdot\|_1(x^*) \Leftrightarrow (I \subset \text{sat}(A^*p) \text{ and } \text{sign}(x_I^*) = A_I^*p), \quad (14.7)$$

where we denoted $I = \text{supp}(x^*)$, and furthermore, for $\lambda > 0$,

$$p^* = \frac{y - Ax^*}{\lambda}.$$

while for $\lambda = 0$, $Ax^* = y$.

Proof. There are several ways to derive the same dual. One can for instance directly use the Fenchel-Rockafeller formula (??). But it is instructive to do the computations using Lagrange duality. One can first consider the following re-writing of the primal problem

$$\min_{x \in \mathbb{R}^N} \{f(z) + \|x\|_1 ; Ax = z\} = \min_{x \in \mathbb{R}^N} \sup_{p \in \mathbb{R}^p} \mathcal{L}(x, z, p) \stackrel{\text{def.}}{=} f_\lambda(z) + \|x\|_1 + \langle z - Ax, p \rangle$$

where $f_\lambda(z) \stackrel{\text{def.}}{=} \frac{1}{2\lambda} \|z - y\|^2$ if $\lambda > 0$ and $f(z) = \iota_{\{y\}}(z)$ if $\lambda = 0$. For $\lambda > 0$ since f_λ and $\|\cdot\|_1$ are continuous, strong duality holds. For $\lambda = 0$, since the constraint appearing in f_0 is linear (actually a singleton), strong duality holds also. Thus using Theorem 32, one can exchange the min and the max and obtains

$$\max_{p \in \mathbb{R}^p} (\min z \langle z, p \rangle + f_\lambda(z)) + (\min x \|x\|_1 - \langle x, A^*p \rangle) = \max_{p \in \mathbb{R}^p} -f_\lambda^*(-p) - (\|\cdot\|_1)^*(A^*p).$$

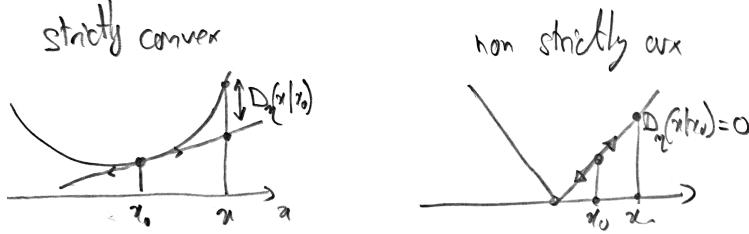


Figure 14.5: Visualization of Bregman divergences.

Using (30), one has that $(\|\cdot\|_1^* = \iota_{\|\cdot\|_\infty \leq 1})$. For $\lambda > 0$, one has using Proposition 31 that

$$f_\lambda^* = (\frac{1}{2\lambda} \|\cdot - y\|^2)^* = \frac{1}{\lambda} (\frac{1}{2} \|\cdot - y\|^2)^*(\lambda \cdot) = \frac{1}{2\lambda} \|\lambda \cdot\|^2 + \langle \cdot, y \rangle$$

which gives the desired dual problem. The first order optimality conditions read $Ax^* = z^*$ and

$$0 \in \partial \|\cdot\|_1(x^*) - A^* p^* \quad \text{and} \quad 0 \in \partial f_\lambda(z^*) + p^*.$$

The first condition is equivalent to (14.7). For $\lambda > 0$, f_λ is smooth, and the second condition is equivalent to

$$p^* = \frac{y - A^* x^*}{\lambda} \quad \text{and} \quad A^* p^* \in \partial \|\cdot\|_1(x^*)$$

which are the desired formula. For $\lambda = 0$, the second condition holds as soon as $z^* = Ax^* = y$. \square

Note that in the case $\lambda > 0$, (14.6) is strongly convex, and in fact the optimal solution p_λ is computed as an orthogonal projection

$$p_\lambda \in \operatorname{argmin}_{p \in \mathbb{R}^P} \{\|p - y/\lambda\| ; \|A^* p\|_\infty \leq 1\}.$$

The sup in (14.6) is thus actually a max if $\lambda > 0$. If $\lambda > 0$, in case $\ker(A^*) = \operatorname{Im}(A)^\perp = \{0\}$, the constraint set of the dual is bounded, so that the sup is also a max.

14.2 Consistency and Sparsistency

14.2.1 Bregman Divergence Rates for General Regularizations

Here we consider the case of a general regularization of the form

$$\min_{x \in \mathbb{R}^N} \frac{1}{2\lambda} \|Ax - y\|^2 + J(x) \tag{14.8}$$

for a convex regularizer J .

For any $\eta \in \partial J(x_0)$, we define the associated Bregman divergence as

$$D_\eta(x|x_0) \stackrel{\text{def.}}{=} J(x) - J(x_0) - \langle \eta, x - x_0 \rangle.$$

One has $D_\eta(x_0|x_0)$, and since J is convex, one has $D_\eta(x|x_0) \geq 0$ [ToDo: put here drawings].

In the case where J is differentiable, since $\partial J(x_0) = \{\nabla J(x_0)\}$, this divergence simply reads

$$D(x|x_0) \stackrel{\text{def.}}{=} J(x) - J(x_0) - \langle \nabla J(x_0), x - x_0 \rangle.$$

If furthermore J is strictly convex, then $D(x|x_0) = 0$ if and only if $x = x_0$, so that $D(\cdot|\cdot)$ is similar to a distance function (but it does not necessarily satisfies the triangular inequality).

If $J = \|\cdot\|^2$, then $D(x|x_0) = \|x - x_0\|^2$ is the Euclidean norm. If $J(x) = \sum_i x_i(\log(x_i) - 1) + \iota_{\mathbb{R}^+}(x_i)$ is the entropy, then

$$D(x|x_0) = \sum_i x_i \log\left(\frac{x_i}{x_{0,i}}\right) + x_{0,i} - x_i$$

is the so-called Kulback-Leibler divergence on \mathbb{R}_+^N .

The following theorem, which is due to Burger-Osher, state a linear rate in term of this Bregman divergence.

Theorem 42. *If there exists*

$$\eta = A^*p \in \text{Im}(A^*) \cap \partial J(x_0), \quad (14.9)$$

then one has for any x_λ solution of (14.8)

$$D_\eta(x_\lambda|x_0) \leq \frac{1}{2} \left(\frac{\|w\|}{\sqrt{\lambda}} + \sqrt{\lambda}\|p\| \right)^2. \quad (14.10)$$

Futhermore, one has the useful bound

$$\|Ax_\lambda - y\| \leq \|w\| + (\sqrt{2} + 1)\|p\|\lambda. \quad (14.11)$$

Proof. The optimality of x_λ for (14.8) implies

$$\frac{1}{2\lambda} \|Ax_\lambda - y\|^2 + J(x_\lambda) \leq \frac{1}{2\lambda} \|Ax_0 - y\|^2 + J(x_0) = \frac{1}{2\lambda} \|w\|^2 + J(x_0).$$

Hence, using $\langle \eta, x_\lambda - x_0 \rangle = \langle p, Ax_\lambda - Ax_0 \rangle = \langle p, Ax_\lambda - y + w \rangle$, one has

$$\begin{aligned} D_\eta(x_\lambda|x_0) &= J(x_\lambda) - J(x_0) - \langle \eta, x_\lambda - x_0 \rangle \leq \frac{1}{2\lambda} \|w\|^2 - \frac{1}{2\lambda} \|Ax_\lambda - y\|^2 - \langle p, Ax_\lambda - y \rangle - \langle p, w \rangle \\ &= \frac{1}{2\lambda} \|w\|^2 - \frac{1}{2\lambda} \|Ax_\lambda - y + \lambda p\|^2 + \lambda\|p\|^2 - \langle p, w \rangle \\ &\leq \frac{1}{2\lambda} \|w\|^2 + \frac{\lambda}{2} \|p\|^2 + \|p\| \|w\| = \frac{1}{2} \left(\frac{\|w\|}{\sqrt{\lambda}} + \sqrt{\lambda}\|p\| \right)^2. \end{aligned}$$

From the second line above, since $D_\eta(x_\lambda|x_0) \geq 0$, one has using Cauchy-Schwartz

$$\|Ax_\lambda - y + \lambda p\|^2 \leq \|w\|^2 + 2\lambda^2\|p\|^2 + 2\lambda\|p\|\|w\| \leq \|w\|^2 + 2\sqrt{2}\|p\|\|w\|\lambda + 2\lambda^2\|p\|^2 = \left(\|w\| + \sqrt{2}\lambda\|p\| \right)^2.$$

Hence

$$\|Ax_\lambda - y\| \leq \|Ax_\lambda - y + \lambda p\| + \lambda\|p\| \leq \|w\| + \sqrt{2}\lambda\|p\| + \lambda\|p\|.$$

□

Choosing $\lambda = \|w\|/\|p\|$ in (14.10), one thus obtain a linear rate in term of Bregman divergence $D_\eta(x_\lambda|x_0) \leq 2\|w\|\|p\|$. For the simple case of a quadratic regularized $J(x) = \|x\|^2/2$, as used in Section ??, one sees that the source conditions (14.9) simply reads

$$x_0 \in \text{Im}(A^*)$$

which is equivalent to (10.12) with exponent $\beta = \frac{1}{2}$, and under this condition, (14.10) gives the following sub-linear rate in term of the ℓ^2 norm

$$\|x_0 - x_\lambda\| \leq 2\sqrt{\|w\|\|p\|}.$$

[ToDo: This seems inconsistent, this should corresponds to $\beta = 1$ to obtain the same rates in both theorems!]

Note that the “source condition” (14.9) is equivalent to x_0 such that $Ax_0 = y$ is a solution to the constraint problem

$$\min_{Ax=y} J(x).$$

So simply being a solution of the constraint noiseless problem thus implies a linear rate for the resolution of the noisy problem in term of the Bregman divergence.

14.2.2 Linear Rates in Norms for ℓ^1 Regularization

The issue with the control (14.10) of the error in term of Bregman divergence is that it is not “distance-like” for regularizers J which are not strictly convex. This is in particular the case for the ℓ^1 norm $J = \|\cdot\|_1$ which we now study.

The following fundamental lemma shows however that this Bregman divergence for ℓ^1 behave like a distance (and in fact controls the ℓ^1 norm) on the indexes where η does not saturate.

Lemma 4. For $\eta \in \partial\|\cdot\|_1(x_0)$, let $J \stackrel{\text{def}}{=} \text{sat}(\eta)$. Then

$$D_\eta(x|x_0) \geq (1 - \|\eta_{J^c}\|_\infty) \|(x - x_0)_{J^c}\|. \quad (14.12)$$

Proof. Note that $x_{0,J^c} = 0$ since $\text{supp}(x_0) \subset \text{sat}(\eta)$ by definition of the subdifferential of the ℓ^1 norm. Since the ℓ^1 norm is separable, each term in the sum defining $D_\eta(x|x_0)$ is positive, hence

$$\begin{aligned} D_\eta(x|x_0) &= \sum_i |x_i| - |x_{0,i}| - \eta_i(x_i - x_{0,i}) \geq \sum_{i \in J^c} |x_i| - |x_0| - \eta_i(x_i - x_{0,i}) \\ &= \sum_{i \in J^c} |x_i| - \eta_i x_i \geq \sum_{i \in J^c} (1 - |\eta_i|) |x_i| \geq (1 - \|\eta_{J^c}\|_\infty) \sum_{i \in J^c} |x_i| = (1 - \|\eta_{J^c}\|_\infty) \sum_{i \in J^c} |x_i - x_{0,i}|. \end{aligned}$$

□

The quantity $1 - \|\eta_{J^c}\|_\infty > 0$ controls how much η is “inside” the subdifferential. The larger this coefficients, the better is the control of the Bregman divergence.

The following theorem uses this lemma to state the convergence rate of the sparse regularized solution, under the same hypothesis has Proposition 48 (with $x^\star = x_0$).

Theorem 43. If there exists

$$\eta \in \mathcal{D}_0(Ax_0, x_0) \quad (14.13)$$

and $\ker(A_J) = \{0\}$ where $J \stackrel{\text{def}}{=} \text{sat}(\eta)$ then choosing $\lambda = c\|w\|$, there exists C (depending on c) such that any solution x_λ of $\mathcal{P}(Ax_0 + w)$ satisfies

$$\|x_\lambda - x_0\| \leq C\|w\|. \quad (14.14)$$

Proof. We denote $y = Ax_0 + w$. The optimality of x_λ in $(\mathcal{P}_\lambda(y))$ implies

$$\frac{1}{2\lambda} \|Ax_\lambda - y\|^2 + \|x_\lambda\|_1 \leq \frac{1}{2\lambda} \|Ax_0 - y\|^2 + \|x_0\|_1 = \frac{1}{2\lambda} \|w\|^2 + \|x_0\|_1$$

and hence

$$\|Ax_\lambda - y\|^2 \leq \|w\|^2 + 2\lambda\|x_0\|_1$$

Using the fact that A_J is injective, one has $A_J^+ A_J = \text{Id}_J$, so that

$$\begin{aligned} \|(x_\lambda - x_0)_J\|_1 &= \|A_J^+ A_J(x_\lambda - x_0)_J\|_1 \leq \|A_J^+\|_{1,2} \|A_J x_{\lambda,J} - y + w\| \leq \|A_J^+\|_{1,2} (\|A_J x_{\lambda,J} - y\| + \|w\|) \\ &\leq \|A_J^+\|_{1,2} (\|Ax_\lambda - y\| + \|A_{J^c} x_{\lambda,J^c}\| + \|w\|) \\ &\leq \|A_J^+\|_{1,2} (\|Ax_\lambda - y\| + \|A_{J^c}\|_{2,1} \|x_{\lambda,J^c} - x_{0,J^c}\|_1 + \|w\|) \\ &\leq \|A_J^+\|_{1,2} \left(\|w\| + (\sqrt{2} + 1)\|p\|\lambda + \|A_{J^c}\|_{2,1} \|x_{\lambda,J^c} - x_{0,J^c}\|_1 + \|w\| \right) \end{aligned}$$

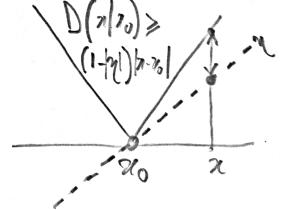


Figure 14.6: Controlling Bregman divergence with the ℓ^1 norm when η is not saturating.

where we used $x_{0,J^c} = 0$ and (14.11). One plug this bound in the decomposition, and using (14.12) and (14.10)

$$\begin{aligned}\|x_\lambda - x_0\|_1 &= \|(x_\lambda - x_0)_J\|_1 + \|(x_\lambda - x_0)_{J^c}\|_1 \\ &\leq \|(x_\lambda - x_0)_{J^c}\|_1 (1 + \|A_J^+\|_{1,2} \|A_{J^c}\|_{2,1}) + \|A_J^+\|_{1,2} ((\sqrt{2} + 1) \|p\| \lambda + 2 \|w\|) \\ &\leq \frac{D_\eta(x|x_0)}{1 - \|\eta_{J^c}\|_\infty} (1 + \|A_J^+\|_{1,2} \|A_{J^c}\|_{2,1}) + \|A_J^+\|_{1,2} ((\sqrt{2} + 1) \|p\| \lambda + 2 \|w\|) \\ &\leq \frac{\frac{1}{2} \left(\frac{\|w\|}{\sqrt{\lambda}} + \sqrt{\lambda} \|p\| \right)^2}{1 - \|\eta_{J^c}\|_\infty} (1 + \|A_J^+\|_{1,2} \|A_{J^c}\|_{2,1}) + \|A_J^+\|_{1,2} ((\sqrt{2} + 1) \|p\| \lambda + 2 \|w\|).\end{aligned}$$

Thus setting $\lambda = c\|w\|$, one obtains the constant

$$C \stackrel{\text{def.}}{=} \frac{\frac{1}{2} \left(\frac{1}{\sqrt{c}} + \sqrt{c} \|p\| \right)^2}{1 - \|\eta_{J^c}\|_\infty} (1 + \|A_J^+\|_{1,2} \|A_{J^c}\|_{2,1}) + \|A_J^+\|_{1,2} ((\sqrt{2} + 1) \|p\| c + 2).$$

□

Note that this theorem does not imply that x_λ is a unique solution, only x_0 is unique in general. The condition (14.13) is often called a “source condition”, and is strengthen by imposing a non-degeneracy $\ker(A_J) = \{0\}$. This non-degeneracy imply some stability in ℓ^2 sense (14.14). The result (14.14) shows a linear rate, i.e. the (possibly multi-valued) inverse map $y \mapsto x_\lambda$ is Lipschitz continuous.

It should be compared with Theorem 31 on linear methods for inverse problem regularization, which only gives sub-linear rate. The sources conditions in the linear (10.12) and non-linear (14.13) cases are however very different. In the linear case, for $\beta = 1/2$, it reads $x_0 \in \text{Im}(A^*) = \ker(A)^\perp$, which is mandatory because linear method cannot recover anything in $\ker(A)$. On contrary, the non-linear source condition only requires that η to be in $\text{Im}(A^*)$, and is able (in the favorable cases of course) to recover information in $\ker(A)$.

14.2.3 Sparsistency

Theorem 43 is abstract in the sense that it rely on hypotheses which are hard to check. The crux of the problem, to be able to apply this theorem, is to be able to “construct” a valid certificate (14.13). We now give a powerful “recipe” which – when it works – not only give a sufficient condition for linear rate, but also provides “support stability”.

For any solution x_λ of $(\mathcal{P}_\lambda(y))$, as already done in (14.2), we define the (unique, independent of the chosen solution) dual certificate

$$\eta_\lambda \stackrel{\text{def.}}{=} A^* p_\lambda \quad \text{where} \quad p_\lambda \stackrel{\text{def.}}{=} \frac{y - Ax_\lambda}{\lambda}.$$

The following proposition shows that p_λ converge to a very specific dual certificate of the constrained problem, which we coined “minimal norm” certificate.

Proposition 49. *If $y = Ax_0$ where x_0 is a solution to $(\mathcal{P}_\lambda(y = Ax_0))$, one has*

$$p_\lambda \rightarrow p_0 \stackrel{\text{def.}}{=} \underset{p \in \mathbb{R}^P}{\operatorname{argmin}} \{ \|p\| ; A^* p \in \mathcal{D}_0(y, x_0) \}. \quad (14.15)$$

The vector $\eta_0 \stackrel{\text{def.}}{=} A^* p_0$ is called the “minimum norm certificate”.

Proof. This follows from the fact that p_λ is the unique solution to (14.6) and then applying the same proof as the one done in Proposition 23 to study the small λ limit of penalized problems. □

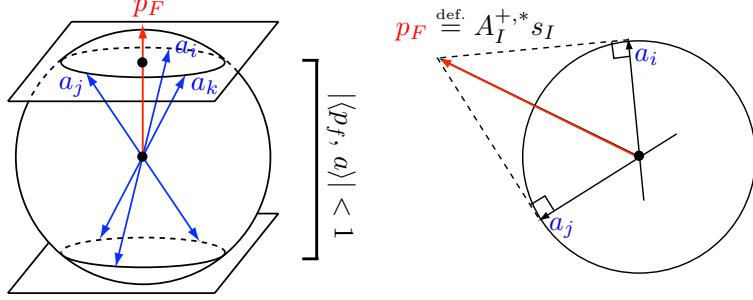


Figure 14.7: Visualization of the condition that $\|\eta_F\|_\infty \leq 1$ as a spherical Delaunay triangulation constraint that all Delaunay spherical caps indexes by identifiable vector should be empty of $(\pm a_i)_i$.

This proposition shows that, while dual certificate $\mathcal{D}_0(y, x_0)$ for $\lambda = 0$ are non-unique, taking the limit as $\lambda \rightarrow 0$ singles-out a specific one, which is of paramount importance to study stability of the support when the noise and λ are small.

A major difficulty in computing (14.24) is that it should satisfy the non-linear constraint $\|\eta_0\|_\infty$. One thus can “simplify” this definition by removing this ℓ^∞ constraint and define the so-called “minimum norm certificate”

$$\eta_F \stackrel{\text{def.}}{=} A^* p_F \quad \text{where} \quad p_F \stackrel{\text{def.}}{=} \underset{p \in \mathbb{R}^P}{\operatorname{argmin}} \{\|p\| ; A_I^* p = \operatorname{sign}(x_{0,I})\}. \quad (14.16)$$

The notation “ η_F ” refers to the “Fuchs” certificate, which we named in honour of J-J. Fuchs who first used it to study ℓ^1 minimization.

We insist that p_F is not necessarily a valid certificate (hence the naming “pre-certificate”) since one does not have in general $\|\eta_F\|_\infty \leq 1$. The vector p_F is a least square solution to the linear system $A_I^* p = \operatorname{sign}(x_{0,I})$, and it can thus be compute in closed form using the pseudo-inverse $p_F = A_I^{*,+} \operatorname{sign}(x_{0,I})$ (see Proposition (21)). In case $\ker(A_I) = \{0\}$, one has the simple formula

$$p_F = A_I (A_I^* A_I)^{-1} \operatorname{sign}(x_{0,I}).$$

Denoting $C \stackrel{\text{def.}}{=} A^* A$ the “correlation” matrix, one has the nice formula

$$\eta_F = C_{\cdot, I} C_{I, I}^{-1} \operatorname{sign}(x_{0,I}). \quad (14.17)$$

The following proposition relates η_F to η_0 , and shows that η_F can be used as a “proxy” for η_0

Proposition 50. *If $\|\eta_F\|_\infty \leq 1$, then $p_F = p_0$ and $\eta_F = \eta_0$.*

The condition $\|\eta_F\|_\infty \leq 1$ implies that x_0 is solution to $(\mathcal{P}_0(y))$. The following theorem shows that if one strengthen this condition to impose a non-degeneracy on η_F , then one has linear rate with a stable support in the small noise regime.

Before proceeding to the proof, let us note that the constraint $\|\eta_F\|_\infty \leq 1$ corresponds to the definition of the spherical Delaunay triangulation, as highlighted by Figure 14.7. This remark was made to us by Charles Dossal.

Remark 8 (Operator norm). In the proof, we use the $\ell^p - \ell^q$ matrix operator norm, which is defined as

$$\|B\|_{p,q} \stackrel{\text{def.}}{=} \max \{\|Bu\|_q ; \|u\|_p \leq 1\}.$$

For $p = q$, we denote $\|B\|_p \stackrel{\text{def.}}{=} \|B\|_{p,p}$. For $p = 2$, $\|B\|_2$ is the maximum singular value, and one has

$$\|B\|_1 = \max_j \sum_i |B_{i,j}| \quad \text{and} \quad \|B\|_\infty = \max_i \sum_j |B_{i,j}|.$$

Theorem 44. If

$$\|\eta_F\|_\infty \leq 1 \quad \text{and} \quad \|\eta_{F,I^c}\|_\infty < 1,$$

and $\ker(A_I) = \{0\}$, then there exists C, C' such that if $\max(\|w\|, \|w\|/\lambda) \leq C$, then the solution x_λ of $(\mathcal{P}_\lambda(y))$ is unique, is supported in I , and in fact

$$x_{\lambda,I} = x_{0,I} + A_I^+ w - \lambda(A_I^* A_I)^{-1} \operatorname{sign}(x_{0,I}^*). \quad (14.18)$$

In particular, $\|x_\lambda - x_0\| = O(\|A^* w\|_\infty) = O(\|w\|)$.

Proof. In the following we denote $T \stackrel{\text{def.}}{=} \min_{i \in I} |x_{0,i}|$ the signal level, and $\delta \stackrel{\text{def.}}{=} \|A^* w\|_\infty$ which is the natural way to measure the noise amplitude in the sparse setting. We define $s \stackrel{\text{def.}}{=} \operatorname{sign}(x_0)$, and consider the “ansatz” (14.18) and thus define the following candidate solution

$$\hat{x}_I \stackrel{\text{def.}}{=} x_{0,I} + A_I^+ w - \lambda(A_I^* A_I)^{-1} s_I, \quad (14.19)$$

and $\hat{x}_{I^c} = 0$. The goal is to show that \hat{x} is indeed the unique solution of $(\mathcal{P}_\lambda(y))$.

Step 1. The first step is to show sign consistency, i.e. that $\operatorname{sign}(\hat{x}) = s$. This is true if $\|x_{0,I} - \hat{x}_I\|_\infty < T$, and is thus implied by

$$\|x_{0,I} - \hat{x}_I\|_\infty \leq K \|A_I^* w\|_\infty + K\lambda < T \quad \text{where} \quad K \stackrel{\text{def.}}{=} \|(A_I^* A_I)^{-1}\|_\infty, \quad (14.20)$$

where we used the fact that $A_I^+ = (A_I^* A_I)^{-1} A_I^*$.

Step 2. The second step is to check the first order condition of Proposition 47, i.e. $\|\hat{\eta}_{I^c}\|_\infty < 1$, where $\lambda \hat{\eta} = A^*(y - Ax)$. This implies indeed that \hat{x} is the unique solution of $(\mathcal{P}_\lambda(y))$. One has

$$\begin{aligned} \lambda \hat{\eta} &= A^*(A_I x_{0,I} + w - A_I(x_{0,I} + A_I^+ w - \lambda(A_I^* A_I)^{-1} s_I)) \\ &= A^*(A_I A_I^+ - \operatorname{Id})w + \lambda \eta_F. \end{aligned}$$

The condition $\|\hat{\eta}_{I^c}\|_\infty < 1$ is thus implied by

$$\|A_{I^c}^* A_I (A_I^* A_I)^{-1}\|_\infty \|A_I^* w\|_\infty + \|A_{I^c}^* w\|_\infty + \lambda \|\eta_{F,I^c}\|_\infty \leq R \|A_I^* w\|_\infty - S\lambda < 0 \quad (14.21)$$

$$R \stackrel{\text{def.}}{=} KL + 1 \quad \text{and} \quad S \stackrel{\text{def.}}{=} 1 - \|\eta_{F,I^c}\|_\infty > 0$$

where we denoted $L \stackrel{\text{def.}}{=} \|A_{I^c}^* A_I\|_\infty$, and also we used the hypothesis $\|\eta_{F,I^c}\|_\infty < 1$.

Conclusion. Putting (14.20) and (14.21) together shows that \hat{x} is the unique solution if (λ, w) are such that the two linear inequations are satisfied

$$\mathcal{R} = \left\{ (\delta, \lambda) ; \delta + \lambda < \frac{T}{K} \quad \text{and} \quad R\delta - S\lambda < 0 \right\}$$

This region \mathcal{R} is triangular-shaped, and includes the following “smaller” simpler triangle

$$\tilde{\mathcal{R}} = \left\{ (\delta, \lambda) ; \frac{\delta}{\lambda} < \frac{S}{R} \quad \text{and} \quad \lambda < \lambda_{\max} \right\} \quad \text{where} \quad \lambda_{\max} \stackrel{\text{def.}}{=} \frac{T(KL + 1)}{K(R + S)}. \quad (14.22)$$

□

It is important to realize that Theorem 44 operates in a “small noise” regime, i.e. $\|w\|$ (and hence λ) needs to be small enough for the support to be identifiable (otherwise small amplitude comment of x_0 will be killed by the regularization). In contrast, Theorem 43 is “global” and holds for any noise level $\|w\|$. The price to pay is that one has no controls about the support (and one does not even know whether x_λ is unique) and than the constant involved are more pessimistic.

A nice feature of this proof is that it gives access to explicit constant, involving the three key parameter K, L, S , which controls:

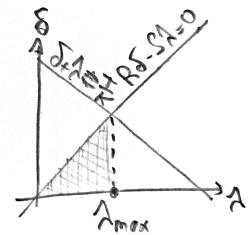


Figure 14.8: Zone in the (λ, δ) where sign consistency occurs.

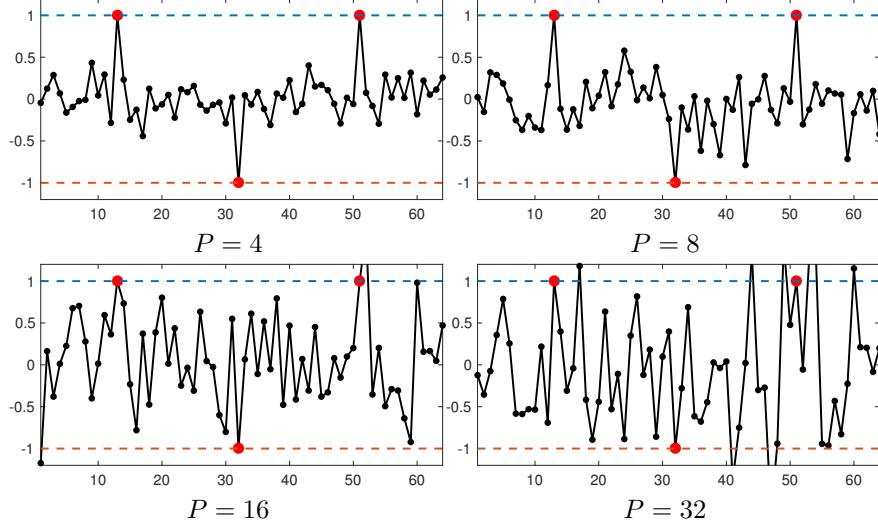


Figure 14.9: Display of certificate η_F for a $A \in \mathbb{R}^{p \times n}$, $n = 64$, with independent Gaussian entries.

- K accounts for the conditionning of the operator on the support I ;
- L accounts for the worse correlation between atoms inside and outside the support ;
- S accounts for how much the certificates η_F is non-degenerate.

The constant on $\|A^*w\|/\lambda$ and on λ are given by (14.22). Choosing (which is in practice impossible, because it requires knowledge about the solution) the smallest possible λ gives $\lambda = \delta \frac{S}{R}$ and in this regime the error is bounded in ℓ^∞ (using other error norms would simply leads to using other matrix norm)

$$\|x_0 - x_\lambda\|_\infty \leq \left(1 + \frac{KL+1}{S}\right) K\delta.$$

The crux of the analysis of the performance (in term of support stability) of ℓ^1 regularization is to be able to say wether, for some class of signal x_0 of interest, η_F is a valid certificate, i.e. $\|\eta_F\|_\infty \leq 1$. Figure 14.9 displays numerically what one obtains when A is random. One see that η_F is non-degenerate when P is large enough. Section 15.2 performs a mathematical analysis of this phenomena.

14.3 Sparse Deconvolution Case Study

Chapter 15 studies the particular case where A is random, in which case it is possible to make very precise statement about wether η_F is a valid certificate.

Another interesting case study, which shows the limitation of this approach, is the case of “super-resolution”. It corresponds to inverse problems where the columns $(a_i)_i$ of A are highly correlated, since typically they are obtained by sampling a smooth kernel.

We thus consider the case where $a_i = \varphi(z_i)$ where the $(z_i)_i \subset \mathbb{X}$ is a sampling grid of a domain \mathbb{X} and $\varphi : \mathbb{X} \rightarrow \mathcal{H}$ is a smooth map. One has

$$Ax = \sum_i x_i \varphi(z_i).$$

Since we seek for sparse x , one can view x as representing the weights of a discrete measure $m_x \stackrel{\text{def.}}{=} \sum_{i=1}^N x_i \delta_{z_i}$ where the dirac masses are constraint to be on the sampling grid.

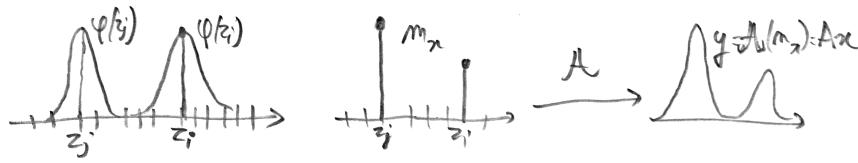


Figure 14.10: Convolution operator.

The matrix A is a discretized version of an infinite dimensional operator mapping Radon measures to vectors of observations $\mathcal{A} : m \in \mathcal{M}(\mathbb{X}) \mapsto y = \mathcal{A}m \in \mathcal{H}$

$$\mathcal{A}(m) \stackrel{\text{def.}}{=} \int_{\mathbb{X}} \varphi(x) dm(x).$$

Indeed, one has for discrete measure $\mathcal{A}(m_x) = Ax$.

A typical example is when using $\mathcal{H} = L^2(\mathbb{X})$ with $\mathbb{X} = \mathbb{R}^d$ or $\mathbb{X} = \mathbb{T}^d$ and $\varphi(z) = \tilde{\varphi}(z - \cdot)$, which corresponds to a convolution

$$(\mathcal{A}m)(z) = \int \tilde{\varphi}(z - x) dm(x) = (\tilde{\varphi} \star m)(z).$$

Note that here \mathcal{H} is infinite dimensional, and to get finite dimensional observations, it suffices to sample the output and consider $\varphi(z) = (\varphi(z - r_j))_{j=1}^P$ (note that the observation grid $r \in \mathbb{X}^P$ can be different from the recovery grid $z \in \mathbb{X}^N$).

Another example, actually very much related, is when using $\varphi(z) = (e^{ikz})_{k=-f_c}^{f_c}$ on $\mathbb{X} = \mathbb{T}$, so that \mathcal{A} corresponds to computing the f_c low-frequencies of the Fourier transform of the measure

$$\mathcal{A}(m) = \left(\int_{\mathbb{T}} e^{ikx} dm(x) \right)_{k=-f_c}^{f_c}.$$

The operator $\mathcal{A}^* \mathcal{A}$ is a convolution against an ideal low pass (Dirichlet) kernel. By weighting the Fourier coefficients, one can this way model any low pass filtering on the torus.

Yet another interesting example on $\mathbb{X} = \mathbb{R}^+$ is the Laplace transform

$$\mathcal{A}(m) = z \mapsto \int_{\mathbb{R}^+} e^{-xz} dm(x).$$

We denote the “continuous” covariance as

$$\forall (z, z') \in \mathbb{X}^2, \quad \mathcal{C}(z, z') \stackrel{\text{def.}}{=} \langle \varphi(z), \varphi(z') \rangle_{\mathcal{H}}.$$

Note that this \mathcal{C} is the kernel associated to the operator $\mathcal{A}^* \mathcal{A}$. The discrete covariance, defined on the computational grid is $C = (\mathcal{C}(z_i, z'_i))_{(i,i')} \in \mathbb{R}^{N \times N}$, while its restriction to some support set I is $C_{I,I} = (\mathcal{C}(z_i, z'_i))_{(i,i') \in I^2} \in \mathbb{R}^{I \times I}$.

Using (14.17), one sees that η_F is obtained as a sampling on the grid of a “continuous” certificate $\tilde{\eta}_F$

$$\eta_F = (\tilde{\eta}_F(z_i))_{i=1}^N \in \mathbb{R}^N,$$

$$\text{where } \tilde{\eta}_F(x) = \sum_{i \in I} b_i \mathcal{C}(x, z_i) \quad \text{where } b_I = C_{I,I}^{-1} \text{sign}(x_0, I), \quad (14.23)$$

so that η_F is a linear combination of I basis functions $(\mathcal{C}(x, z_i))_{i \in I}$.

The question is whether $\|\eta_F\|_{\ell^\infty} \leq 1$. If the grid is fine enough, i.e. N large enough, this can only hold if $\|\tilde{\eta}_F\|_{L^\infty} \leq 1$. The major issue is that $\tilde{\eta}_F$ is only constrained by construction to interpolate $\text{sign}(x_0, I)$

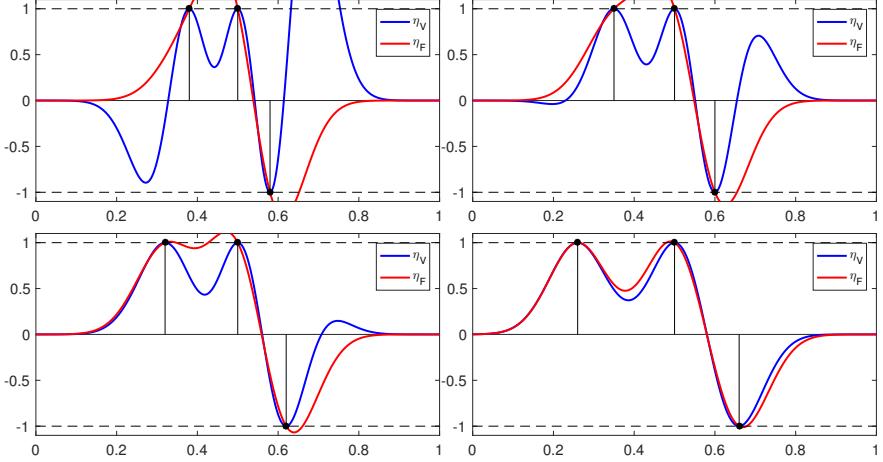


Figure 14.11: Display of “continuous” certificate η_F and η_V for A being a convolution operator.

are points $z_{0,i}$ for $i \in I$. So nothing prevents $\tilde{\eta}_F$ to go outside $[-1, 1]$ around each interpolation point. Figure 14.11 illustrates this fact.

In order to guarantee this property of “local” non-degeneracy around the support, one has to impose on the certificate the additional constraint $\eta'(z_i) = 0$ for $i \in I$. This leads to consider a minimum pre-certificate with vanishing derivatives

$$\eta_V \stackrel{\text{def.}}{=} A^* p_V \quad \text{where} \quad p_V \underset{p \in L^2(\mathbb{R})}{\operatorname{argmin}} \left\{ \|p\|_{L^2(\mathbb{R})} ; \tilde{\eta}(z_I) = \operatorname{sign}(x_{0,I}), \tilde{\eta}'(z_I) = \mathbf{0}_I \right\}. \quad (14.24)$$

where we denoted $\tilde{\eta} = \bar{\psi} \star p$. Similarly to (14.23), this vanishing pre-certificate can be written as a linear combination, but this time of $2|I|$ basis functions

$$\tilde{\eta}_V(x) = \sum_{i \in I} b_i \mathcal{C}(x, z_i) + c_i \partial_2 \mathcal{C}(x, z_i),$$

where $\partial_2 \mathcal{C}$ is the derivative of \mathcal{C} with respect to the second variable, and (b, c) are solution of a $2|I| \times 2|I|$ linear system

$$\begin{pmatrix} b \\ c \end{pmatrix} = \begin{pmatrix} (\mathcal{C}(x_i, x_{i'}))_{i, i' \in I^2} & (\partial_2 \mathcal{C}(x_i, x_{i'}))_{i, i' \in I^2} \\ (\partial_1 \mathcal{C}(x_i, x_{i'}))_{i, i' \in I^2} & (\partial_1 \partial_2 \mathcal{C}(x_i, x_{i'}))_{i, i' \in I^2} \end{pmatrix}^{-1} \begin{pmatrix} \operatorname{sign}(x_{0,I}) \\ \mathbf{0}_I \end{pmatrix}.$$

The associated continuous pre-certificate is $\tilde{\eta}_V = \bar{\psi} \star p_V$, and η_V is a sampling on the grid of $\tilde{\eta}_V$. Figure 14.9 shows that this pre-certificate η_V is much better behaved than η_F . If $\|\eta_V\|_\infty \leq 1$, one can apply (43) and thus obtain a linear convergence rate with respect to the ℓ^2 norm on the grid. But for very fine grid, since one is interested in sparse solution, the ℓ^2 norm becomes meaningless (because the L^2 norm is not defined on measures). Since η_V is different from η_F , one cannot directly applies Theorem 44: the support is not stable on discrete grids, which is a fundamental property of super-resolution problems (as opposed to compressed sensing problems). The way to recover interesting results is to use and analyze methods without grids. Indeed, after removing the grid, one can show that η_V becomes the minimum norm certificate (and is the limit of η_λ).

Chapter 15

Compressed Sensing

This chapter details an important class of inverse problems, which corresponds to using “random” forward operators Φ . This is interesting from an applicative point of view since it allows to model a novel class of imaging devices which can potentially have improved resolution with respect to traditional operators (e.g. low-pass filters for usual cameras) when using in conjunction with sparse regularization technics. This is also interesting from a theoretical point of view, since the mathematical analysis becomes much simpler than with deterministic operators, and one can have good recovery and stability performances. Let us however stress that the “physical” creation of hardware that fulfils the theoretical hypothesis, in particular in medical imaging, is still largely open (put aside some restricted areas), although the theory gives many insightful design guides to improve imaging devices.

The main references for this chapter are [28, 22, 35].

15.1 Motivation and Potential Applications

15.1.1 Single Pixel Camera

In order to illustrate the exposition, we will discuss the “single pixel camera” prototype developed at Rice University [?], and which is illustrated by the figure 15.1 (left). It is an important research problem of developing a new class of cameras allowing to obtain both the sampling and the compression of the image. Instead of first sampling very finely (ie with very large Q) the analog signal \tilde{f} to obtain a $f \in \mathbb{R}^Q$ image then compressing enormously (ie with M small) using (??), we would like to dispose directly of an economic representation $y \in \mathbb{R}^P$ of the image, with a budget P as close to M and such that one is able to “decompress” y to obtain a good approximation of the image f_0 .

The “single-pixel” hardware performs the compressed sampling of an observed scene \tilde{f}_0 (the letter “R” in Figure 15.1), which is a continuous function indicating the amount of light $\tilde{f}_0(s)$ reaching each point $s \in \mathbb{R}^2$ of the focal plane of the camera. To do this, the light is focused against a set of Q micro-mirrors aligned on the focal plane. These micro-mirrors are not sensors. Unlike conventional sampling (described in Section ??), they do not record any information, but they can each be positioned to reflect or absorb light. To obtain the complete sampling/compression process, one very quickly changes P times the configurations of the micro-mirrors. For $p = 1, \dots, P$, one sets $\Phi_{p,q} \in \{0,1\}$, depending on whether the micromirror at position q has been placed in the absorbing (0) or reflective (value 1) position at step p of the acquisition. The total light reflected at step p is then accumulated into a single sensor (hence the name “single pixel”, in fact it is rather a “single sensor”), which achieves a linear sum of the reflected intensities to obtain the recorded $y_p \in \mathbb{R}$ value. In the end, if the light intensity arriving on the surface c_q of the mirror indexed by $f_q = \int_{c_q} \tilde{f}_0(s) ds$ is denoted (as in the ?? section) as q , the equation that links the discrete image $f \in \mathbb{R}^Q$

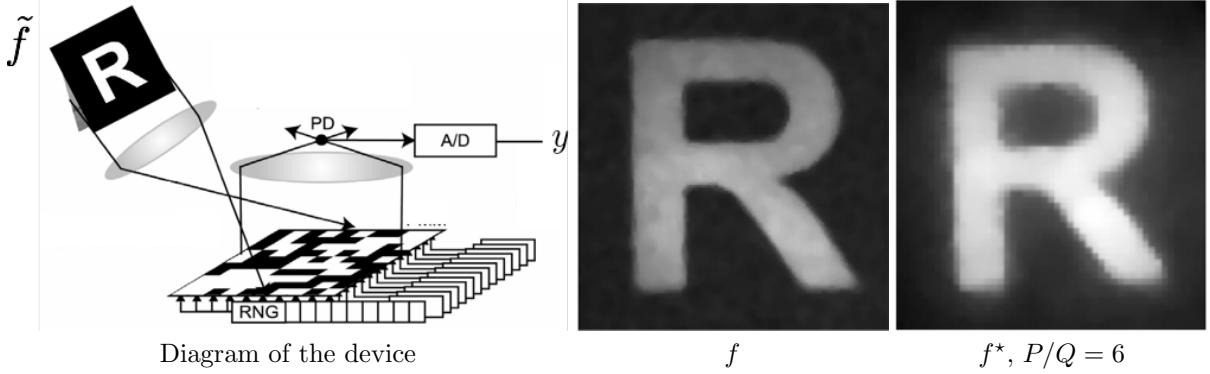


Figure 15.1: Left: diagram of the single-pixel acquisition method. Center: image $f_0 \in \mathbb{R}^Q$ “ideal” observed in the focal plane of the micro-mirrors. Right: image $f_0^* = \Psi x^*$ reconstructed from observation $y \in \mathbb{R}^P$ with a compression factor $P/Q = 6$ using ℓ^1 -type regularization.

“seen through the mirrors” to the P measures $y \in \mathbb{R}^P$ is

$$\forall p = 1, \dots, P, \quad y_p \approx \sum_q \Phi_{p,q} \int_{c_n} \tilde{f}_0(s) ds = (\Phi f_0)_p,$$

(here \approx accounts for some noise), which corresponds to the usual forward model of inverse problems

$$y = \Phi f_0 + w \in \mathbb{R}^P$$

where w is the noise vector. It is important to note that the mirrors do not record anything, so in particular the f_0 discrete image is never calculated or recorded, since the device directly calculates the compressed representation y from the analog signal \tilde{f}_0 . The term w models here the acquisition imperfections (measurement noise). The compressed sampling therefore corresponds to the transition from the observed scene \tilde{f}_0 to the compressed vector y . The “decompression” corresponds to the resolution of an inverse problem, whose goal is to find a good approximation of f_0 (the discrete image “ideal” as seen by the micro-mirrors) from y .

15.1.2 Sparse Recovery

In order to reconstruct an approximation of the (unknown) image f_0 , following Section 11.2, we assume it is sparse in some dictionary Ψ . Denoting $A \stackrel{\text{def}}{=} \Psi \Phi \in \mathbb{R}^{P \times N}$, this leads us to consider the usual ℓ^1 regularized problem (11.10)

$$x_\lambda \in \operatorname{argmin}_{x \in \mathbb{R}^N} \frac{1}{2\lambda} \|y - Ax\|^2 + \|x\|_1, \quad (\mathcal{P}_\lambda(y))$$

so that the reconstructed image is $f_\lambda = \Psi x_\lambda$. We also sometimes consider the constraint problem

$$x_\varepsilon \in \operatorname{argmin}_{\|Ax-y\| \leq \varepsilon} \|x\|_1, \quad (\mathcal{P}^\varepsilon(y))$$

where, for the sake of simplicity, we set $\varepsilon = \|w\|$ (which we assume is known). From a mathematical point of view, these problems are equivalent in the sense that there exists a bijection between λ and ε which links its solution. But in practice, this bijection is not explicitly known and depends on y .

Here, it is important to remember that A is drawn from a random matrix ensemble. For an arbitrary Ψ , it is hard to analyze this random distribution. If Ψ is orthogonal, and the distribution of the columns of Φ are invariant by rotation (which is the case if the entries are i.i.d. Gaussian), then A has the same distribution as Φ . In the following, we thus directly models the distribution of A and assumes it has some nice property (typically it is close to being Gaussian i.i.d.).

15.2 Dual Certificate Theory and Non-Uniform Guarantees

15.2.1 Random Projection of Polytopes

When there is no noise, $w = 0$ a way to tackle the problem is to use the characterization of solutions of $(\mathcal{P}^0(Ax_0)) = (\mathcal{P}_0(Ax_0))$ given in Section 14.1.2. According to Proposition 43, identifiable vectors with sparsity $\|x_0\|_0 = s$ corresponds to s -dimensional faces of the ℓ^1 balls B_1 which are mapped to face of the projected polytope AB_1 . This leads to a combinatorial problems to count the number of face of random polytope. Donoho and Tanner were able to perform a sharp analysis of this problem. They showed the existence of two regimes, using two functions C_A, C_M so that, with high probability (i.e. a probability converging exponentially fast to 1 with (n, p)) on the matrix A

- All x_0 so that $\|x_0\|_0 \leq C_A(P/N)P$ are identifiable.

- Most x_0 so that $\|x_0\|_0 \leq C_M(P/N)P$ are identifiable.

For instance, they show that $C_A(1/4) = 0.065$ and $C_M(1/4) = 0.25$. Figure 15.5 illustrates numerically these two phase transitions. This analysis can be shown to be sharp in high dimension, i.e. when $\|x_0\|_0 > C_M(P/N)P$, then x_0 is not identifiable with high probability (this corresponds to a phase transition phenomena). For large dimensions (N, P) , the scaling given by C_M describe very well what one observe in practice. For $P = N/4$ (compression of a factor 4), one retrieve in practice all vector with sparsity smaller than P/N . The function C_M can be computed numerically, and it can be shown to have a logarithmic grows $C_M(r) \sim \log(r)$ for small r . This suggests that for high compression regime, one recovers with ℓ^1 minimization almost all vector with a sparsity $\|x_0\|_0$ proportional (up to log factor) to the number of measurements P .

15.2.2 Random Matrices

The analysis of the performance ℓ^1 minimization to solve compressed sensing problem is made possible because of the very precise understanding of the distribution of the singular values of certain random ensemble let us illustrate this in the Gaussian case, which is the simplest, and is very illustrative.

An important part of the recovery proof relies on controlling the correlation matrix $A_I^*A_I$ of selected columns, and even more importantly, its inverse $(A_I A_I)^{-1}$. These random matrices are called Wishart matrices and inverse Wishart matrices. Such a matrix $B = A_I$ is of size (P, s) and is also drawn from the Gaussian ensemble. Fixing s and letting $P \rightarrow +\infty$, one has thanks to the law of large numbers $B^*B \rightarrow \text{Id}_s$ almost surely. This is however not a very realistic setting, since in general, one hope to have s almost equal, up to log factor, to P .

Linear growth $P = s/\beta$. A quite extreme setting is when s grows proportionally to P , and impose $s/P = \beta$. In this case, the eigenvalues of B^*B are, for large p , essentially contained in the interval $[\lambda_-, \lambda_+]$ where $\lambda_{\pm} = (1 \pm \sqrt{\beta})^2$, $\beta \stackrel{\text{def.}}{=} s/p$, in the sense that the probability distribution of eigenvalues converges (in the weak sense of measures) toward the Marcenko-Pastur law

$$\forall (u, v) \in \mathbb{R}_+^2, \quad \mathbb{P}(\text{eig}(B^\top B) \in [u, v]) \xrightarrow{p \rightarrow +\infty} \int_u^v f_\beta(\lambda) d\lambda$$

where one fix the ratio $\beta = s/P$, and the Marcenko-Pastur law is

$$f_\beta(\lambda) \stackrel{\text{def.}}{=} \frac{1}{2\pi\beta\lambda} \sqrt{(\lambda - \lambda_-)_+(\lambda_+ - \lambda)} \mathbf{1}_{[\lambda_-, \lambda_+] }(\lambda).$$

Figure (15.3) illustrates this convergence.

Super-linear grows $P = s \log(\dots)$. In order to have a better concentration of the singular values of $A_I^*A_I$ around 1, one needs to have a slightly super-linear growth of P with s . In this setting one has that $A_I^*A_I$. In order to derive non-asymptotic (i.e. with explicit constants) results, one can use a celebrated

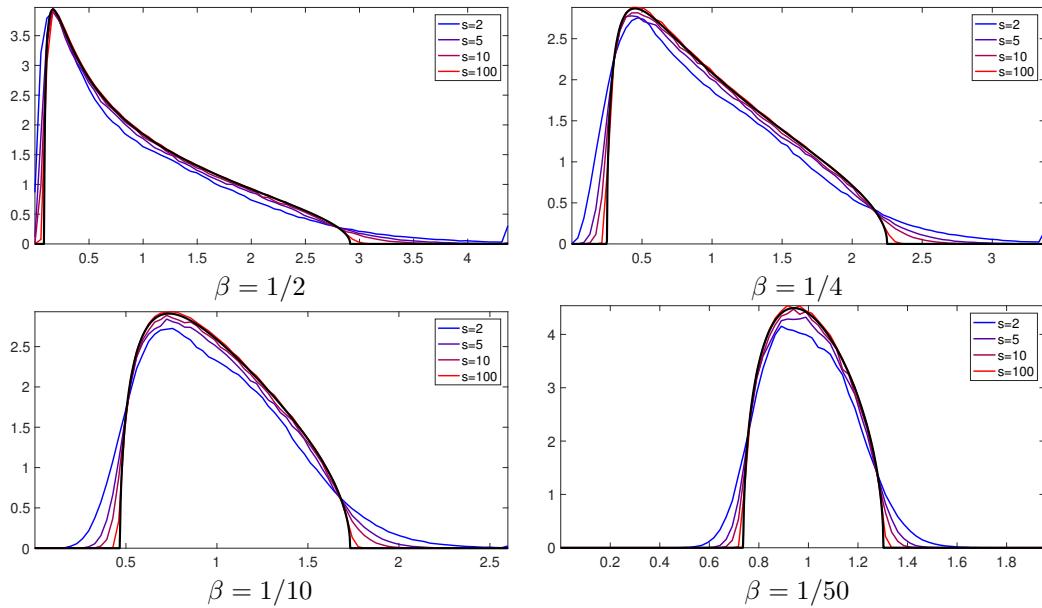


Figure 15.2: Illustration of the convergence toward the Marcenko-Pastur law.

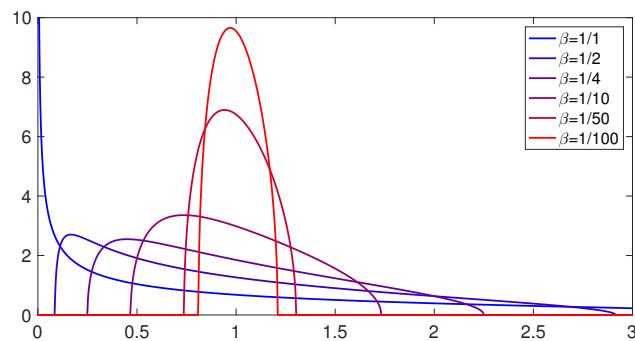


Figure 15.3: Display of the Marcenko-Pastur distribution f_β for various β .

concentration inequality due to Talagrand, which assert that one has a fast concentration of this randomized covariance $A_I^* A_I$ toward its expectation Id_s .

$$\mathbb{P} \left(\|A_I^* A_I - \text{Id}_s\|_{\text{op}} \geq t + \sqrt{\frac{s}{P}} \right) \leq e^{-\frac{t^2 s}{2}}. \quad (15.1)$$

15.2.3 Dual Certificates

In order to analyze recovery performance, one can looks not only for ℓ^2 stability ($\|x_\lambda - x_0\| \sim \|w\|$) but also that x_λ has the same support as x_0 when $\|w\|$ is small. As detailed in Section 14.2.3, this requires to ensure that the pre-certificate η_F defined in (14.24) is non-degenerated, i.e.

$$\|\eta_F\|_\infty \leq 1 \quad \text{where } \eta_F = A^* A_I (A_I^* A_I)^{-1} \text{sign}(x_{0,I}). \quad (15.2)$$

Figure 14.9 suggests that this should be the case if P is large enough with respect to $\|x_0\|$. This theorem backup this observation.

Coherence-based analysis. We first perform a crude analysis using the so-called coherence of the matrix $A = (a_j)_{j=1}^N$ where the $a_j \in \mathbb{R}^P$ are the columns of A , which we assume to be normalized $\|a_j\| = 1$

$$\mu \stackrel{\text{def.}}{=} \max_{i \neq j} |\langle a_i, a_j \rangle| = \|A^* A - \text{Id}_N\|_\infty \quad (15.3)$$

where $\|C\|_\infty = \max_{i,j} |C_{i,j}|$. The coherence is 0 for an orthogonal matrix, and is always smaller than 1, $\mu \in [0, 1]$. The smaller the coherence, the better conditioned the inverse problem $Ax = y$ is, and the more likely is the certificate η_F to be non-degenerate, as shown by the following proposition.

Proposition 51. *One has, denoting $s = \|x_0\|_0 = |I|$ where $I = \text{supp}(x_0)$, for $\mu < \frac{1}{s-1}$,*

$$\|\eta_{F,I^c}\|_\infty \leq \frac{s\mu}{1-(s-1)\mu} \quad \text{and} \quad \|p_F\|^2 \leq \frac{s}{1-(s-1)\mu}. \quad (15.4)$$

In particular, if $s < \frac{1}{2} \left(1 + \frac{1}{\mu}\right)$, $\|\eta_{F,I^c}\|_\infty < 1$ and one can thus apply the recovery Theorem 44.

Proof. We recall that the ℓ^∞ operator norm (see Remark 8) is

$$\|B\|_\infty = \max_i \sum_j |B_{i,j}|.$$

We denote $C = A^* A$. One has

$$\|A_{I^c}^* A_I\|_\infty = \max_{j \in I^c} \sum_{i \in I} C_{i,j} \leq s\mu \quad \text{and} \quad \|\text{Id}_s - A_I^* A_I\|_\infty = \max_{j \in I} \sum_{i \in I, i \neq j} C_{i,j} \leq (s-1)\mu$$

One also has

$$\begin{aligned} \|(A_I^* A_I)^{-1}\|_\infty &= \|(\text{Id}_s - A_I^* A_I)^{-1}\|_\infty = \left\| \sum_{k \geq 0} (\text{Id}_s - A_I^* A_I)^k \right\|_\infty \\ &\leq \sum_{k \geq 0} \|\text{Id}_s - A_I^* A_I\|_\infty^k \leq \sum_{k \geq 0} ((s-1)\mu)^k = \frac{1}{1-(s-1)\mu} \end{aligned}$$

which is legit because the matrix series indeed converge since $(s-1)\mu < 1$. Using these two bounds, one has

$$\|\eta_{F,I^c}\|_\infty = \|A_{I^c}^* A_I (A_I^* A_I)^{-1} \text{sign}(x_{0,I})\|_\infty \leq \|A_{I^c}^* A_I\|_\infty \|(A_I^* A_I)^{-1}\|_\infty \|\text{sign}(x_{0,I})\|_\infty \leq (s\mu) \times \frac{1}{1-(s-1)\mu} \times 1.$$

One has

$$\frac{s\mu}{1-(s-1)\mu} \iff 2s\mu < 1 + \mu$$

which gives the last statement. One also has

$$\|p_F\|^2 = \langle (A_I^* A_I)^{-1} s_I, s_I \rangle \leq \| (A_I^* A_I)^{-1} \|_\infty \|s\|_1 \leq \frac{s}{1-(s-1)\mu}$$

□

Note that this proof actually shows that if $s < \frac{1}{2} \left(1 + \frac{1}{\mu}\right)$, all certificate η_F are valid, for any sign pattern $\text{sign}(x_0)$. This actually implies a much stronger stability in the sense that whatever the noise w (which might not be small), the support of x_λ is included (not necessarily equal) in the one of x_0 .

One can show that one always has

$$\mu \geq \sqrt{\frac{N-P}{P(N-1)}} \quad (15.5)$$

which is equivalent to $1/\sqrt{P}$ for $N \gg P$. For Gaussian matrix $A \in \mathbb{R}^{P \times N}$, one has for large $(N, P) \rightarrow +\infty$

$$\mu \sim \sqrt{\log(PN)/P}$$

which shows that Gaussian matrix are close to being optimal for the bound (15.5) if $N \gg P$. Applying Proposition 51 thus shows that ℓ^1 regularization is able to recover with a stable support vector with less than $s \sim O(\sqrt{P})$ (ignoring log terms) non-zero coefficients. In fact, we will show now that it does much better and recover a proportional number $s \sim O(P)$. This is because the coherence bound (15.4) is very pessimistic.

Randomized analysis of the Fuchs certificate. We consider here a class of sufficiently “random” matrices.

Definition 22 (sub-Gaussian random matrix). *A random matrix $\sqrt{P} A$ is said to be sub-Gaussian if its entries are independent such that $\mathbb{E}(A_{i,j}) = 0$ (zero mean) $\mathbb{E}(A_{i,j}^2) = 1/P$ and*

$$\mathbb{P}(|\sqrt{P} A_{i,j}| \geq t) \leq \beta e^{-\kappa t^2}.$$

Note that its entries does not needs to be identically distributed, but the sub-Gaussianity parameter (β, κ) should not depend on (i, j) . Note also the presence of the normalization factor \sqrt{P} , which is here to ensure $\mathbb{E}(\|a_j\|^2) = 1$ where a_j are the columns of A .

Typical example of sub-Gaussian random matrix are Gaussian or Bernoulli matrices.

Theorem 45. *For a given $x_0 \in \mathbb{R}^N$, denoting $s = \|x_0\|_0$, and assuming A is sub-Gaussian, then for any $0 < \varepsilon < 1$ provided that*

$$P \geq \frac{4c}{1-\delta} s \log(2N/\varepsilon) \quad \text{where} \quad \delta^2 \stackrel{\text{def}}{=} \frac{C}{4c} \left(\frac{7}{\log(2N/\varepsilon)} + \frac{2}{s} \right)$$

condition (15.2) holds with probability $1-\varepsilon$, so that x_λ has the same support and sign as x_0 when $(\|w\|, \|w\|/\lambda)$ is small enough. The constant $C, c, C = \frac{2}{3\tilde{c}}$ where $\tilde{c} \stackrel{\text{def}}{=} \frac{\kappa^2}{4\beta+2\kappa}$ only depends on the sub-Gaussianity parameter (β, κ) appearing (22), and for Gaussian or Benoulli matrices, $c = 1/2$.

For a Gaussian matrix, the scaling is that one should have $P \geq (2 + \delta)s \log(N)$ with a high probability.

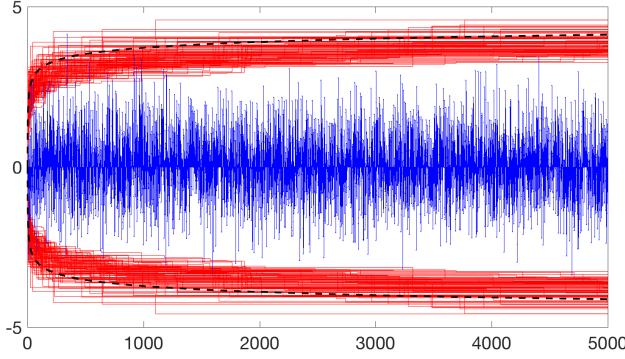


Figure 15.4: Graphical display of how the maximum of N i.i.d. gaussians concentrates tightly just bellow the $\sqrt{2 \log(N)}$ dashed curve.

Proof. We only give the main insight for the proof. Its crux relies on the fact that $\|A_{I^c}^* p_F\|_\infty \leq 1$ reads

$$\max_{j \notin I} |\langle a_j, p_F \rangle| \leq 1$$

where $p_F = A_I^{*+} s_{0,I}$ is *independent* from the vectors $(a_j)_{j \notin I}$ it is correlated against. This allows one to check this condition by first controlling $\|p_F\|$ and then making as if p_F was a fixed deterministic vector. Noticing

$$\|p_F\|^2 = \langle A_I(A_I^* A_I)^{-1} s_{0,I}, A_I(A_I^* A_I)^{-1} s_{0,I} \rangle = \langle (A_I^* A_I)^{-1} s_{0,I}, s_{0,I} \rangle,$$

the heuristic reasoning is that, following what we said in Section (15.2.2), if P grows slightly (logarithmically) faster than s , $A_I^* A_I$ is close to Id_s (see Talagrand inequality (15.1)), so that

$$\|p_F\|^2 \sim \|s_{0,I}\|^2 = s \quad (15.6)$$

For a *fixed* p_F , one has that $\langle a_j, p_F \rangle$ is Gaussian distributed with variance $\|p_F\|^2/P$, and we use the well known fact (already encountered for denoising using thresholding) that the maximum of $P-s$ such vectors concentrates just bellow the universal threshold $\|p_F\|\sqrt{2\log(N-s)/P}$. Using the estimate (15.6), one sees that $\|p_F\|_\infty \leq 1$ is implied by $2\log(N)s/P \leq 1$, which gives the sharp scaling $P \geq 2\log(N)s$.

In order to get robustness to noise, one needs to impose that $\|A_{I^c}^* p_F\| < 1$, which can be achieved by using a slightly stronger scaling $P \geq 2(1+\delta)\log(N)s$ for a small δ .

One can actually make this reasoning very precise, because quite surprisingly, it turns out that $Ps/\|p_F\|^2$ is actually distributed according to a χ^2 variable with $P-s+1$ degrees of freedom (i.e. the sum of $P-s+1$ squares of Gaussian variables). Indeed, for $s=1$, one immediately sees that $P/\|p_F\|^2$ is χ^2 with P degrees of freedom. The general case is more involved, and the proof relies on the fact that the isotropy of the distribution implies that $P/\|p_F\|^2$ is the square of the distance between the first column of A_I and the linear space spanned by the other columns (hence $P-s+1$ degrees of freedom). Since one has a very good understanding of the clustering of such a χ^2 variable around its means, one can thus show that $\|p_F\| \leq (1-\delta)\sqrt{s}$ with high precision for arbitrary small $\delta > 0$. \square

For $(N, s) \rightarrow +\infty$, one has $\delta \rightarrow 0$, so that informally, the scaling is

$$P \geq 2s\log(2N/\varepsilon). \quad (15.7)$$

This theorem states a non-uniform recovery guarantee, in the sense that one first chooses a vector x_0 , then draws the matrix A , and the recovery results holds with high probability. This should be contrasted with the RIP theory developed in Section 15.3 which provides stronger uniform guarantees.

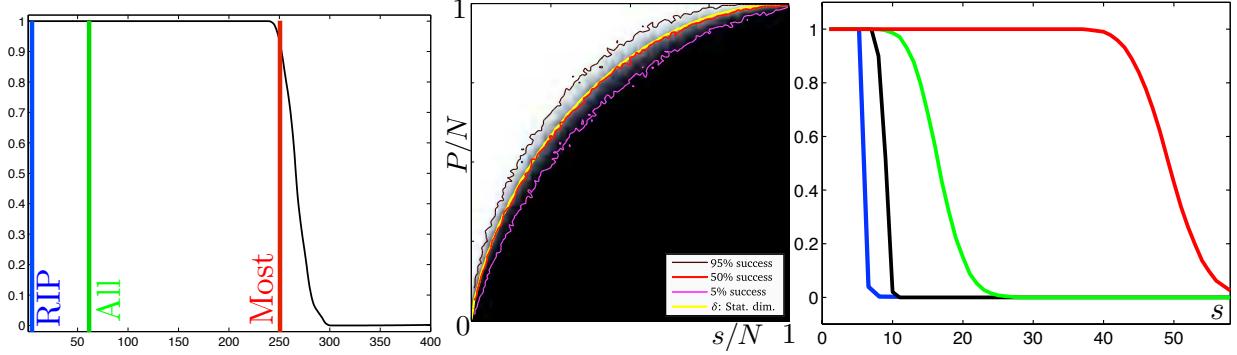


Figure 15.5: Phase transitions. For the figure on the right shows probability as function of sparsity that certain criteria hold true, blue: w-ERC, black: ERC, green $|\eta_F| \leq 1$, red: identifiability.

15.3 RIP Theory for Uniform Guarantees

15.3.1 Restricted Isometry Constants

The Restricted Isometry constant δ_s of a matrix $A \in \mathbb{R}^{P \times N}$ is defined as

$$\forall z \in \mathbb{R}^N, \quad \|z\|_0 \leq s \implies (1 - \delta_s)\|z\|^2 \leq \|Az\|^2 \leq (1 + \delta_s)\|z\|^2, \quad (15.8)$$

and one usually chose the smallest δ_s so that these relation hold.

A related concept is the (s, s') restricted orthogonality constant $\theta_{s,s'}$, which is such that for all (x, x') with $\|x\|_0 \leq s$, $\|x'\|_0 \leq s'$ and disjoint support , one has

$$|\langle Ax, Ax' \rangle| \leq \theta_{s,s'} \|x\| \|x'\|$$

The following lemma shows that RI and RO constants are tightly related.

Lemma 5. *One has*

$$\theta_{s,s'} \leq \delta_{s+s'} \leq \theta_{s,s'} + \max(\delta_s, \delta_{s'}).$$

Proof. We prove the first inequality (which is the most important). We prove that if z and z' have disjoints supports and $\|z\| \leq s$ and $\|z'\|_0 \leq s$,

$$|\langle Az, Az' \rangle| \leq \delta_{2s} \|z\| \|z'\|.$$

Using the RIP (15.8) since $z \pm z'$ has support of size $s + s'$ and the fact that $\|z \pm z'\|^2 = \|z\|^2 + \|z'\|^2$, one has

$$(1 - \delta_{s+s'}) (\|z\|^2 + \|z'\|^2) \leq \|Az \pm Az'\|^2 \leq (1 + \delta_{s+s'}) (\|z\|^2 + \|z'\|^2).$$

One thus has using the parallelogram equality

$$|\langle Az, Az' \rangle| = \frac{1}{4} |\|Az + Az'\|^2 - \|Az - Az'\|^2| \leq \delta_{s+s'} \|z\| \|z'\|.$$

□

The following theorem states that for a sub-Gaussian random matrix, these RIP constants grow slowly with s . Let us stress that, although this is an abuse of notation, here we assume that A is a random matrix, and not a deterministic one as previously considered.

Theorem 46. *If A is a sub-Gaussian random matrix, then provided*

$$P \geq C\delta^{-2}s \log(eN/s) \quad (15.9)$$

it satisfies $\delta_s \leq \delta$ with probability $1 - 2e^{-\delta^2 \frac{m}{2C}}$, where C only depends on the sub-Gaussianity parameters appearing in (22).

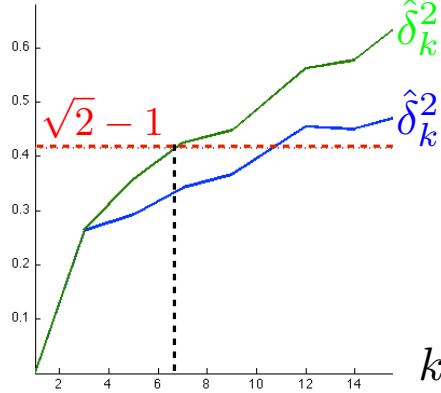


Figure 15.6: Left: evolution of lower bounds $\hat{\delta}_k$ on the RIP constant.

We do not prove this Theorem, and simply give the main intuition. The proof of this theorem relies on results regarding the distribution of the singular values of Gaussian matrices. Indeed, the RIP condition (15.8) is equivalent to having the bound $\text{eig}(A_I^* A_I) \subset [1 - \delta_s, 1 + \delta_s]$ for all Gaussian matrices A_I extracted from A . In the Gaussian case (actually this holds for any random matrix with i.i.d. entries and unit covariance), one has a very good understanding of the distribution of the singular values of covariance matrices $B^* B \in \mathbb{R}^{s \times s}$ of a Gaussian matrix B of size (P, s) , $B \sim \text{randn}(P, s)/\sqrt{P}$, as detailed in Section 15.2.2. In particular, using Talagrand concentration inequality (15.1), one obtains the desired controls over the δ_s constants. The intuition is that, if we assume that $s/P = \beta$ is constant and P is large, then one has that the eigenvalue of $A_I^* A_I$, and for instance its smaller one should be of the order of $2\sqrt{s/P} - s/P$, so that δ_s should be a function of s/P , and hence P should scale proportionally to the s . The log term comes from the exponential number of such matrix to control, and one need to use a non-asymptotic lower bound in place of the Marcenko-Pastur asymptotic law.

15.3.2 RIP implies dual certificates

The following theorem ensures that having small enough restricted isometry constant implies the existence of a valid dual certificates. This means that one can apply Theorem 43, which in turn ensures that one has a stable recovery of sparse signals.

Theorem 47. *If $\delta_s + \theta_{s,s} + \theta_{s,2s} < 1$, then for any x_0 with $\|x_0\| \leq s$, there exists $\eta \in \mathcal{D}_0(Ax_0, x_0)$, i.e. $\eta \in \text{Im}(A^*) \cap \partial\|x_0\|_1$. More precisely, one has, denoting $I = \text{supp}(x_0)$, $\eta_I = \text{sign}(x_0, I)$ and*

$$\|\eta_{I^c}\|_\infty \leq \frac{\theta_{s,s}}{1 - \delta_s - \delta_{s,2s}} < 1 \quad \text{and} \quad \|p\| \leq \frac{\delta_{s,s}}{1 - \delta_s - \theta_{s,2s}} \frac{\sqrt{s}}{\sqrt{1 - \delta_s}}.$$

Note that thanks to Lemma 5, condition

$$\delta_s + \theta_{s,s} + \theta_{s,2s} \leq \delta_s + \delta_{2s} + \delta_{3s} \leq 3\delta_{3s}$$

so that condition $\delta_s + \theta_{s,s} + \theta_{s,2s} < 1$ is implied by $\delta_{3s} < 1/3$. It is furthermore possible to refine the proof of this theorem to obtain alternative (often much sharper) sufficient condition such as $\delta_{2d} \leq \sqrt{2} - 1$. These sufficient conditions involving RI constants are often called “Restricted Isometry Properties” (RIP). As we will illustrate next, however, the constant involved on the sufficient sparsity to guarantee that such uniform RIP conditions holds are large, of the order of a few hundred.

To prove this theorem, it is not enough to directly consider the pre-certificate η_F defined in (15.2). Indeed, using such a certificate leads to slightly suboptimal log-factors in the asymptotic. The proof strategy, developed by Candes and Tao (in “Decoding with Linear Programming”) consists in iteratively “improving”

this certificate by removing a vector interpolating the largest s entries outside I . In order to study η_F and to perform the improvement, it is crucial the behavior of least square solution of interpolation problems using random matrices.

Lemma 6. *We assume $\delta_s < 1$. Let $c \in \mathbb{R}^n$ with $\|c\|_0 \leq s$ and $\text{supp}(c) = J$. Let $\eta = A^*p$ be the least square solution of $\eta_J = A_J^*p = c_J$, i.e.*

$$p \stackrel{\text{def.}}{=} A_J^{*,+}c_J = A_J(A_J^*A_J)^{-1}c_J$$

(note that A_J is full rank because $\delta_s < 1$). Then, denoting $K \subset J^c$ the s' largest entries in magnitude of η_{J^c} , one has

$$\|\eta_{(J \cup K)^c}\|_\infty \leq \frac{\theta_{s,s'}\|c\|}{(1-\delta_s)\sqrt{s}} \quad \text{and} \quad \|\eta_K\| \leq \frac{\theta_{s,s'}\|c\|}{1-\delta_s} \quad \text{and} \quad \|p\| \leq \frac{\|c\|}{\sqrt{1-\delta_s}}$$

Proof. Since $|J| \leq s$, one has that $\lambda_{\min}(A_J^*A_J) \geq 1 - \delta_s$ and hence

$$\|(A_JA_J^*)^{-1}\| \leq \frac{1}{1-\delta_s}$$

One has

$$\|p\|^2 = \langle A_J(A_J^*A_J)^{-1}c_J, A_J(A_J^*A_J)^{-1}c_J \rangle = \langle (A_J^*A_J)^{-1}c_J, c_J \rangle \leq \frac{\|c\|^2}{1-\delta_s}.$$

Let L be any set with $L \cap J = \emptyset$ and $|L| \leq s'$. One has

$$\|\eta_L\|^2 = |\langle \eta_L, \eta_L \rangle| = |\langle A_J(A_J^*A_J)^{-1}c_J, A_L\eta_L \rangle| \leq \theta_{s,s'}\|(A_J^*A_J)^{-1}c_J\|\|\eta_L\| \leq \frac{\theta_{s,s'}}{1-\delta_s}\|c_J\|\|\eta_L\|,$$

so that this gives

$$\|\eta_L\| \leq \frac{\theta_{s,s'}}{1-\delta_s}\|c_J\|. \tag{15.10}$$

Let us denote $\bar{K} = \{k \in J^c ; |\eta_k| > T\}$ where $T \stackrel{\text{def.}}{=} \frac{\theta_{s,s'}}{(1-\delta_s)\sqrt{s'}}\|c_J\|$. One necessarily has $|\bar{K}| \leq s'$, otherwise one would have, taking $K \subset \bar{K}$ the s' largest entries (in fact any s' entries)

$$\|\eta_K\| > \sqrt{s'T^2} = \frac{\theta_{s,s'}}{1-\delta_s}\|c_J\|$$

which contradicts (15.10). This shows that the entries in η_{J^c} after the rank s' are smaller than T . \square

We can now prove the Theorem 47.

Proof. We denote $I_0 = \text{supp}(x_0)$. We first consider $\eta_1 = \eta_F$, $p_1 = p_F$, and we use (6) with $c_I = \sigma_I \stackrel{\text{def.}}{=} \text{sign}(x_{0,I})$, $J = I$, $s' = s$, to control this first pre-certificate. This lemma defines a second set I_1 (denoted K in the lemma) which are the s largest entries of η_{1,I_1^c} , with

$$I_0 \cap I_1 = \emptyset, \quad |I_1| \leq s, \quad \eta_{1,I_0} = s_I, \quad \|\eta_{1,(I_0 \cup I_1)^c}\|_\infty \leq \frac{\theta_{s,s}}{1-\delta_s}, \quad \|\eta_{I_1}\| \leq \frac{\theta_{s,s}\sqrt{s}}{1-\delta_s}, \quad \|p_1\| \leq \frac{\sqrt{s}}{\sqrt{1-\delta_s}}.$$

Now we proceed recursively. Having defined the vectors ($\eta_1 = A^*p_1, \dots, \eta_n = A^*p_n$) with associated sets (I_1, \dots, I_n) , we define $\eta_{n+1} = A^*p_{n+1}$ and I_{n+1} by applying (6) to $J = I_0 \cup I_n$ with $c = (0_{I_0}, \eta_{n,I_n})$ (and thus $(2s, s)$ in place of (s, s')), which hence satisfy

$$I_{n+1} \cap (I_0 \cup I_n) = \emptyset, \quad |I_{n+1}| \leq s, \quad \eta_{n+1,I_0 \cup I_n} = (0_{I_0}, \eta_{n,I_n}), \quad \text{and}$$

$$\|\eta_{n+1,(I_0 \cup I_n \cup I_{n+1})^c}\|_\infty \leq \frac{\theta_{s,s}}{1-\delta_s}Q^n, \quad \|\eta_{n+1,I_{n+1}}\| \leq \frac{\theta_{s,s}\sqrt{s}}{1-\delta_s}Q^n, \tag{15.11}$$

where we denoted $Q \stackrel{\text{def.}}{=} \frac{\theta_{s,2s}}{1-\delta_s}$, and we have

$$\|p_{n+1}\| \leq \frac{\|\eta_{n,I_n}\|}{\sqrt{1-\delta_s}} \leq \frac{1}{\sqrt{1-\delta_s}} \frac{\theta_{s,s}\sqrt{s}}{1-\delta_s} Q^{n-1}.$$

Since $\delta_s + \theta_{s,s} < 1$, one has that $Q < 1$, and thus setting

$$p = \sum_{n=1}^{+\infty} (-1)^{n-1} p_n \quad \text{and} \quad \eta = A^* p$$

defines a convergent series. By construction, since $\eta_{1,I} = \sigma$ and $\eta_{n,I} = 0$ for $n > 1$, one has $\eta_I = \sigma_I$, thus this vector interpolates the sign vector σ . Now consider $j \in I^c$ and define $E_j \stackrel{\text{def.}}{=} \{n \geq 1 ; j \in I_n\} = \{n_1 \leq n_2 \leq n_3 \leq \dots\}$. Since $I_n \cap I_{n+1} = \emptyset$, necessary $n_{k+1} \geq n_k + 2$ (j cannot belong to two consecutive I_n). Furthermore, if $n \in E_j \Leftrightarrow j \in I_n$, then by construction

$$\eta_{n,j} = \eta_{n+1,j}$$

so that these two consecutive terms cancels out in the sum defining η [ToDo: make a drawing], which in turn can thus be written in the form

$$\eta = \sum_{n \in H} (-1)^{n-1} \eta_n.$$

The index set H is composed of $n \notin E_j$ such that $n - 1 \notin E_j$ (because otherwise one could cancel η_n from the sum). So this means that for $n \in H$, one has $j \notin (I_0 \cup I_n \cup I_{n+1})$, thus applying the property (15.11), one has

$$\forall n \in H, \quad |\eta_{j,n}| \leq \frac{\theta_{s,s}}{1-\delta_s} Q^{n-1},$$

so that

$$|\eta_j| \leq \sum_{n \in H} |\eta_{j,n}| \leq \sum_{n=1}^{+\infty} \frac{\theta_{s,s}}{1-\delta_s} Q^{n-1} = \frac{\theta_{s,s}}{1-\delta_s} \frac{1}{1 - \theta_{s,2s}(1-\delta_s)^{-1}} = \frac{\theta_{s,s}}{1-\delta_s - \theta_{s,2s}}.$$

Note that one also has the bound

$$\|p\| \leq \sum_n \|p_n\| \leq \sum_n \frac{1}{\sqrt{1-\delta_s}} \frac{\theta_{s,s}\sqrt{s}}{1-\delta_s} Q^{n-1} = \frac{\delta_{s,s}}{1-\delta_s - \theta_{s,2s}} \frac{\sqrt{s}}{\sqrt{1-\delta_s}}.$$

□

15.3.3 RIP implies stable recovery

Putting together Theorems 46 and ??, and using the general inverse problem stability theorem ??, one obtains the following recovery guarantee.

Theorem 48. *If A is a sub-Gaussian random matrix, then there exists constants (C, C') such that provided*

$$P \geq Cs \log(N/s) \tag{15.12}$$

with probability $1 - 2e^{-C'P}$ on the draw of A , one has that for every s -sparse signal x_0 ,

$$\|x_\lambda - x_0\| = O(\|w\|)$$

where x_λ is the unique solution of (11.10) with measurements $y = Ax_0 + w$ when choosing $\lambda \sim \|w\|$.

It is possible to extend this theorem when x_0 is not exactly s -sparse but only approximately. Defining $x_{0,s}$ the best s -term approximation, the easiest way to go is to write $y = Ax_{0,s} + \tilde{w}$ where $\tilde{w} = w + A(x_0 - x_{0,s})$ and applying the previous result to obtain

$$\|x_\lambda - x_{0,s}\| = O(\|w\| + \|A\| \|x_0 - x_{0,s}\|).$$

It is possible to obtain better scaling in term of the non-linear approximation error $\|x_0 - x_{0,s}\|$ by doing a more careful proof.

This theorem provides a uniform recovery guarantee, in the sense that it means

$$\mathbb{P}(\forall s-\text{sparse } x_0, x^* = x_0) \text{ goes fast to 1 when } P \rightarrow +\infty.$$

In contrast, theorem 45 proves a weaker non-uniform guarantee, in the sense that it means

$$\forall s-\text{sparse } x_0, \mathbb{P}(x^* = x_0) \text{ goes fast to 1 when } P \rightarrow +\infty.$$

The recovery performance analysis based on RIP constants proves a better scaling in term of log-factors. This is because the analysis using η_F does not only imply stable recovery, it also provides stability of the support (sparsistency). Note however that the constants involved in the RIP analysis are very large (of the order of a few hundreds, as highlighted by Figure 15.6, left, and by Figure 15.5, right), while the constant appearing in (15.7) are small and known to be sharp.

Also, one can show that having small enough RIP constant implies the existence of a valid dual certificate (but this certificate is not necessarily η_F).

15.3.4 Fourier sampling RIP

A practical issue is that doing hardware implementing random operators A is very difficult, specially if this operator is “fully” random, i.e. if its entries are i.i.d. A more practical option is to use structured sampling operator, which are in some sense “less random”. A possibility is to consider a random sub-sampling of orthogonal projection of the signal in some ortho-basis $\Xi = (\xi_\omega)_{\omega=1}^N$ of \mathbb{R}^N , so that

$$Ax \stackrel{\text{def.}}{=} (\langle x, \xi_\omega \rangle)_{\omega \in \Omega} \in \mathbb{R}^P \quad (15.13)$$

where $\Omega \subset \{1, \dots, N\}$ is drawn uniformly at random among all sets of size P . The following theorem ensure that such an operator satisfies the RIP properties for large s (proportional to P up to log factors) if the atomes φ_ω are “spread enough”, i.e. have a small magnitude, as measured by

$$\rho(\Xi) \stackrel{\text{def.}}{=} \sqrt{N} \max_{1 \leq \omega \leq N} \|\xi_\omega\|_\infty.$$

Theorem 49 (Rudelson-Vershynin). *For any $0 < c < 1$, there exists C , such that provided that*

$$P \geq C\rho(\Xi)^2 s \log(N)^4 \quad (15.14)$$

with high probability on Ω , then A defined as in (15.13) satisfies $\delta_{2s} \leq c$.

One always has $1 \leq \rho(\Xi)^2 \leq N$. The worse case is ξ_ω to be Sirac atoms (i.e. $\Xi = \text{Id}_N$), having $\rho(\Xi)^2 = N$, in which case P needs to be as large as N . In sharp contrast, optimal sampling bases are for instance Fourier atoms $\xi_\omega = (e^{\frac{2i\pi}{N}\omega n})_{n=1}^N \in \mathbb{C}^N$, for which $\rho(\Xi) = 1$ (it is also possible to consider a Hadamard basis for instance). In this case, up to log-factors, the scaling (15.14) is similar to the one for sub-Gaussian matrices (15.12).

Theorem 49 deals with cases where the data x_0 to recover is sparse in the Dirac basis. If the data f_0 is sparse in another basis $\Psi = (\psi_m)_{m=1}^N$, one can do a change of variable $x = \Psi^* f$ (x being the coefficients of f in the basis Ψ), in which case $\rho(\Xi)$ appearing in (15.14) should be replaced by the mutual coherence between the sampling and the sparsity bases

$$\rho(\Psi^* \Xi) \stackrel{\text{def.}}{=} \sqrt{N} \max_{1 \leq \omega, m \leq N} |\langle \psi_m, \xi_\omega \rangle|.$$

Good recovery performances are thus reached by (sampling,sparsity) pairs which are incoherent. The (Fourier,Dirac) pair is maximally incoherent. In contrast, Wavelet and Fourier are highly coherent. There exists explicit construction of “noiselets” bases which are almost maximally incoherent with wavelets. Note however that in contrast to Gaussian matrices, these structured measurement matrices are not universal, in the sense that there compressed sensing recovery performances depend on the sparsity basis Ψ which is used.

Chapter 16

Machine Learning

This chapter gives a rapid overview of the main concepts in machine learning. The goal is not to be exhaustive, but to highlight representative problems and insist on the distinction between unsupervised (vizualization and clustering) and supervised (regression and classification) setups. We also shed light on the tight connexions between machine learning and inverse problems.

While imaging science problems are generally concern with processing a single data (e.g. an image), machine learning problem is rather concern with analysing large collection of data. The focus (goal and performance measures) is thus radically different, but quite surprisingly, it uses very similar tools and algorithm (in particular linear models and convex optimization).

16.1 Unsupervised Learning

In unsupervised learning setups, one observes n points $(x_i)_{i=1}^n$. The problem is now to infer some properties for this points, typically for vizualization or unsupervised classification (often called clustering). For simplicity, we assume the data are points in Euclidean space $x_i \in \mathbb{R}^p$ (p is the so-called number of features). These points are conveniently stored as the rows of a matrix $X \in \mathbb{R}^{n \times d}$.

16.1.1 Dimensionality Reduction and PCA

Dimensionality reduction is useful for vizualization. It can also be understood as the problem of feature extraction (determining which are the relevant parameters) and this can be later used for doing other tasks more efficiently (faster and/or with better performances). The simplest method is the Principal Component Analysis (PCA), which performs an orthogonal linear projection on the principal axes (eigenvectors) of the covariance matrix.

The empirical mean is defined as

$$\hat{m} \stackrel{\text{def.}}{=} \frac{1}{n} \sum_{i=1}^n x_i \in \mathbb{R}^p$$

and covariance

$$\hat{C} \stackrel{\text{def.}}{=} \frac{1}{n} \sum_{i=1}^n (x_i - \hat{m})(x_i - \hat{m})^* \in \mathbb{R}^{p \times p}. \quad (16.1)$$

Denoting $\tilde{X} \stackrel{\text{def.}}{=} X - 1_p \hat{m}^*$, one has $\hat{C} = \tilde{X}^* \tilde{X} / n$.

Note that if the points $(x_i)_i$ are modelled as i.i.d. variables, and denoting \mathbf{x} one of these random variables, one has, using the law of large numbers, the almost sure convergence as $n \rightarrow +\infty$

$$\hat{m} \rightarrow m \stackrel{\text{def.}}{=} \mathbb{E}(\mathbf{x}) \quad \text{and} \quad \hat{C} \rightarrow C \stackrel{\text{def.}}{=} \mathbb{E}((\mathbf{x} - m)(\mathbf{x} - m)^*). \quad (16.2)$$

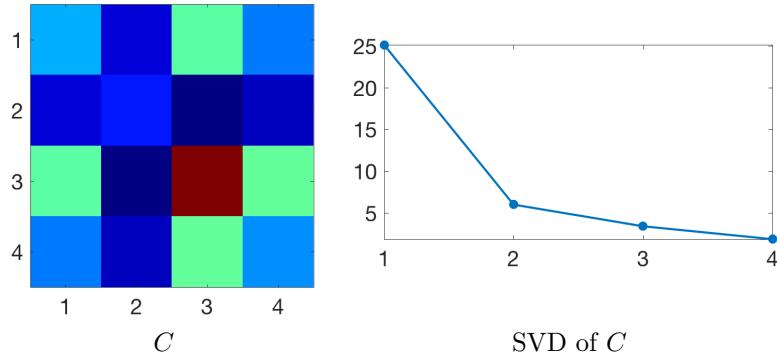


Figure 16.1: Empirical covariance of the data and its associated singular values.

Denoting μ the distribution (Radon measure) on \mathbb{R}^p of \mathbf{x} , one can alternatively write

$$m = \int_{\mathbb{R}^p} x d\mu(x) \quad \text{and} \quad C = \int_{\mathbb{R}^p} (x - m)(x - m)^* d\mu(x).$$

The PCA ortho-basis, already introduced in Section 20, corresponds to the right singular vectors of the centred data matrix, as defined using the (reduced) SVD decomposition

$$\tilde{X} = U \operatorname{diag}(\sigma) V^*$$

where $U \in \mathbb{R}^{n \times r}$ and $V \in \mathbb{R}^{p \times r}$, and where $r = \operatorname{rank}(\tilde{X}) \leq \min(n, p)$. We denote $V = (v_k)_{k=1}^r$ the orthogonal columns (which forms a orthogonal system of eigenvectors of \hat{C}), $v_k \in \mathbb{R}^p$. The intuition is that they are the main axes of “gravity” of the point cloud $(x_i)_i$ in \mathbb{R}^p . We assume the singular values are ordered, $\sigma_1 \geq \dots \geq \sigma_r$, so that the first singular values capture most of the variance of the data.

Figure 16.1 displays an example of covariance and its associated spectrum σ . The points $(x_i)_i$ correspond to the celebrated IRIS dataset¹ of Fisher. This dataset consists of 50 samples from each of three species of Iris (Iris setosa, Iris virginica and Iris versicolor). The dimensionality of the features is $p = 4$, and the dimensions corresponds to the length and the width of the sepals and petals.

The PCA dimensionality reduction embedding $x_i \in \mathbb{R}^p \mapsto z_i \in \mathbb{R}^d$ in dimension $d \leq p$ is obtained by projecting the data on the first d singular vector

$$z_i \stackrel{\text{def.}}{=} (\langle x_i - m, v_k \rangle)_{k=1}^d \in \mathbb{R}^d.$$

From these low-dimensional embedding, one can reconstruct back an approximation as

$$\tilde{x}_i \stackrel{\text{def.}}{=} m + \sum_k z_{i,k} v_k \in \mathbb{R}^p.$$

One has that $\tilde{x}_i = \operatorname{Proj}_{\tilde{T}}(x_i)$ where $\tilde{T} \stackrel{\text{def.}}{=} m + \operatorname{Span}_{k=1}^d(v_k)$ is an affine space. The following proposition shows that PCA is optimal in term of ℓ^2 distance if one consider only affine spaces.

Proposition 52. *One has*

$$(\tilde{x}, \tilde{T}) \in \operatorname{argmin}_{(\bar{x}, T)} \left\{ \sum_i \|x_i - \bar{x}_i\|^2 ; \forall i, \bar{x}_i \in \bar{T} \right\}$$

where \tilde{T} is constrained to be a d -dimensional affine space.

Figure 16.3 shows an example of PCA for 2-D and 3-D visualization.

¹https://en.wikipedia.org/wiki/Iris_flower_data_set



Figure 16.2: PCA main axes capture variance

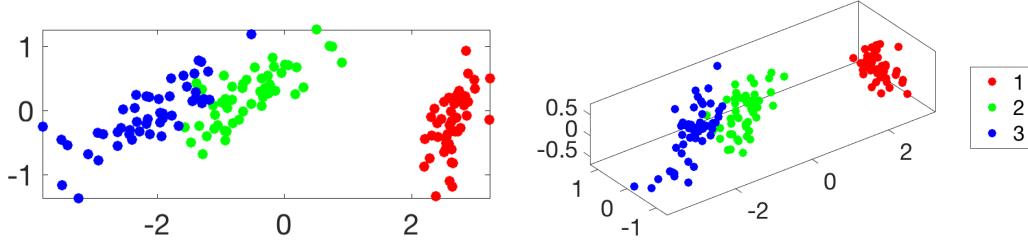


Figure 16.3: 2-D and 3-D PCA visualization of the input clouds.

16.1.2 Clustering and k -means

A typical unsupervised learning task is to infer a class label $y_i \in \{1, \dots, k\}$ for each input point x_i , and this is often called a clustering problem (since the set of points associated to a given label can be thought as a cluster).

k -means A way to infer these labels is by assuming that the clusters are compact, and optimizing some compactness criterion. Assuming for simplicity that the data are in Euclidean space (which can be relaxed to an arbitrary metric space, although the computations become more complicated), the k -means approach minimizes the distance between the points and their class centroids $c = (c_\ell)_{\ell=1}^k$, where each $c_\ell \in \mathbb{R}^p$. The corresponding variational problem becomes

$$\min_{(y,c)} \mathcal{E}(y, c) \stackrel{\text{def.}}{=} \sum_{\ell=1}^k \sum_{i:y_i=\ell} \|x_i - c_\ell\|^2.$$

The k -means algorithm can be seen as a block coordinate relaxation, which alternatively updates the class labels and the centroids. The centroids c are first initialized (more on this later), for instance, using a well-spread set of points from the samples. For a given set c of centroids, minimizing $y \mapsto \mathcal{E}(y, c)$ is obtained in closed form by assigning as class label the index of the closest centroids

$$\forall i \in \{1, \dots, n\}, \quad y_i \leftarrow \operatorname{argmin}_{1 \leq \ell \leq k} \|x_i - c_\ell\|. \quad (16.3)$$

For a given set y of labels, minimizing $c \mapsto \mathcal{E}(y, c)$ is obtained in closed form by computing the barycenter of each class

$$\forall \ell \in \{1, \dots, k\}, \quad c_\ell \leftarrow \frac{\sum_{i:y_i=\ell} x_i}{|\{i ; y_i = \ell\}|} \quad (16.4)$$

If during the iterates, one of the cluster associated to some c_ℓ becomes empty, then one can either decide to destroy it and replace k by $k - 1$, or try to “teleport” the center c_ℓ to another location (this might increase the objective function \mathcal{E} however).

Since the energy \mathcal{E} is decaying during each of these two steps, it is converging to some limit value. Since there is a finite number of possible labels assignments, it is actually constant after a finite number of iterations, and the algorithm stops.

Of course, since the energy is non-convex, little can be said about the property of the clusters output by k -means. To try to reach lower energy level, it is possible to “teleport” during the iterations centroids c_ℓ associated to clusters with high energy to locations within clusters with lower energy (because optimal solutions should somehow balance the energy).

Figure 16.5 shows an example of k -means iterations on the Iris dataset.

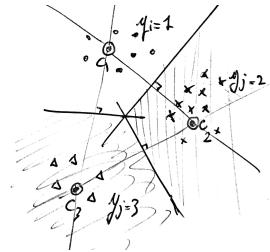


Figure 16.4: k -means clusters according to Voronoi cells.

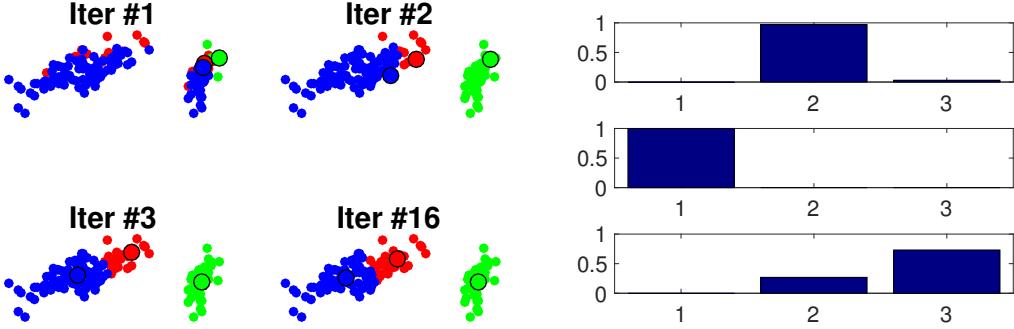


Figure 16.5: Left: iteration of k -means algorithm. Right: histogram of points belonging to each class after the k -means optimization.

k -means++ To obtain good results when using k -means, it is crucial to have an efficient initialization scheme. In practice, the best results are obtained by seeding them as far as possible from one another (a greedy strategy works great in practice).

Quite surprisingly, there exists a randomized seeding strategy which can be shown to be close to optimal in term of value of \mathcal{E} , even without running the k -means iterations (although in practice it still needs to be used to polish the results). The corresponding k -means++ initialization is obtained by selecting c_1 uniformly at random among the x_i , and then assuming c_ℓ has been seeded, drawing $c_{\ell+1}$ among the sample according to the probability $\pi^{(\ell)}$ on $\{1, \dots, n\}$ proportional to the square of the distance to the previously seeded points

$$\forall i \in \{1, \dots, n\}, \quad \pi_i^{(\ell)} \stackrel{\text{def.}}{=} \frac{d_i^2}{\sum_{j=1}^n d_j^2} \quad \text{where} \quad d_j \stackrel{\text{def.}}{=} \min_{1 \leq r \leq \ell-1} \|x_i - c_r\|.$$

This means that points which are located far away from the preciously seeded centers are more likely to be picked.

The following results, due to David Arthur and Sergei Vassilvitskii, shows that this seeding is optimal up to log factor on the energy. Note that finding a global optimum is known to be NP-hard.

Theorem 50. *For the centroids c^* defined by the k -means++ strategy, denoting y^* the associated nearest neighbor labels defined as in (16.3), one has*

$$\mathbb{E}(\mathcal{E}(y^*, c^*)) \leq 8(2 + \log(k)) \min_{(y, c)} \mathcal{E}(y, v),$$

where the expectation is on the random draws performed by the algorithm.

Lloyd algorithm and continuous densities. The k -means iterations are also called “Lloyd” algorithm, which also find applications to optimal vector quantization for compression. It can also be used in the “continuous” setting where the empirical samples $(x_i)_i$ are replaced by an arbitrary measure over \mathbb{R}^p . The energy to minimize becomes

$$\min_{(\mathcal{V}, c)} \sum_{\ell=1}^k \int_{\mathcal{V}_\ell} \|x - c_\ell\|^2 d\mu(x)$$

where $(\mathcal{V}_\ell)_\ell$ is a partition of the domain. Step (16.3) is replaced by the computation of a Voronoi cell

$$\forall \ell \in \{1, \dots, k\}, \quad \mathcal{V}_\ell \stackrel{\text{def.}}{=} \{x ; \forall \ell' \neq \ell, \|x - c_\ell\| \leq \|x - c_{\ell'}\|\}.$$

These Voronoi cells are polyhedra delimited by segments of mediatrix between centroids, and this Voronoi segmentation can be computed efficiently using tools from algorithmic geometry in low dimension. Step (16.4) are then replaced by

$$\forall \ell \in \{1, \dots, k\}, \quad c_\ell \leftarrow \frac{\int_{c_\ell} x d\mu(x)}{\int_{c_\ell} d\mu(x)}.$$

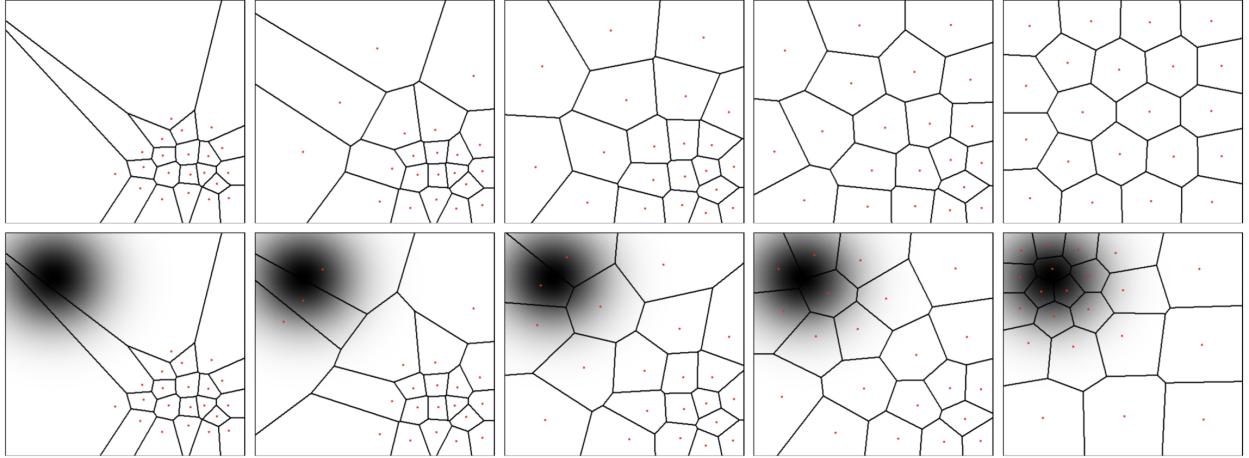


Figure 16.6: Iteration of k -means algorithm (Lloyd algorithm) on continuous densities μ . Top: uniform. Bottom: non-uniform (the densities of μ with respect to the Lebesgue measure is displayed as a grayscale image in the background).

In the case of μ being uniform distribution, optimal solution corresponds to the hexagonal lattice. Figure 16.6 displays two examples of Lloyd iterations on 2-D densities on a square domain.

16.2 Empirical Risk Minimization

Before diving into the specifics of regression and classification problems, let us give describe a generic methodology which can be applied in both case (possibly with minor modification for classification, typically considering class probabilities instead of class labels).

In order to make the problem tractable computationally, and also in order to obtain efficient prediction scores, it is important to restrict the fit to the data $y_i \approx f(x_i)$ using a “small enough” class of functions. Intuitively, in order to avoid overfitting, the “size” of this class of functions should grows with the number n of samples.

16.2.1 Empirical Risk

Denoting \mathcal{F}_n some class of functions (which depends on the number of available samples), one of the most usual way to do the learning is to perform an empirical risk minimization (ERM)

$$\hat{f} \in \operatorname{argmin}_{f \in \mathcal{F}_n} \frac{1}{n} \sum_{i=1}^n L(f(x_i), y_i). \quad (16.5)$$

Here $L : \mathcal{Y}^2 \rightarrow \mathbb{R}^+$ is the so-called loss function, and it should typically satisfies $L(y, y') = 0$ if and only if $y = y'$. The specifics of L depend on the application at hand (in particular, one should use different losses for classification and regression tasks). To highlight the dependency of \hat{f} on n , we occasionally write \hat{f}_n .

16.2.2 Prediction and Consistency

When doing a mathematically analysis, one usually assumes that (x_i, y_i) are drawn from a distribution π on $\mathcal{X} \times \mathcal{Y}$, and the large n limit defines the ideal estimator

$$\bar{f} \in \operatorname{argmin}_{f \in \mathcal{F}_\infty} \int_{\mathcal{X} \times \mathcal{Y}} L(f(x), y) d\pi(x, y) = \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim \pi}(L(f(\mathbf{x}), \mathbf{y})). \quad (16.6)$$

Intuitively, one should have $\hat{f}_n \rightarrow \bar{f}$ as $n \rightarrow +\infty$, which can be captured in expectation of the prediction error over the samples $(x_i, y_i)_i$, i.e.

$$E_n \stackrel{\text{def.}}{=} \mathbb{E}(\tilde{L}(\hat{f}_n(\mathbf{x}), \bar{f}(\mathbf{x}))) \longrightarrow 0.$$

One should be careful that here the expectation is over both \mathbf{x} (distributed according to the marginal $\pi_{\mathcal{X}}$ of π on \mathcal{X}), and also the n i.i.d. pairs $(x_i, y_i) \sim \pi$ used to define \hat{f}_n (so a better notation should rather be $(\mathbf{x}_i, \mathbf{y}_i)_i$). Here \tilde{L} is some loss function on \mathcal{Y} (one can use $\tilde{L} = L$ for instance). One can also study convergence in probability, i.e.

$$\forall \varepsilon > 0, \quad E_{\varepsilon,n} \stackrel{\text{def.}}{=} \mathbb{P}(\tilde{L}(\hat{f}_n(\mathbf{x}), \bar{f}(\mathbf{x})) > \varepsilon) \rightarrow 0.$$

If this holds, then one says that the estimation method is consistent (in expectation or in probability). The question is then to derive convergence rates, i.e. to upper bound E_n or $E_{\varepsilon,n}$ by some explicitly decay rate.

Note that when $\tilde{L}(y, y') = |y - y'|^r$, then convergence in expectation is stronger (implies) than convergence in probability since using Markov's inequality

$$E_{\varepsilon,n} = \mathbb{P}(|\hat{f}_n(\mathbf{x}) - f(\mathbf{x})|^r \geq \varepsilon) \leq \frac{1}{\varepsilon} \mathbb{E}(|\hat{f}_n(\mathbf{x}) - f(\mathbf{x})|^r) = \frac{E_n}{\varepsilon}.$$

16.2.3 Parametric Approaches and Regularization

Instead of directly defining the class \mathcal{F}_n and using it as a constraint, it is possible to rather use a penalization using some prior to favor “simple” or “regular” functions. A typical way to achieve this is by using a parametric model $y \approx f(x, \beta)$ where $\beta \in \mathcal{B}$ parametrizes the function $f(\cdot, \beta) : \mathcal{X} \rightarrow \mathcal{Y}$. The empirical risk minimization procedure (16.5) now becomes

$$\hat{\beta} \in \operatorname{argmin}_{\beta \in \mathcal{B}} \frac{1}{n} \sum_{i=1}^n L(f(x_i, \beta), y_i) + \lambda_n J(\beta). \quad (16.7)$$

where J is some regularization function, for instance $J = \|\cdot\|_2^2$ (to avoid blowing-up of the parameter) or $J = \|\cdot\|_1$ (to perform model selection, i.e. using only a sparse set of feature among a possibly very large pool of p features). Here $\lambda_n > 0$ is a regularization parameter, and it should tend to 0 when $n \rightarrow +\infty$.

Then one similarly defines the ideal parameter $\bar{\beta}$ as in (16.6) so that the limiting estimator as $n \rightarrow +\infty$ is of the form $\bar{f} = f(\cdot, \bar{\beta})$ for $\bar{\beta}$ defined as

$$\bar{\beta} \in \operatorname{argmin}_{\beta} \int_{\mathcal{X} \times \mathcal{Y}} L(f(x, \beta), y) d\pi(x, y) = \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim \pi}(L(f(\mathbf{x}, \beta), \mathbf{y})). \quad (16.8)$$

Prediction vs. estimation risks. In this parametric approach, one could be interested in also studying how close $\hat{\beta}$ is to $\bar{\beta}$. This can be measured by controlling how fast some estimation error $\|\hat{\beta} - \bar{\beta}\|$ (for some norm $\|\cdot\|$) goes to zero. Note however that in most cases, controlling the estimation error is more difficult than doing the same for the prediction error. In general, doing a good parameter estimation implies doing a good prediction, but the converse is not true.

16.2.4 Testing Set and Cross-validation

It is not possible to access E_n or $E_{\varepsilon,n}$ because the optimal \bar{f} is unknown. In order to tune some parameters of the methods (for instance the regularization parameter λ), one rather wants to minimize the risk $\mathbb{E}(L(\hat{f}(\mathbf{x}), \mathbf{y}))$, but this one should not be approximated using the training samples $(x_i, y_i)_i$.

One thus rather resorts to a second set of data $(\bar{x}_j, \bar{y}_j)_{j=1}^{\bar{n}}$, called “testing set”. From a modelling perspective, this set should also be distributed i.i.d. according to π . The validation (or testing) risk is then

$$R_{\bar{n}} = \frac{1}{\bar{n}} \sum_{j=1}^{\bar{n}} L(\hat{f}(\bar{x}_j), \bar{y}_j) \quad (16.9)$$

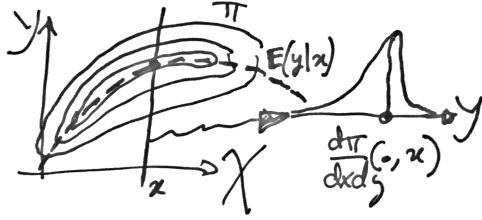


Figure 16.8: Conditional expectation.

which converges to $\mathbb{E}(L(\hat{f}(\mathbf{x}), \mathbf{y}))$ for large \bar{n} . Minimizing $R_{\bar{n}}$ to setup to some meta-parameter of the method (for instance the regularization parameter λ_n) is called “cross validation” in the literature.

16.3 Supervised Learning: Regression

In supervised learning, one has access to training data, consisting in pairs $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$. Here $\mathcal{X} = \mathbb{R}^p$ for simplicity. The goal is to infer some relationship, typically of the form $y_i \approx f(x_i)$ for some deterministic function $f : \mathcal{X} \rightarrow \mathcal{Y}$, in order, when some un-observed data x without associated value in \mathcal{Y} is given, to be able to “predict” the associated value using $y = f(x)$.

If the set \mathcal{Y} is discrete and finite, then this problem is called a supervised classification problem, and this is studied in Section 16.4. The simplest example being the binary classification case, where $\mathcal{Y} = \{0, 1\}$. It finds applications for instance in medical diagnosis, where $y_i = 0$ indicates a healthy subject, why $y_i = 1$ a pathological one. If \mathcal{Y} is continuous (the typical example being $\mathcal{Y} = \mathbb{R}$), then this problem is called a regression problem.

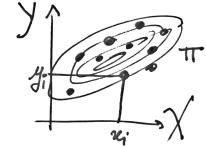


Figure 16.7: Probabilistic modelling.

16.3.1 Linear Regression

We now specialize the empirical risk minimization approach to regression problems, and even more specifically, we consider $\mathcal{Y} = \mathbb{R}$ and use a quadratic loss $L(y, y') = \frac{1}{2}|y - y'|^2$.

Note that non-linear regression can be achieved using approximation in dictionary (e.g. polynomial interpolation), and this is equivalent to using lifting to a higher dimensional space, and is also equivalent to kernelization technics studied in Section 16.5.

Least square and conditional expectation. If one do not put any constraint on f (beside being measurable), then the optimal limit estimator $\bar{f}(x)$ defined in (16.6) is simply averaging the values y sharing the same x , which is the so-called conditional expectation. Assuming for simplicity that π has some density $\frac{d\pi}{dxdy}$ with respect to a tensor product measure $dxdy$ (for instance the Lebegues mesure), one has

$$\forall x \in \mathcal{X}, \quad \bar{f}(x) = \mathbb{E}(\mathbf{y}|\mathbf{x} = x) = \frac{\int_{\mathcal{Y}} y \frac{d\pi}{dxdy}(x, y) dy}{\int_{\mathcal{Y}} \frac{d\pi}{dxdy}(x, y) dy}$$

where (\mathbf{x}, \mathbf{y}) are distributed according to π .

In the simple case where \mathcal{X} and \mathcal{Y} are discrete, denoting $\pi_{x,y}$ the probability of $(\mathbf{x} = x, \mathbf{y} = y)$, one has

$$\forall x \in \mathcal{X}, \quad \bar{f}(x) = \frac{\sum_y y \pi_{x,y}}{\sum_y \pi_{x,y}}$$

and it is unspecified if the marginal of π along \mathcal{X} vanishes at x .

The main issue is that this estimator \hat{f} performs poorly on finite samples, and $f(x)$ is actually undefined if there is no sample x_i equal to x . This is due to the fact that the class of functions is too large, and one should impose some regularity or simplicity on the set of admissible f .

Penalized linear models. A very simple class of models is obtained by imposing that f is linear, and set $f(x, \beta) = \langle x, \beta \rangle$, for parameters $\beta \in \mathcal{B} = \mathbb{R}^p$. Note that one can also treat this way affine functions by remarking that $\langle x, \beta \rangle + \beta_0 = \langle (x, 1), (\beta, \beta_0) \rangle$ and replacing x by $(x, 1)$. So in the following, without loss of generality, we only treat the vectorial (non-affine) case.

Under the square loss, the regularized ERM (16.7) is conveniently rewritten as

$$\hat{\beta} \in \operatorname{argmin}_{\beta \in \mathcal{B}} \frac{1}{2} \langle \hat{C}\beta, \beta \rangle - \langle \hat{u}, \beta \rangle + \lambda_n J(\beta) \quad (16.10)$$

where we introduced the empirical correlation (already introduced in (16.1)) and observations

$$\hat{C} \stackrel{\text{def.}}{=} \frac{1}{n} X^* X = \frac{1}{n} \sum_{i=1}^n x_i x_i^* \quad \text{and} \quad \hat{u} \stackrel{\text{def.}}{=} \frac{1}{n} \sum_{i=1}^n y_i x_i = \frac{1}{n} X^* y \in \mathbb{R}^p.$$

As $n \rightarrow 0$, under weak condition on π , one has with the law of large numbers the almost sure convergence

$$\hat{C} \rightarrow C \stackrel{\text{def.}}{=} \mathbb{E}(\mathbf{x}^* \mathbf{x}) \quad \text{and} \quad \hat{u} \rightarrow u \stackrel{\text{def.}}{=} \mathbb{E}(\mathbf{y} \mathbf{x}). \quad (16.11)$$

When considering $\lambda_n \rightarrow 0$, in some cases, one can shows that in the limit $n \rightarrow +\infty$, one retrieves the following ideal parameter

$$\bar{\beta} \in \operatorname{argmin}_{\beta} \{J(\beta); C\beta = u\}.$$

Problem (16.10) is equivalent to the regularized resolution of inverse problems (10.9), with \hat{C} in place of Φ and \hat{u} in place of $\Phi^* y$. The major, and in fact only difference between machine learning and inverse problems is that the linear operator is also noisy since \hat{C} can be viewed as a noisy version of C . The “noise level”, in this setting, is $1/\sqrt{n}$ in the sense that

$$\mathbb{E}(\|\hat{C} - C\|) \sim \frac{1}{\sqrt{n}} \quad \text{and} \quad \mathbb{E}(\|\hat{u} - u\|) \sim \frac{1}{\sqrt{n}},$$

under the assumption that $\mathbb{E}(\mathbf{y}^4) < +\infty$, $\mathbb{E}(\|\mathbf{x}\|^4) < +\infty$ so ensure that one can use the central limit theorem on \mathbf{x}^2 and \mathbf{xy} . Note that, although we use here linear estimator, one does not need to assume a “linear” relation of the form $\mathbf{y} = \langle \mathbf{x}, \beta \rangle + w$ with a noise w independent from \mathbf{x} , but rather hope to do “as best as possible”, i.e. estimate a linear model as close as possible to $\bar{\beta}$.

The general take home message is that it is possible to generalize Theorems 31, 42 and 43 to cope with the noise on the covariance matrix to obtain prediction convergence rates of the form

$$\mathbb{E}(|\langle \hat{\beta}, \mathbf{x} \rangle - \langle \bar{\beta}, \mathbf{x} \rangle|^2) = O(n^{-\kappa})$$

and estimation rates of the form

$$\mathbb{E}(\|\hat{\beta} - \bar{\beta}\|^2) = O(n^{-\kappa'}),$$

under some suitable source condition involving C and u . Since the noise level is roughly $n^{-\frac{1}{2}}$, the ideal cases are when $\kappa = \kappa' = 1$, which is the so-called linear rate regime. It is also possible to derive sparsistency theorems by extending theorem 44. For the sake of simplicity, we now focus our attention to quadratic penalization, which is by far the most popular regression technic. It is fair to say that sparse (e.g. ℓ^1 type) methods are not routinely used in machine learning, because they typically do not improve the estimation performances, and are mostly useful to do model selection (isolate a few useful coordinates in the features). This is in sharp contrast with the situation for inverse problems in imaging sciences, where sparsity is a key feature because it corresponds to a modelling assumption on the structure of the data to recover.

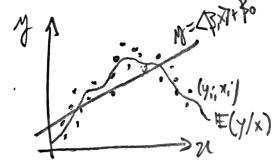


Figure 16.9: Linear regression.

Ridge regression (quadratic penalization). For $J = \|\cdot\|^2/2$, the estimator (16.10) is obtained in closed form as

$$\hat{\beta} = (X^*X + n\lambda_n \text{Id}_p)^{-1}X^*y = (\hat{C} + n\lambda_n \text{Id})^{-1}\hat{u}. \quad (16.12)$$

This is often called ridge regression in the literature. Note that thanks to the Woodbury formula, this estimator can also be re-written as

$$\hat{\beta} = X^*(XX^* + n\lambda_n \text{Id}_n)^{-1}y. \quad (16.13)$$

If $n \gg p$ (which is the usual setup in machine learning), then (16.13) is preferable. In some cases however (in particular when using RKHS technics), it makes sense to consider very large p (even infinite dimensional), so that (16.12) must be used.

If $\lambda_n \rightarrow 0$, then using (16.11), one has the convergence in expectation and probability

$$\hat{\beta} \rightarrow \bar{\beta} = C^+u.$$

Theorems 31 and 42 can be extended to this setting and one obtains the following result.

Theorem 51. If

$$\bar{\beta} = C^\gamma z \quad \text{where} \quad \|z\| \leq \rho \quad (16.14)$$

for $0 < \gamma \leq 2$, then

$$\mathbb{E}(\|\hat{\beta} - \bar{\beta}\|^2) \leq C\rho^{2\frac{1}{\gamma+1}}n^{-\frac{\gamma}{\gamma+1}} \quad (16.15)$$

for a constant C depending only on γ .

It is important to note that, since $\bar{\beta} = C^+u$, the source condition (16.14) is always satisfied. What trully matters here is that the rate (16.15) does not depend on the dimension p of the features, but rather only on ρ , which can be much smaller. This theoretical analysis actually works perfectly fine in infinite dimension $p = \infty$ (which is the setup considered when dealing with RKHS bellow).

16.4 Supervised Learning: Classification

We now focus on the case of discrete labels $y_i \in \mathcal{Y} = \{1, \dots, k\}$, which is the classification setup. We now detail two popular classification methods: nearest neighbors and logistic classification. It is faire to say that a significant part of successful applications of machine learning technics consists in using one of these two approaches, which should be considered as baselines. Note that the nearest neighbors approach, while popular for classification could as well be used for regression.

16.4.1 Nearest Neighbors Classification

Probably the simplest method for supervised classification is R nearest neighbors (R -NN), where R is a parameter indexing the number of neighbors. Increasing R is important to cope with noise and obtain smoother decision boundaries, and hence better generalization performances. It should typically decreases as the number of training samples n increases. Despite its simplicity, k -NN is surprisingly successful in practice, specially in low dimension p .

The class $\hat{f}(x) \in \mathcal{Y}$ predicted for a point x is the one which is the most represented among the R points $(x_i)_i$ which are the closest to x . This is a non-parametric method, and \hat{f} depends on the numbers n of samples (its “complexity” increases with n).

One first compute the Euclidean distance between this x and all other x_i in the training set. Sorting the distances generates an indexing σ (a permutation of $\{1, \dots, n\}$) such that

$$\|x - x_{\sigma(1)}\| \leq \|x - x_{\sigma(2)}\| \leq \dots \leq \|x - x_{\sigma(n)}\|.$$

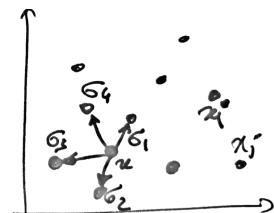


Figure 16.10: Nearest neighbors.

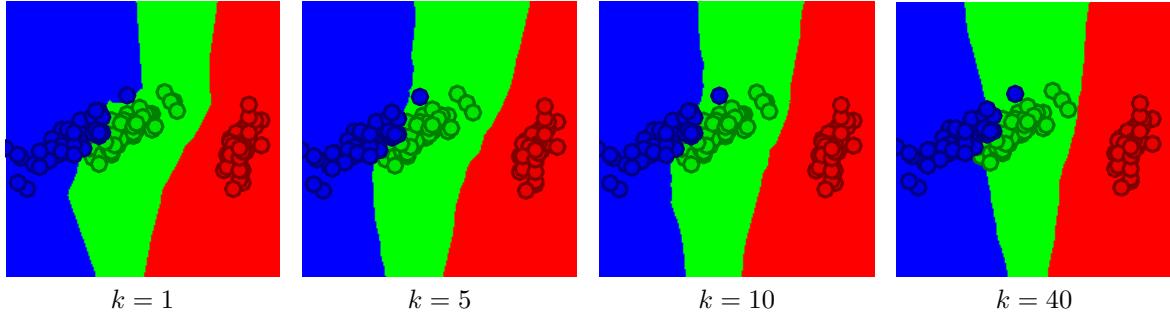


Figure 16.11: k -nearest-neighbor classification boundary function.

For a given R , one can compute the “local” histogram of classes around x

$$h_\ell(x) \stackrel{\text{def.}}{=} \frac{1}{R} \{ i ; y_{\sigma(i)} \in \{1, \dots, R\} \}.$$

The decision class for x is then a maximum of the histogram

$$\hat{f}(x) \in \operatorname{argmax}_\ell h_\ell(x).$$

In practice, the parameter R can be setup through cross-validation, by minimizing the testing risk R_n defined in (16.9), which typically uses a 0-1 loss for counting the number of mis-classifications

$$R_n \stackrel{\text{def.}}{=} \sum_{j=1}^n \delta(\bar{y}_j - \hat{f}(x_i))$$

where $\delta(0) = 0$ and $\delta(s) = 1$ if $s \neq 0$. Of course the method extends to arbitrary metric space in place of Euclidean space \mathbb{R}^p for the features. Note also that instead of explicitly sorting all the Euclidean distance, one can use fast nearest neighbor search methods.

Figure 16.11 shows, for the IRIS dataset, the classification domains (i.e. $\{x ; f(x) = \ell\}$ for $\ell = 1, \dots, k$) using a 2-D projection for visualization. Increasing R leads to smoother class boundaries.

16.4.2 Two Classes Logistic Classification

The logistic classification method (for 2 classes and multi-classes) is one of the most popular (maybe “the” most) popular machine learning technics. This is due in large part of both its simplicity and because it also outputs a probability of belonging to each class (in place of just a class membership), which is useful to (somehow ...) quantify the “uncertainty” of the estimation. Note that logistic classification is actually called “logistic regression” in the literature, but it is in fact a classification method.

Another very popular (and very similar) approach is support vector machine (SVM). SVM is both more difficult to train (because the loss is non-smooth) and does not give class membership probability, so the general rule of thumb is that logistic classification is preferable.

To simplify the expression, classes indexes are set to $y_i \in \mathcal{Y} = \{-1, 1\}$ in the following. Note that for logistic classification, the prediction function $f(\cdot, \beta) \in [0, 1]$ outputs the probability of belonging to the first class, and not the class indexes. With a slight abuse of notation, we still denote it as f .

Logistic classification can be understood as a linear model as introduced in Section 16.3.1, although the decision function $f(\cdot, \beta)$ is not linear. Indeed, one needs to “remap” the linear value $\langle x, \beta \rangle$ in the interval $[0, 1]$. In logistic classification, we define the predicted probability of x belonging to class with label -1 as

$$f(x, \beta) \stackrel{\text{def.}}{=} \theta(\langle x, \beta \rangle) \quad \text{where} \quad \theta(s) \stackrel{\text{def.}}{=} \frac{e^s}{1 + e^s} = (1 + e^{-s})^{-1}, \quad (16.16)$$

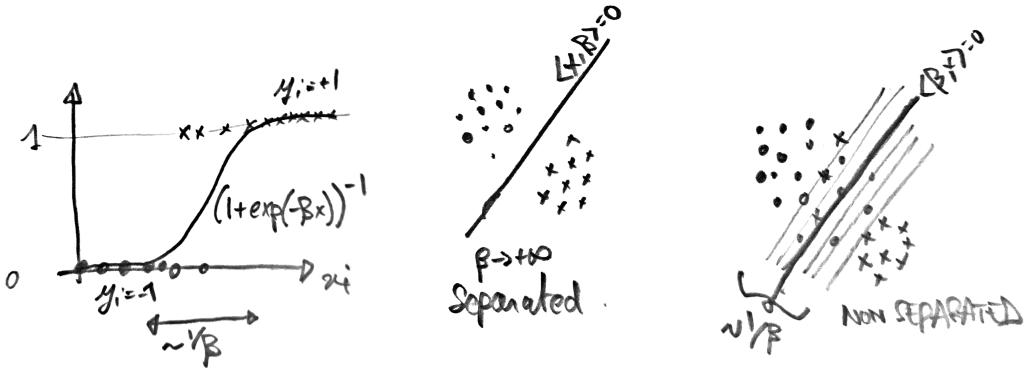


Figure 16.12: 1-D and 2-D logistic classification, showing the impact of $\|\beta\|$ on the sharpness of the classification boundary.

which is often called the “logit” model. Using a linear decision model might seem overly simplistic, but in high dimension p , the number of degrees of freedom is actually enough to reach surprisingly good classification performances. Note that the probability of belonging to the second class is $1 - f(x, \beta) = \theta(-s)$. This symmetry of the θ function is important because it means that both classes are treated equally, which makes sense for “balanced” problem (where the total mass of each class are roughly equal).

Intuitively, $\beta/\|\beta\|$ controls the separating hyperplane direction, while $1/\|\beta\|$ is roughly the fuzziness of the separation. As $\|\beta\| \rightarrow +\infty$, one obtains sharp devision boundary, and logistic classification ressembles SVM.

Note that $f(x, \beta)$ can be interpreted as a single layer perceptron with a logistic (sigmoid) rectifying unit, more details on this in Chapter 17.

Since the (x_i, y_i) are modelled as i.i.d. variables, it makes sense to define $\hat{\beta}$ from the observation using a maximum likelihood, assuming that each y_i conditioned on x_i is a Bernoulli variable with associated probability $(p_i, 1 - p_i)$ with $p_i = f(x_i, \beta)$. The probability of observing $y_i \in \{0, 1\}$ is thus, denoting $s_i = \langle x_i, \beta \rangle$

$$\mathbb{P}(\mathbf{y} = y_i | \mathbf{x} = x_i) = p_i^{1-\bar{y}_i} (1 - p_i)^{\bar{y}_i} = \left(\frac{e^{s_i}}{1 + e^{s_i}} \right)^{1-\bar{y}_i} \left(\frac{1}{1 + e^{s_i}} \right)^{\bar{y}_i}$$

where we denoted $\bar{y}_i = \frac{y_i+1}{2} \in \{0, 1\}$.

One can then minimize minus the sum of the log of the likelihoods, which reads

$$\hat{\beta} \in \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} - \sum_{i=1}^n \log(\mathbb{P}(\mathbf{y} = y_i | \mathbf{x} = x_i)) = \sum_{i=1}^n -(1 - \bar{y}_i) \log \frac{e^{s_i}}{1 + e^{s_i}} - \bar{y}_i \log \frac{1}{1 + e^{s_i}}$$

Some algebraic manipulations shows that this is equivalent to an ERM-type form (16.7) with a logistic loss function

$$\hat{\beta} \in \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} E(\beta) = \frac{1}{n} \sum_{i=1}^n L(\langle x_i, \beta \rangle, y_i) \quad (16.17)$$

where the logistic loss reads

$$L(s, y) \stackrel{\text{def.}}{=} \log(1 + \exp(-sy)). \quad (16.18)$$

Problem (16.17) is a smooth convex minimization. If X is injective, E is also strictly convex, hence it has a single global minimum.

Figure (16.13) compares the binary (ideal) 0-1 loss, the logistic loss and the hinge loss (the one used for SVM).

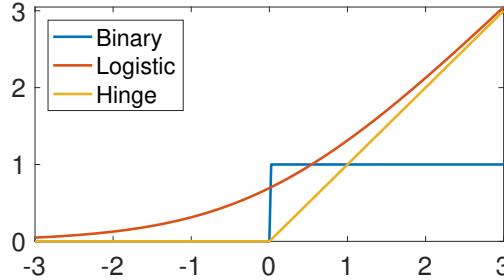


Figure 16.13: Comparison of loss functions.

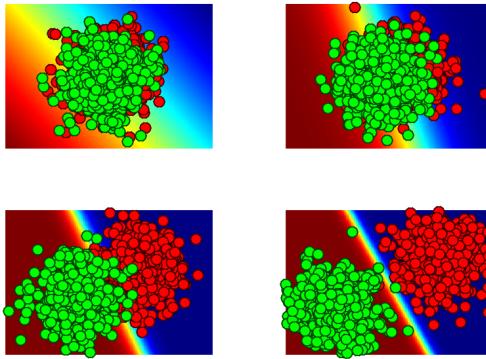


Figure 16.14: Influence on the separation distance between the class on the classification probability.

Re-writing the energy to minimize

$$E(\beta) = \mathcal{L}(X\beta, y) \quad \text{where} \quad \mathcal{L}(s, y) = \frac{1}{n} \sum_i L(s_i, y_i),$$

its gradient reads

$$\nabla E(\beta) = X^* \nabla \mathcal{L}(X\beta, y) \quad \text{where} \quad \nabla \mathcal{L}(s, y) = \frac{y}{n} \odot \theta(-y \odot s),$$

where \odot is the pointwise multiplication operator, i.e. $.*$ in Matlab. Once $\beta^{(\ell=0)} \in \mathbb{R}^p$ is initialized (for instance at 0_p), one step of gradient descent (13.2) reads

$$\beta^{(\ell+1)} = \beta^{(\ell)} - \tau_\ell \nabla E(\beta^{(\ell)}).$$

To understand the behavior of the method, in Figure 16.14 we generate synthetic data distributed according to a mixture of Gaussian with an overlap governed by an offset ω . One can display the data overlaid on top of the classification probability, this highlight the separating hyperplane $\{x ; \langle \beta, x \rangle = 0\}$.

16.4.3 Multi-Classes Logistic Classification

The logistic classification method is extended to an arbitrary number k of classes by considering a family of weight vectors $\beta = (\beta_\ell)_{\ell=1}^k$, which are conveniently stored as columns of a matrix $\beta \in \mathbb{R}^{p \times k}$.

This allows one to model probabilistically the belonging of a point $x \in \mathbb{R}^p$ to the classes using the logit model

$$f(x, \beta) = \left(\frac{e^{-\langle x, \beta_\ell \rangle}}{\sum_m e^{-\langle x, \beta_m \rangle}} \right)_\ell$$

This vector $h(x) \in [0, 1]^k$ describes the probability of x belonging to the different classes, and $\sum_\ell h(x)_\ell = 1$.

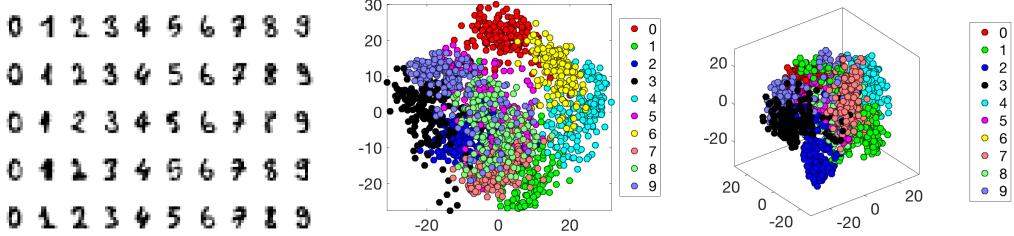


Figure 16.15: 2-D and 3-D PCA visualization of the digits images.

The computation of β is obtained by solving a maximum likelihood estimator

$$\max_{\beta \in \mathbb{R}^{p \times k}} \frac{1}{n} \sum_{i=1}^n \log(f(x_i, \beta)_{y_i})$$

where we recall that $y_i \in \mathcal{Y} = \{1, \dots, k\}$ is the class index of the point x_i .

This is conveniently rewritten as

$$\min_{\beta \in \mathbb{R}^{p \times k}} \mathcal{E}(\beta) \stackrel{\text{def.}}{=} \sum_i \text{LSE}(X\beta)_i - \langle X\beta, D \rangle$$

where $D \in \{0, 1\}^{n \times k}$ is the binary class index matrices

$$D_{i,\ell} = \begin{cases} 1 & \text{if } y_i = \ell, \\ 0 & \text{otherwise.} \end{cases}$$

and LSE is the log-sum-exp operator

$$\text{LSE}(S) = \log \left(\sum_{\ell} \exp(S_{i,\ell}) \right) \in \mathbb{R}^n.$$

Note that in the case of $k = 2$ classes $\mathcal{Y} = \{-1, 1\}$, this model can be shown to be equivalent to the two-classes logistic classifications methods exposed in Section (16.4.2), with a solution vector being equal to $\beta_1 - \beta_2$ (so it is computationally more efficient to only consider a single vector as we did).

The computation of the LSE operator is unstable for large value of $S_{i,\ell}$ (numerical overflow, producing NaN), but this can be fixed by subtracting the largest element in each row, since

$$\text{LSE}(S + a) = \text{LSE}(S) + a$$

if a is constant along the rows. This is often referred to as the “LSE trick” and is very important to use in practice (in particular if some classes are well separated, since the corresponding β_ℓ vector might become large).

The gradient of the LSE operator is the soft-max operator

$$\nabla \text{LSE}(S) = \text{SM}(S) \stackrel{\text{def.}}{=} \left(\frac{e^{S_{i,\ell}}}{\sum_m e^{S_{i,m}}} \right)$$

Similarly to the LSE, it needs to be stabilized by subtracting the maximum value along rows before computation.

Once D matrix is computed, the gradient of \mathcal{E} is computed as

$$\nabla \mathcal{E}(\beta) = \frac{1}{n} X^* (\text{SM}(X\beta) - D).$$

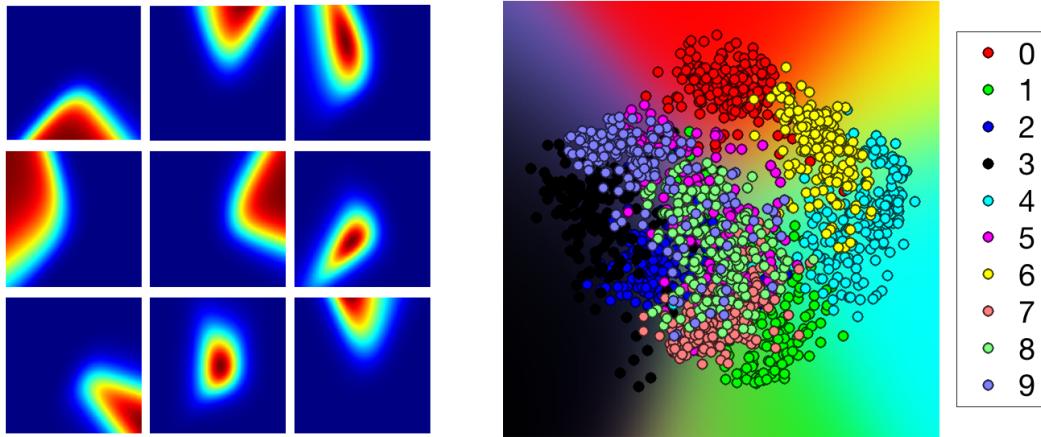


Figure 16.16: Results of digit classification Left: probability $h(x)_l$ of belonging to each of the 9 first classes (displayed over a 2-D PCA space). Right: colors reflect probability $h(x)$ of belonging to classes.

and one can minimize \mathcal{E} using for instance a gradient descent scheme.

To illustrate the method, we use a dataset of n images of size $p = 8 \times 8$, representing digits from 0 to 9 (so there are $k = 10$ classes). Figure 16.15 displays a few representative examples as well as 2-D and 3-D PCA projections. Figure (16.16) displays the “fuzzy” decision boundaries by visualizing the value of $h(x)$ using colors on an image regular grid.

16.5 Kernel Methods

Linear methods are parametric and cannot generate complex regression or decision functions. The linearity assumption is often too restrictive and in some case the geometry of the input functions or classes is not well captured by these models. In many cases (e.g. for text data) the input data is not even in a linear space, so one cannot even apply these model.

Kernel method is a simple yet surprisingly powerful remedy for these issues. By lifting the features to a high dimensional embedding space, it allows to generate non-linear decision and regression functions, but still re-use the machinery (linear system solvers or convex optimization algorithm) of linear models. Also, by the use of the so-called “kernel-trick”, the computation cost does not depend on the dimension of the embedding space, but of the number n of points. It is the perfect example of so-called “non-parametric” methods, where the number of degrees of freedom (number of variables involved when fitting the model) grows with the number of samples. This is often desirable when one wants the precisions of the result to improve with n , and also to mathematically model the data using “continuous” models (e.g. functional spaces such as Sobolev).

The general rule of thumb is that any machine learning algorithm which only makes use of inner products (and not directly of the features x_i themselves) can be “kernelized” to obtain a non-parametric algorithm. This is for instance the case for linear and nearest neighbor regression, SVM classification, logistic classification and PCA dimensionality reduction. We first explain the general machinery, and instantiate this in two representative setup (ridge regression, nearest-neighbor regression and logistic classification)

16.5.1 Reproducing Kernel Hilbert Space

We consider a general lifting $\varphi : x \in \mathbb{R}^p \rightarrow \bar{x} = \varphi(x) \in \mathcal{H}$ where \mathcal{H} is a Hilbert space. We denote $\bar{X} = (\bar{x}_i^* \stackrel{\text{def}}{=} \varphi(x_i)^*)_{i=1}^n$ the “matrix” where each row is a lifted feature $\varphi(x_i)$. For instance, if $\mathcal{H} = \mathbb{R}^{\bar{p}}$ is finite dimensional, one can view this as a matrix $\bar{X} \in \mathbb{R}^{n \times \bar{p}}$, but the rows of the matrix can be infinite dimensional vectors.

The following proposition is the crux of the RKHS approaches. When using a regularization which is a squared Euclidean norm, $\|\cdot\|_{\mathcal{H}}^2$, it states that the solutions actually belongs to a data-driven linear sub-space of dimension n . Although the proof is straightforward, its implications are very profound, since it leads to tractable algorithms even when using an infinite dimensional lifting space \mathcal{H} . as we elaborate next. It is often called the “representer” theorem in RKHS theory.

Proposition 53. *The solution $\beta^* \in \mathcal{H}$ of*

$$\min_{\beta \in \mathcal{H}} \mathcal{L}(\bar{X}\beta, y) + \frac{\lambda}{2} \|\beta\|_{\mathcal{H}}^2 \quad (16.19)$$

is unique and can be written as

$$\beta = \bar{X}^* q^* = \sum_i q_i^* \varphi(x_i) \in \mathcal{H} \quad (16.20)$$

where $q \in \mathbb{R}^N$ is a solution of

$$\min_{q \in \mathbb{R}^N} \mathcal{L}(Kp, y) + \frac{\lambda}{2} \langle Kq, q \rangle_{\mathbb{R}^n} \quad (16.21)$$

where we defined

$$K \stackrel{\text{def.}}{=} \bar{X}^* \bar{X} = (\langle \varphi(x_i), \varphi(x_j) \rangle_{\mathcal{H}})_{i,j=1}^n \in \mathbb{R}^{n \times n}.$$

Proof. The first order condition of (16.19) reads

$$0 \in \bar{X}^* \partial \mathcal{L}(\bar{X}^* \beta^*, y) + \lambda \beta^* = 0$$

i.e. there exists $u^* \in \partial \mathcal{L}(\bar{X}^* \beta^*, y)$ such that

$$\beta^* = -\frac{1}{\lambda} \bar{X}^* u^* \in \text{Im}(\bar{X}^*)$$

which is the desired result. \square

Equation (16.20) expresses the fact that the solution only lives in the n dimensional space spanned by the lifted observed points $\varphi(x_i)$. A crucial by product of this results is that all the computations as well as the prediction procedure can be expressed using the so-called kernel $\kappa : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ associated to φ

$$\forall (x, x') \in \mathcal{X}^2, \quad \kappa(x, x') \stackrel{\text{def.}}{=} \langle \varphi(x), \varphi(x') \rangle_{\mathcal{H}}.$$

Indeed, one has $K = (\kappa(x_i, x_j))_{i,j}$ and the prediction operator, as a function of x and not $\varphi(x)$ (which makes it non-linear) is a weighted sum of kernel functions centered at the x_i

$$\langle \bar{x}, \beta^* \rangle_{\mathcal{H}} = \sum_{i=1}^n p_i^* \langle \varphi(x), \varphi(x_i) \rangle_{\mathcal{H}} = \sum_{i=1}^n p_i^* \kappa(x_i, x). \quad (16.22)$$

This means that one actually never needs to manipulate quantities in \mathcal{H} (which can be infinite dimensional).

But more importantly, one can reverse the process, and instead of starting from a lifting φ , directly consider a kernel $\kappa(x, x')$. This is actually the way this is done in practice, since it is easier to design kernel and think in term of their geometrical properties (for instance, one can sum kernels). In order for this to make sense, the kernel needs to be positive definite, i.e. one should have that $(\kappa(x_i, x_j))_{i,j}$ should be symmetric positive definite for any choice of sampling points $(x_i)_i$. This can be shown to be equivalent to the existence of a lifting function φ generating the kernel. Note that such a kernel can be defined on arbitrary space (not necessarily Euclidean).

When using the linear kernel $\kappa(x, y) = \langle x, y \rangle$, one retrieves the linear models studied in the previous section, and the lifting is trivial $\varphi(x) = x$. A family of popular kernels are polynomial ones, $\kappa(x, x') =$

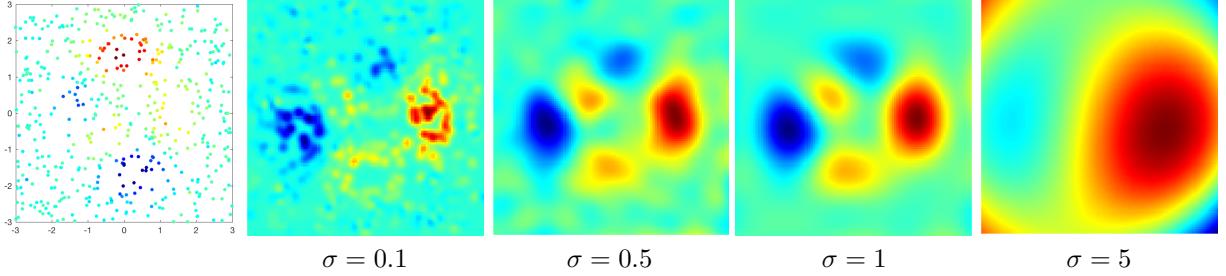


Figure 16.17: Regression using a Gaussian kernel.

$(\langle x, y \rangle + c)^a$ for $a \in \mathbb{N}^*$ and $c > 0$, which corresponds to a lifting in finite dimension. For instance, for $a = 2$ and $p = 2$, one has a lifting in dimension 6

$$\kappa(x, x') = (x_1 x'_1 + x_1 x'_1 + c)^2 = \langle \varphi(x), \varphi(x') \rangle \quad \text{where} \quad \varphi(x) = (x_1^2, x_2^2, \sqrt{2}x_1 x_2, \sqrt{2}cx_1, \sqrt{2}cx_2, c)^* \in \mathbb{R}^6.$$

In Euclidean spaces, the gaussian kernel is the most well known and used kernel

$$\kappa(x, y) \stackrel{\text{def.}}{=} e^{-\frac{\|x-y\|^2}{2\sigma^2}}. \quad (16.23)$$

The bandwidth parameter $\sigma > 0$ is crucial and controls the locality of the model. It is typically tuned through cross validation. It corresponds to an infinite dimensional lifting $x \mapsto e^{-\frac{\|x-\cdot\|^2}{2(\sigma/2)^2}} \in L^2(\mathbb{R}^p)$. Another related popular kernel is the Laplacian kernel $\exp(-\|x-y\|/\sigma)$. More generally, when considering translation invariant kernels $\kappa(x, x') = k(x - x')$ on \mathbb{R}^p , being positive definite is equivalent to $\hat{k}(\omega) > 0$ where \hat{k} is the Fourier transform, and the associated lifting is obtained by considering $\hat{h} = \sqrt{\hat{k}}$ and $\varphi(x) = h(x - \cdot) \in L^2(\mathbb{R}^p)$.

16.5.2 Examples of Kernelized Algorithms

We illustrate this general machinery by applying it to three typical problems.

Kernelized ridge regression. The simplest instantiation of this kernelization approach is when using the square loss $L(y, y') = \frac{1}{2}|y - y'|^2$, which is the ridge regression problem studied in Section 16.3.1. The obtain regression model (16.22) corresponds to approximating the data using a weighted sum of data-centered kernel function $\kappa(x_i, \cdot)$. When using a Gaussian kernel (16.23), the bandwidth σ controls the smoothness of the approximation. This is illustrated in Figure 16.17.

In this special case of a square loss, one can solve in closed form (16.21) by solving a $n \times n$ linear system

$$q^* = (KK + \lambda K)^{-1}Ky = (K + \lambda \text{Id}_N)^{-1}y$$

This expression matches exactly (16.13) when using K in place of \hat{C}

Kernelized logistic classification. Logistic classification tries to separate the classes using a linear separating hyperplane $\{x ; \langle \beta, x \rangle = 0\}$. In order to generate a non-linear decision boundary, one can replace the parametric linear model by a non-linear non-parametric model, thanks to kernelization. This allows in particular to generate decision boundaries of arbitrary complexity.

In the two class problem, as detailed in Section 16.4.2, one solves (16.21) using the logistic loss (16.18). This can be for instance achieved by a gradient descent method. Once the solution q^* is obtained, the probability of x belonging to the first class is then

$$\theta\left(\sum_{i=1}^n q_i^* \kappa(x_i, x)\right).$$

Figure 16.18 illustrate such a non-linear decision function on a simple 2-D problem.

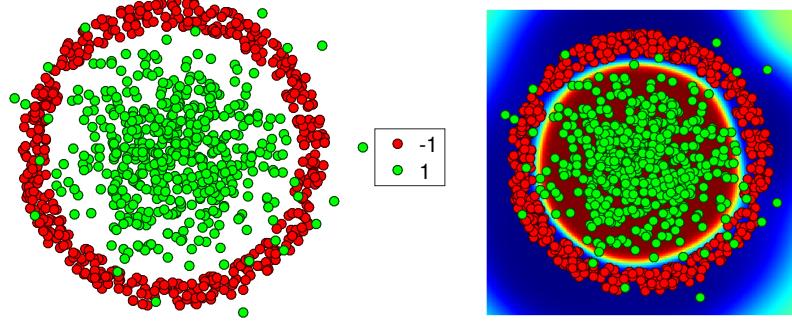


Figure 16.18: Non-linear classification using a Gaussian kernel.

Kernelized nearest-neighbors. It is also possible to extend nearest neighbor classification (as detailed in Section 16.4.1) and regression over a lifted space by making use only of kernel evaluation, simply noticing that

$$\|\varphi(x_i) - \varphi(x_j)\|_{\mathcal{H}}^2 = \kappa(x_i, x_i) + \kappa(x_j, x_j) - 2\kappa(x_i, x_j).$$

Kernel on strings. [ToDo: write me]

Chapter 17

Deep Learning

Before detailing deep architectures and their use, we start this chapter by presenting two essential computational tools that are used to train these models: stochastic optimization methods and automatic differentiation. In practice, they work hand-in-hand to be able to learn painlessly complicated non-linear models on large-scale datasets.

17.1 Stochastic Optimization

We detail some important stochastic Gradient Descent methods, which enable to perform optimization in the setting where the number of samples n is large and even infinite.

17.1.1 Minimizing Sums and Expectation

A large class of functionals in machine learning can be expressed as minimizing large sums of the form

$$\min_{\beta \in \mathbb{R}^p} \mathcal{E}(\beta) \stackrel{\text{def.}}{=} \frac{1}{n} \sum_{i=1}^n \mathcal{E}_i(\beta) \quad (17.1)$$

or even expectations of the form

$$\min_{\beta \in \mathbb{R}^p} \mathcal{E}(\beta) \stackrel{\text{def.}}{=} \mathbb{E}_{\mathbf{z} \sim \pi}(\mathcal{E}(\beta, \mathbf{z})) = \int_{\mathcal{Z}} \mathcal{E}(\beta, z) d\pi(z). \quad (17.2)$$

Problem (17.1) can be seen as a special case of (17.2), when using a discrete empirical uniform measure $\pi = \sum_{i=1}^n \delta_i$ and setting $\mathcal{E}(x, i) = \mathcal{E}_i(x)$. One can also view (17.1) as a discretized “empirical” version of (17.2) when drawing $(z_i)_i$ i.i.d. according to \mathbf{z} and defining $\mathcal{E}_i(x) = \mathcal{E}(x, z_i)$. In this setup, (17.1) converges to (17.2) as $n \rightarrow +\infty$.

A typical example of such a class of problems is empirical risk minimization (here without regularization $J = 0$ for simplicity) (16.7) and its expectation version (16.8), where in these cases

$$\mathcal{E}_i(\beta) = L(f(x_i, \beta), y_i) \quad \text{and} \quad \mathcal{E}(\beta, z) = L(f(x, \beta), y) \quad (17.3)$$

for $z = (x, y) \in \mathcal{Z} = (\mathcal{X} = \mathbb{R}^p) \times (\mathcal{Y} = \mathbb{R}^q)$ (typically $q = 1$). We illustrate below the methods on binary logistic classification, where

$$L(s, y) \stackrel{\text{def.}}{=} \log(1 + \exp(-sy)) \quad \text{and} \quad f(x, \beta) = \langle x, \beta \rangle, \quad (17.4)$$

see Section 16.4.2 for details. But this extends to arbitrary parametric models, and in particular deep neural networks as detailed in Section 17.3.

While some algorithms (in particular batch gradient descent) are specific to finite sums (17.1), the stochastic methods we detail next work verbatim (with the same convergence guarantees) in the expectation case (17.2). For the sake of simplicity, we however do the exposition for the finite sums case, which is sufficient in the vast majority of cases. But one should keep in mind that n can be arbitrarily large, so it is not acceptable in this setting to use algorithms whose complexity per iteration depend on n .

The general idea underlying stochastic optimization methods is *not* to have faster algorithms with respect to traditional optimization schemes such as those detailed in Chapter 13. In almost all cases, if n is not too large so that one afford the price of doing a few non-stochastic iterations, then deterministic methods are faster. But if n is so large that one cannot do even a single deterministic iteration, then stochastic methods allow one to have a fine grained scheme by breaking the cost of deterministic iterations in smaller chunks. Another advantage is that they are quite easy to parallelize.

17.1.2 Batch Gradient Descent (BGD)

The usual deterministic (batch) gradient descent (BGD) is studied in details in Section 13.1. Its iterations read

$$\beta^{(\ell+1)} = \beta^{(\ell)} - \tau_\ell \nabla \mathcal{E}(\beta^{(\ell)})$$

and the step size should be chosen as $0 < \tau_{\min} < \tau_\ell < \tau_{\max} \stackrel{\text{def.}}{=} 2/L$ where L is the Lipschitz constant of the gradient $\nabla \mathcal{E}$. In particular, in this deterministic setting, this step size should not go to zero and this ensures quite fast convergence (even linear rates if \mathcal{E} is strongly convex).

The computation of the gradient in our setting reads

$$\nabla \mathcal{E}(\beta) = \frac{1}{n} \sum_{i=1}^n \nabla \mathcal{E}_i(\beta) \quad (17.5)$$

so it typically has complexity $O(np)$ if computing $\nabla \mathcal{E}_i$ has linear complexity in p .

In the ERM setting (17.3), the gradient reads

$$\nabla \mathcal{E}_i(\beta) = [\partial f(x_i, \beta)]^\top (\nabla L(f(x_i, \beta), y_i)), \quad (17.6)$$

where $\partial f(x, \beta) \in \mathbb{R}^{q \times p}$ is the Jacobian of the mapping $\beta \in \mathbb{R}^p \mapsto f(x, \beta) \in \mathbb{R}^q$, while $\nabla L(y, y') \in \mathbb{R}^q$ is the gradient with respect to the first variable, i.e. the gradient of the map $y \in \mathbb{R}^q \mapsto L(y, y') \in \mathbb{R}$.

In the case of a linear model such as (17.4), this gradient computation simply reads

$$\nabla \mathcal{E}_i(\beta) = L'(\langle x_i, \beta \rangle, y_i) x_i$$

where L' is the differential of L with respect to the first variable. For the logistic loss, it is simply

$$L'(s, y) = -s \frac{e^{-sy}}{1 + e^{-sy}}.$$

17.1.3 Stochastic Gradient Descent (SGD)

For very large n , computing the full gradient $\nabla \mathcal{E}$ as in (17.5) is prohibitive. The idea of SGD is to trade this exact full gradient by an inexact proxy using a single functional \mathcal{E}_i where i is drawn uniformly at random. The main idea that makes this work is that this sampling scheme provides an unbiased estimate of the gradient, in the sense that

$$\mathbb{E}_{\mathbf{i}} \nabla \mathcal{E}_{\mathbf{i}}(\beta) = \nabla \mathcal{E}(\beta) \quad (17.7)$$

where \mathbf{i} is a random variable distributed uniformly in $\{1, \dots, n\}$.

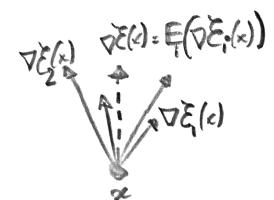


Figure 17.2: Unbiased gradient estimate

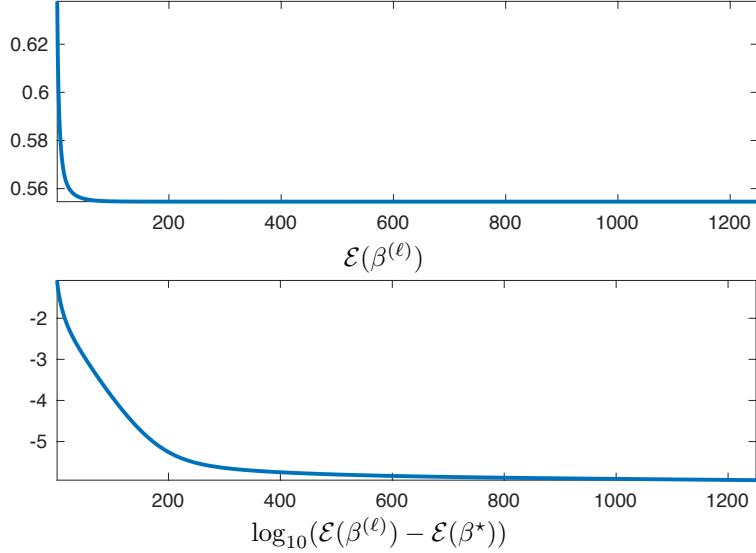


Figure 17.1: Evolution of the error of the BGD for logistic classification.

Starting from some $\beta^{(0)}$, the iterations of stochastic gradient descent (SGD) read

$$\beta^{(\ell+1)} = \beta^{(\ell)} - \tau_\ell \nabla \mathcal{E}_{i(\ell)}(\beta^{(\ell)})$$

where, for each iteration index ℓ , $i(\ell)$ is drawn uniformly at random in $\{1, \dots, n\}$. It is important that the iterates $\beta^{(\ell+1)}$ are thus random vectors, and the theoretical analysis of the method thus studies whether this sequence of random vectors converges (in expectation or in probability for instance) toward a deterministic vector (minimizing \mathcal{E}), and at which speed.

Note that each step of a batch gradient descent has complexity $O(np)$, while a step of SGD only has complexity $O(p)$. SGD is thus advantageous when n is very large, and one cannot afford to do several passes through the data. In some situations, SGD can provide accurate results even with $\ell \ll n$, exploiting redundancy between the samples.

A crucial question is the choice of step size schedule τ_ℓ . It must tend to 0 in order to cancel the noise induced on the gradient by the stochastic sampling. But it should not go too fast to zero in order for the method to keep converging.

A typical schedule that ensures both properties is to have asymptotically $\tau_\ell \sim \ell^{-1}$ for $\ell \rightarrow +\infty$. We thus propose to use

$$\tau_\ell \stackrel{\text{def.}}{=} \frac{\tau_0}{1 + \ell/\ell_0} \quad (17.8)$$

where ℓ_0 indicates roughly the number of iterations serving as a “warmup” phase.

Figure 17.4 shows a simple 1-D example to minimize $\mathcal{E}_1(\beta) + \mathcal{E}_2(\beta)$ for $\beta \in \mathbb{R}$ and $\mathcal{E}_1(\beta) = (\beta - 1)^2$ and $\mathcal{E}_2(\beta) = (\beta + 1)^2$. One can see how the density of the distribution of $\beta^{(\ell)}$ progressively clusters around the minimizer $\beta^* = 0$. Here the distribution of $\beta^{(0)}$ is uniform on $[-1/2, 1/2]$.

The following theorem shows the convergence in expectation with a $1/\sqrt{\ell}$ rate on the objective.

Theorem 52. *We assume \mathcal{E} is μ -strongly convex as defined in (S _{μ}) (i.e. $\text{Id}_{N \times N} \preceq \partial^2 \mathcal{E}(\beta)$ if \mathcal{E} is C^2), and is such that $\|\nabla \mathcal{E}_i(x)\|^2 \leq C^2$. For the step size choice $\tau_\ell = \frac{1}{\mu(\ell+1)}$, one has*

$$\mathbb{E}(\|\beta^{(\ell)} - \beta^*\|^2) \leq \frac{R}{\ell + 1} \quad \text{where} \quad R = \max(\|\beta^{(0)} - \beta^*\|, C^2/\mu^2), \quad (17.9)$$

where \mathbb{E} indicates an expectation with respect to the i.i.d. sampling performed at each iteration.



Figure 17.3:
Schematic view
of SGD iterates

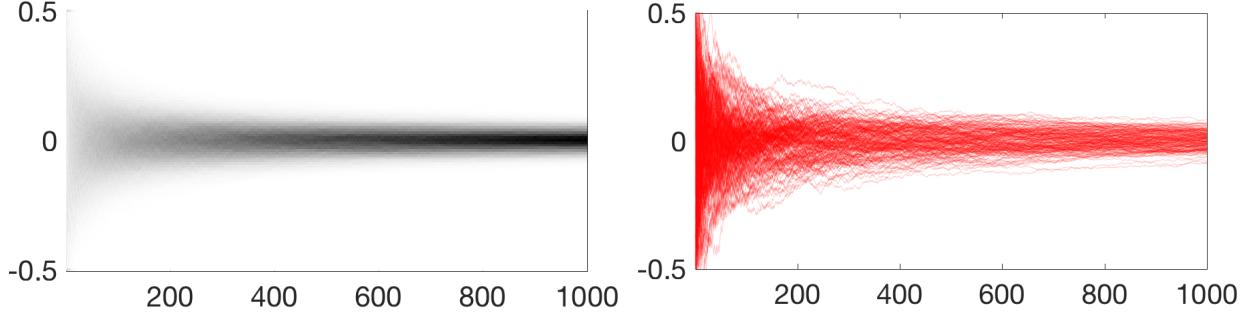


Figure 17.4: Display of a large number of trajectories $\ell \mapsto \beta^{(\ell)} \in \mathbb{R}$ generated by several runs of SGD. On the top row, each curve is a trajectory, and the bottom row displays the corresponding density.

Proof. By strong convexity, one has

$$\begin{aligned}\mathcal{E}(\beta^*) - \mathcal{E}(\beta^{(\ell)}) &\geq \langle \nabla \mathcal{E}(\beta^{(\ell)}), \beta^* - \beta^{(\ell)} \rangle + \frac{\mu}{2} \|\beta^{(\ell)} - \beta^*\|^2 \\ \mathcal{E}(\beta^{(\ell)}) - \mathcal{E}(\beta^*) &\geq \langle \nabla \mathcal{E}(\beta^*), \beta^{(\ell)} - \beta^* \rangle + \frac{\mu}{2} \|\beta^{(\ell)} - \beta^*\|^2.\end{aligned}$$

Summing these two inequalities and using $\nabla \mathcal{E}(\beta^*) = 0$ leads to

$$\langle \nabla \mathcal{E}(\beta^{(\ell)}) - \nabla \mathcal{E}(\beta^*), \beta^{(\ell)} - \beta^* \rangle = \langle \nabla \mathcal{E}(\beta^{(\ell)}), \beta^{(\ell)} - \beta^* \rangle \geq \mu \|\beta^{(\ell)} - \beta^*\|^2. \quad (17.10)$$

Considering only the expectation with respect to the ransom sample of $i(\ell) \sim \mathbf{i}_\ell$, one has

$$\begin{aligned}\mathbb{E}_{\mathbf{i}_\ell}(\|\beta^{(\ell+1)} - \beta^*\|^2) &= \mathbb{E}_{\mathbf{i}_\ell}(\|\beta^{(\ell)} - \tau_\ell \nabla \mathcal{E}_{\mathbf{i}_\ell}(\beta^{(\ell)}) - \beta^*\|^2) \\ &= \|\beta^{(\ell)} - \beta^*\|^2 + 2\tau_\ell \langle \mathbb{E}_{\mathbf{i}_\ell}(\nabla \mathcal{E}_{\mathbf{i}_\ell}(\beta^{(\ell)})), \beta^* - \beta^{(\ell)} \rangle + \tau_\ell^2 \mathbb{E}_{\mathbf{i}_\ell}(\|\nabla \mathcal{E}_{\mathbf{i}_\ell}(\beta^{(\ell)})\|^2) \\ &\leq \|\beta^{(\ell)} - \beta^*\|^2 + 2\tau_\ell \langle \nabla \mathcal{E}(\beta^{(\ell)}), \beta^* - \beta^{(\ell)} \rangle + \tau_\ell^2 C^2\end{aligned}$$

where we used the fact (17.7) that the gradient is unbiased. Taking now the full expectation with respect to all the other previous iterates, and using (17.10) one obtains

$$\mathbb{E}(\|\beta^{(\ell+1)} - \beta^*\|^2) \leq \mathbb{E}(\|\beta^{(\ell)} - \beta^*\|^2) - 2\mu\tau_\ell \mathbb{E}(\|\beta^{(\ell)} - \beta^*\|^2) + \tau_\ell^2 C^2 = (1 - 2\mu\tau_\ell) \mathbb{E}(\|\beta^{(\ell)} - \beta^*\|^2) + \tau_\ell^2 C^2. \quad (17.11)$$

We show by recursion that the bound (17.9) holds. We denote $\varepsilon_\ell \stackrel{\text{def.}}{=} \mathbb{E}(\|\beta^{(\ell)} - \beta^*\|^2)$. Indeed, for $\ell = 0$, this is true that

$$\varepsilon_0 \leq \frac{\max(\|\beta^{(0)} - \beta^*\|, C^2/\mu^2)}{1} = \frac{R}{1}.$$

We now assume that $\varepsilon_\ell \leq \frac{R}{\ell+1}$. Using (17.11) in the case of $\tau_\ell = \frac{1}{\mu(\ell+1)}$, one has, denoting $m = \ell + 1$

$$\begin{aligned}\varepsilon_{\ell+1} &\leq (1 - 2\mu\tau_\ell)\varepsilon_\ell + \tau_\ell^2 C^2 = \left(1 - \frac{2}{m}\right)\varepsilon_\ell + \frac{C^2}{(\mu m)^2} \\ &\leq \left(1 - \frac{2}{m}\right)\frac{R}{m} + \frac{R}{m^2} = \left(\frac{1}{m} - \frac{1}{m^2}\right)R = \frac{m-1}{m^2}R = \frac{m^2-1}{m^2(m+1)}R \leq \frac{R}{m+1}\end{aligned}$$

□

A weakness of SGD (as well as the SGA scheme studied next) is that it only weakly benefit from strong convexity of \mathcal{E} . This is in sharp contrast with BGD, which enjoy a fast linear rate for strongly convex functionals, see Theorem 35.

Figure 17.5 displays the evolution of the energy $\mathcal{E}(\beta^{(\ell)})$. It overlays on top (black dashed curve) the convergence of the batch gradient descent, with a careful scaling of the number of iteration to account for the fact that the complexity of a batch iteration is n times larger.

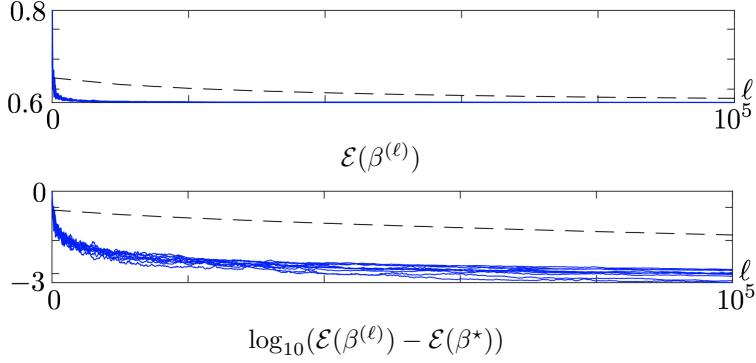


Figure 17.5: Evolution of the error of the SGD for logistic classification (dashed line shows BGD).

17.1.4 Stochastic Gradient Descent with Averaging (SGA)

Stochastic gradient descent is slow because of the fast decay of τ_ℓ toward zero. To improve somehow the convergence speed, it is possible to average the past iterate, i.e. run a “classical” SGD on auxiliary variables $(\tilde{\beta}^{(\ell)})_\ell$

$$\tilde{\beta}^{(\ell+1)} = \tilde{\beta}^{(\ell)} - \tau_\ell \nabla \mathcal{E}_{i(\ell)}(\tilde{\beta}^{(\ell)})$$

and output as estimated weight vector the Cesaro average

$$\beta^{(\ell)} \stackrel{\text{def.}}{=} \frac{1}{\ell} \sum_{k=1}^{\ell} \tilde{\beta}^{(k)}.$$

This defines the Stochastic Gradient Descent with Averaging (SGA) algorithm.

Note that it is possible to avoid explicitly storing all the iterates by simply updating a running average as follow

$$\beta^{(\ell+1)} = \frac{1}{\ell} \tilde{\beta}^{(\ell)} + \frac{\ell-1}{\ell} \beta^{(\ell)}.$$

In this case, a typical choice of decay is rather of the form

$$\tau_\ell \stackrel{\text{def.}}{=} \frac{\tau_0}{1 + \sqrt{\ell/\ell_0}}.$$

Notice that the step size now goes much slower to 0, at rate $\ell^{-1/2}$.

Typically, because the averaging stabilizes the iterates, the choice of (ℓ_0, τ_0) is less important than for SGD.

Bach proves that for logistic classification, it leads to a faster convergence (the constant involved are smaller) than SGD, since on contrast to SGD, SGA is adaptive to the local strong convexity of E .

17.1.5 Stochastic Averaged Gradient Descent (SAG)

For problem size n where the dataset (of size $n \times p$) can fully fit into memory, it is possible to further improve the SGA method by bookkeeping the previous gradients. This gives rise to the Stochastic Averaged Gradient Descent (SAG) algorithm.

We store all the previously computed gradients in $(G^i)_{i=1}^n$, which necessitates $O(n \times p)$ memory. The iterates are defined by using a proxy g for the batch gradient, which is progressively enhanced during the iterates.

The algorithm reads

$$h \leftarrow \nabla \mathcal{E}_{i(\ell)}(\tilde{\beta}^{(\ell)}),$$

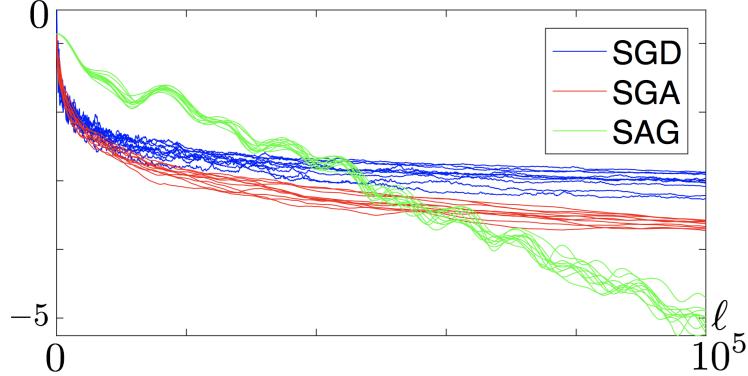


Figure 17.6: Evolution of $\log_{10}(\mathcal{E}(\beta^{(\ell)}) - \mathcal{E}(\beta^*))$ for SGD, SGA and SAG.

$$\begin{aligned} g &\leftarrow g - G^{i(\ell)} + h, \\ G^{i(\ell)} &\leftarrow h, \\ \beta^{(\ell+1)} &= \beta^{(\ell)} - \tau g. \end{aligned}$$

Note that in contrast to SGD and SGA, this method uses a fixed step size τ . Similarly to the BGD, in order to ensure convergence, the step size τ should be of the order of $1/L$ where L is the Lipschitz constant of \mathcal{E} .

This algorithm improves over SGA and SGD since it has a convergence rate of $O(1/\ell)$ as does BGD. Furthermore, in the presence of strong convexity (for instance when X is injective for logistic classification), it has a linear convergence rate, i.e.

$$\mathbb{E}(\mathcal{E}(\beta^{(\ell)})) - \mathcal{E}(\beta^*) = O(\rho^\ell),$$

for some $0 < \rho < 1$.

Note that this improvement over SGD and SGA is made possible only because SAG explicitly uses the fact that n is finite (while SGD and SGA can be extended to infinite n and more general minimization of expectations (17.2)).

Figure 17.9 shows a comparison of SGD, SGA and SAG.

17.2 Automatic Differentiation

The main computational bottleneck of these gradient descent methods (batch or stochastic) is the evaluation of the elementary gradients $\nabla \mathcal{E}_i$. The gradient formula (17.6) shows that it requires to remap the gradient of the loss $\nabla L(f(x_i, \beta), y_i)$ through the adjoint of the Jacobian $\partial f(x_i, \beta)$. The general idea is that for complicated model this computation should be broken in simpler sub-computation, which ultimately should corresponds to elementary operators (binary operators such as $+$ or $*$ and unary operators such as \exp , \log , etc.) for which the differential are trivial to compute.

17.2.1 Reverse Differentiation on a Feedforward Graph

To give a concrete examples which is actually found in many practical situation (and in particular for simple deep architectures, as detailed in Section 17.3.1), if the functional to be differentiated has the form

$$\mathcal{E}(\beta) = \mathcal{L} \circ \mathcal{F}_{L-1} \circ \mathcal{F}_{L-2} \circ \dots \circ \mathcal{F}_0(\beta) \quad (17.12)$$

(often called a “feedforward” model) where $\mathcal{F}_\ell : \mathbb{R}^{n_\ell} \rightarrow \mathbb{R}^{n_{\ell+1}}$ and $\mathcal{L} : \mathbb{R}^{n_L} \rightarrow \mathbb{R}$, then one can compute the gradient $\nabla \mathcal{E}(\beta)$ in two steps:

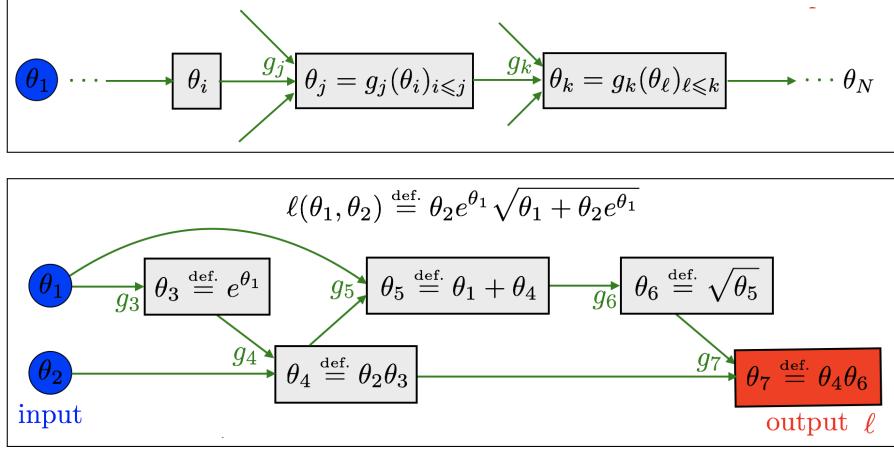


Figure 17.7: Top: elementary component of the DAG computational graph. Bottom: example of DAG computational graph.

- A forward pass, where one evaluates the function itself and keep track of all the intermediate computations, i.e., initializing $\beta_0 = \beta$, one computes

$$\beta_{\ell+1} \stackrel{\text{def.}}{=} \mathcal{F}_\ell(\beta_\ell) \quad \text{and} \quad \mathcal{E}(\beta) = \mathcal{L}(\beta_L).$$

- A backward pass, where one use the chain rule together with the Jacobian transposition

$$\nabla \mathcal{E}(\beta) = [\partial \mathcal{F}_0(\beta_0)]^* \circ \dots \circ [\partial \mathcal{F}_{L-1}(\beta_{L-1})]^* (\nabla \mathcal{L}(\beta_L)) \quad (17.13)$$

to define the following backward recursion, initialized by $h_L = \nabla \mathcal{L}(\beta_L)$, and then

$$h_{\ell-1} \stackrel{\text{def.}}{=} [\partial \mathcal{F}_{\ell-1}(\beta_{\ell-1})]^*(h_\ell) \quad \text{and} \quad \nabla \mathcal{E}(\beta) = h_0. \quad (17.14)$$

The main issue here is when these adjoint Jacobian $[\partial \mathcal{F}_{\ell-1}(\beta_{\ell-1})]^*$ are difficult to apply (and it is out of question in most cases to actually store them on a computer, since it would occurs an enormous storage requirement and typical quadratic time scaling for the algorithms). We now detail a finer grained analysis which enable to tackle building blocks of arbitrary complexity.

The computation (17.13) should be compared with the “forward” accumulation

$$\nabla \mathcal{E}(\beta) = \partial \mathcal{L}(\beta_L) \circ \partial \mathcal{F}_{L-1}(\beta_{L-1}) \dots \circ \partial \mathcal{F}_0(\beta_0).$$

Computing these matrix product would be extremely costly, although it would require no memory overhead because the computation would be carried over in parallel to the evaluation of the function.

17.2.2 Reverse Differentiation on a Generic Computational Graph

One can generalize the idea above to differentiate automatically any function which can be implemented on a computer. What is even more surprising is that the computational cost is the same as the one of evaluating the function itself. This fundamental computational fact (that gradient evaluation and function evaluation have the same computational cost) is not so well known, but is of paramount practical interest when it comes to differentiating complicated recursive functions. We will apply it in a very simple setup for deep-architectures, but it can be applied to much more involved computational architectures.

Note that this results only applies for function which output scalar values. For functions which output vector values, one can of course re-use this idea for each output, but this is in general vastly sub-optimal, because it ignore the redundancy between the computation of each output. The determination of optimal

strategy in this case is known to be NP-hard. This include for instance the computation of the Hessian of a scalar valued function (since it corresponds to the differentiation of the gradient, which is itself a vector-valued function). Fortunately, for machine learning application, one is often interested in differentiating only empirical losses functions, which are scalar valued.

Forward pass as a DAG traversal. The crux of this idea is that the computational flow of any computable function ℓ can be represented as a directed acyclic graph. We denote $(\theta_i)_{i=1}^R$ the set of all scalar variables (input, output and intermediary) manipulated by the computational program. Without loss of generality, we impose that the first variable $(\theta_1, \dots, \theta_M)$ are the M input variables, while the last θ_R is the output variable. The function to be computed is thus of the form

$$\theta_R = \ell(\theta_1, \dots, \theta_M)$$

where $\ell : \mathbb{R}^M \rightarrow \mathbb{R}$ is broken in $R - M + 1$ intermediate steps corresponding to all the remaining variable $(\theta_i)_{M < i < R}$. The successive execution of the program defines an ordering of all the intermediate variables, so that, after initializing the input variables $(\theta_1, \dots, \theta_M)$, the forward pass computes the value of θ_r for $r = M + 1, \dots, R$ as

$$\theta_r = g_r(\theta_{\pi(r)})$$

for some scalar valued function $g_r : \mathbb{R}^{|\pi(r)|} \rightarrow \mathbb{R}$, where $\pi(r) \subset \{1, \dots, r - 1\}$ is the set of “parent” node of r in a directed acyclic graph (DAG). Figure 17.7 shows an example of such a computational DAG.

From a symbolic computation point of view, variables θ_j (for $j > M$) in the graph can be interpreted either as variables (i.e. which can be assigned scalar values) and functions depending on input variables θ_m for $m \leq M$. The beauty of this DAG representation is that one can also view θ_j as depending on any other intermediate variable θ_i as long as $i < j$.

Direct mode auto-diff. The goal is to compute the gradient vector, which reads

$$\nabla \ell = \left(\frac{\partial \theta_R}{\partial \theta_m} \right)_{m=1}^M.$$

The naive way to compute this gradient vector would thus be to compute for each of the M input variable θ_m the differential $\frac{\partial \theta_j}{\partial \theta_m}$ of all function θ_j with respect to θ_m . Without loss of generality, we consider $m = 1$. This can be achieved by using the standard chain rule

$$\frac{\partial \theta_j}{\partial \theta_1} = \sum_{i \in \pi(j)} \frac{\partial \theta_j}{\partial \theta_i} \frac{\partial \theta_i}{\partial \theta_1}. \quad (17.15)$$

Here the multipliers involved are actually differential of the elementary functions

$$\frac{\partial \theta_j}{\partial \theta_i} = \partial_i g_j$$

Note that this writing is abusive, since $\frac{\partial \theta_j}{\partial \theta_i}$ really means that in practice such a differential is evaluated assuming all the variable $(\theta_r)_{r < j}$ on which the function θ_j depends are defined to their respective value (which have been computed by the forward pass, which here can be run in parallel to the forward DAG traversal).

By traversing forward the DAG, iterating this formula compute all the derivative, and in particular $(\nabla \ell)_1 = \frac{\partial \theta_R}{\partial \theta_1}$. This approach, while being the most natural, is however vastly sub-optimal because its complexity is M times the one of the evaluation of the function.

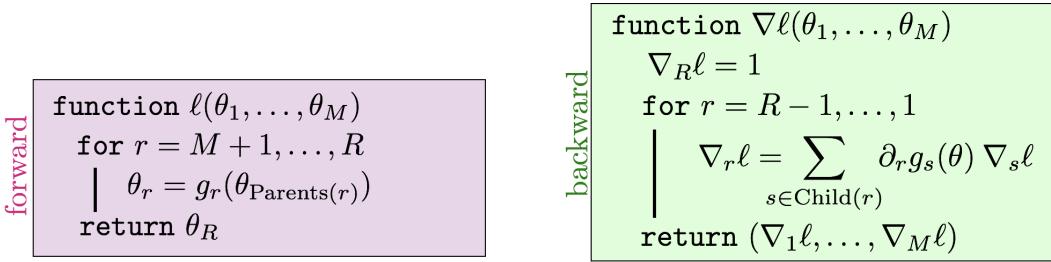


Figure 17.8: Recap of the two step of the automatic differentiation procedure.

Reverse mode auto-diff. Instead of computing the quantities $(\frac{\partial \theta_j}{\partial \theta_1})_j$, a radically different approach consists in rather computing the quantities $\frac{\partial \theta_L}{\partial \theta_j} = (\nabla \ell)_j$. In place of the “forward” chain rule, one needs to the backward one

$$\frac{\partial \theta_R}{\partial \theta_j} = \sum_{k \in \pi(k)} \frac{\partial \theta_R}{\partial \theta_k} \frac{\partial \theta_k}{\partial \theta_j}. \quad (17.16)$$

Note that here the summation is done over k which are “child” of j in the DAG. Here the multiplier appearing in the formula are differential of the elementary function since $\frac{\partial \theta_k}{\partial \theta_j} = \partial_j g_k$. The main interest of this reverse recursion (17.16) with respect to the direct one (17.16) is that it only needs to be run once, so that the overall complexity is the same as the one of the forward pass to compute the function itself.

Figure 17.8 recaps the two passes of the reverse mode automatic differentiation method.

The main bottleneck of this backward automatic differentiation technic is the memory consumption. Indeed, since all intermediate results need to be computed and stored explicitly before applying the backward pass, memory grows proportionally to execution time. This can be unacceptable for very large machine learning model. Fortunately, it is possible to trade time vs. memory and only keep track of a fraction of intermediate results, and retrieve the missing result locally by small forward passes. Doing this approach recursively allows to only have a logarithmic overhead in term of both time and memory, showing the vast superiority of automatic differentiation method with respect to any other alternative for differentiation. We could not insist more on the crucial importance and impact of this class of technics on modern data science.

17.3 Deep Discriminative Models

17.3.1 Deep Network Structure

Deep learning are estimator $f(x, \beta)$ which are built as composition of simple building blocks. In their simplest form (non-recursive), they corresponds to a simple linear computational graph as already defined in (17.12) (without the loss \mathcal{L}), and we write this as

$$f(\cdot, \beta) = f_{L-1}(\cdot, \beta_1) \circ f_{L-2}(\cdot, \beta_2) \circ \dots \circ f_0(\cdot, \beta_0)$$

where $\beta = (\beta_0, \dots, \beta_{L-1})$ is the set of parameters, and

$$f_\ell(\cdot, \beta_\ell) : \mathbb{R}^{n_\ell} \rightarrow \mathbb{R}^{n_{\ell+1}}$$

While it is possible to consider more complicated architecture (in particular recurrent ones), we restrict here our attention to these simple linear graph computation structures (so-called feedforward networks).

The supervised learning of these parameters β is usually done by empirical risk minimization (16.7) using SGD-type methods as explained in Section 17.1. Note that this results in highly non-convex optimization problems. In particular, strong convergence guarantees such as Theorem 52 do not hold anymore, and only weak convergence (toward stationary points) holds. SGD type technics are however found to work

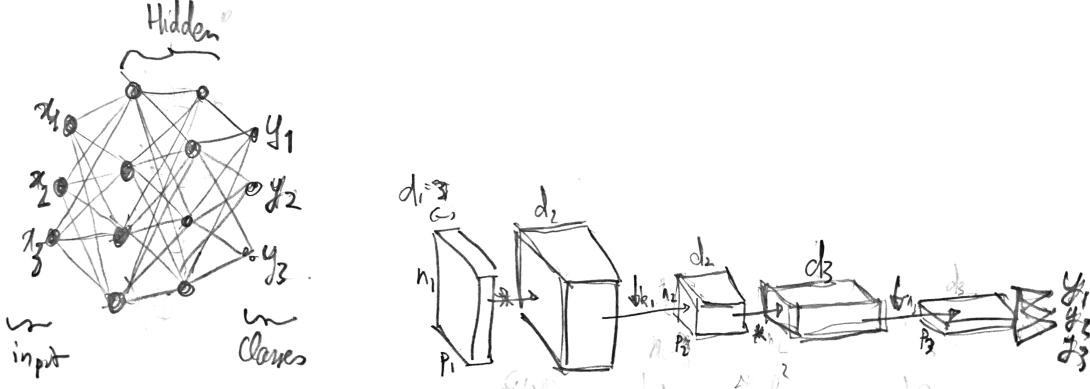


Figure 17.9: Left: example of fully connected network. Right: example of convolutional neural network.

surprisingly well in practice, and it now believe that the success of these deep-architecture approaches (in particular the ability of these over-parameterized model to generalize well) are in large part due to the dynamics of the SGD itself, which induce an implicit regularization effect.

For these simple linear architectures, the gradient of the ERM loss (17.6) can be computed using the reverse mode computation detailed in Section 17.2.1. In particular, in the context of deep learning, formula (17.18). One should however keep in mind that for more complicated (e.g. recursive) architectures, such a simple formula is not anymore available, and one should resort to reverse mode automatic differentiation (see Section 17.2.2), which, while being conceptually simple, is actually implementing possibly highly non-trivial and computationally optimal recursive differentiation.

In most successful applications of deep-learning, each computational block $f_\ell(\cdot, \beta_\ell)$ is actually very simple, and is the composition of

- an affine map, $B_\ell \cdot + b_\ell$ with a matrix $B_\ell \in \mathbb{R}^{n_\ell \times \tilde{n}_\ell}$ and a vector $b_\ell \in \mathbb{R}^{\tilde{n}_\ell}$ parametrized (in most case linearly) by β_ℓ ,
- a fixed (not depending on β_ℓ) non-linearity $\rho_\ell : \mathbb{R}^{\tilde{n}_\ell} \rightarrow \mathbb{R}^{n_{\ell+1}}$

which we write as

$$\forall x_\ell \in \mathbb{R}^{n_\ell}, \quad f_\ell(x_\ell, \beta_\ell) = \rho_\ell(B_\ell x_\ell + b_\ell) \in \mathbb{R}^{n_{\ell+1}}. \quad (17.17)$$

In the simplest case, the so-called “fully connected”, one has $(B_\ell, b_\ell) = \beta_\ell$, i.e. B_ℓ is a full matrix and its entries (together with the bias b_ℓ) are equal to the set of parameters β_ℓ . Also in the simplest cases ρ_ℓ is a pointwise non-linearity $\rho_\ell(z) = (\tilde{\rho}_\ell(z_k))_k$, where $\tilde{\rho}_\ell : \mathbb{R} \rightarrow \mathbb{R}$ is non-linear. The most usual choices are the rectified linear unit (ReLU) $\tilde{\rho}_\ell(s) = \max(s, 0)$ and the sigmoid $\tilde{\rho}_\ell(s) = \theta(s) = (1 + e^{-s})^{-1}$.

The important point here is that the interleaving of non-linear map progressively increases the complexity of the function $f(\cdot, \beta)$.

The parameter $\beta = (B_\ell, b_\ell)_\ell$ of such a deep network are then trained by minimizing the ERM functional (16.7) using SGD-type stochastic optimization method. The gradient can be computed efficiently (with complexity proportional to the application of the model, i.e. $O(\sum_\ell n_\ell^2)$) by automatic differentiation. Since such models are purely feedforward, one can directly use the back-propagation formula (17.12).

For regression tasks, one can directly use the output of the last layer (using e.g. a ReLu non-linearity) in conjunction with a ℓ^2 squared loss L . For classification tasks, the output of the last layer needs to be transformed into class probabilities by a multi-class logistic map (??).

An issue with such a fully connected setting is that the number of parameters is too large to be applicable to large scale data such as images. Furthermore, it ignores any prior knowledge about the data, such as for instance some invariance. This is addressed in more structured architectures, such as for instance convolutional networks detailed in Section 17.3.3.

17.3.2 Perceptron and Shallow Models

Before going on with the description of deep architectures, let us re-interpret the logistic classification method detailed in Sections 16.4.2 and 16.4.3.

The two-class logistic classification model (16.16) is equal to a single layer ($L = 1$) network of the form (17.17) (ignoring the constant bias term) where

$$B_0x = \langle x, \beta \rangle \quad \text{and} \quad \tilde{\lambda}_0(u) = \theta(u).$$

The resulting one-layer network $f(x, \beta) = \theta(\langle x, \beta \rangle)$ (possibly including a bias term by adding one dummy dimension to x) is trained using the loss, for binary classes $y \in \{0, 1\}$

$$L(t, y) = -\log(t^y(1-t)^{1-y}) = -y \log(t) - (1-y) \log(1-t).$$

In this case, the ERM optimization is of course a convex program.

Multi-class models with K classes are obtained by computing $B_0x = (\langle x, \beta_k \rangle)_{k=1}^K$, and a normalized logistic map

$$f(x, \beta) = \mathcal{N}((\exp(\langle x, \beta_k \rangle))_k) \quad \text{where} \quad \mathcal{N}(u) = \frac{u}{\sum_k u_k}$$

and assuming the classes are represented using vectors y on the probability simplex, one should use as loss

$$L(t, y) = -\sum_{k=1}^K y_k \log(t_k).$$

17.3.3 Convolutional Neural Networks

In order to be able to tackle data of large size, and also to improve the performances, it is important to leverage some prior knowledge about the structure of the typical data to process. For instance, for signal, images or videos, it is important to make use of the spacial location of the pixels and the translation invariance (up to boundary handling issues) of the domain.

Convolutional neural networks are obtained by considering that the manipulated vectors $x_\ell \in \mathbb{R}^{n_\ell}$ at depth ℓ in the network are of the form $x_\ell \in \mathbb{R}^{\bar{n}_\ell \times d_\ell}$, where \bar{n}_ℓ is the number of “spatial” positions (typically along a 1-D, 2-D, or 3-D grid) and d_ℓ is the number of “channels”. For instance, for color images, one starts with \bar{n}_ℓ being the number of pixels, and $d_\ell = 3$.

The linear operator $B_\ell : \mathbb{R}^{\bar{n}_\ell \times d_\ell} \rightarrow \mathbb{R}^{\bar{n}_{\ell+1} \times d_{\ell+1}}$ is then (up to boundary artefact) translation invariant and hence a convolution along each channel (note that the number of channels can change between layers). It is thus parameterized by a set of filters $(\psi_{\ell,r,s})_{s=1,\dots,d_\ell}^{r=1,\dots,d_{\ell+1}}$. Denoting $x_\ell = (x_{\ell,s,\cdot})_{s=1}^{d_\ell}$ the different layers composing x_ℓ , the linear map reads

$$\forall r \in \{1, \dots, d_{\ell+1}\}, \quad (B_\ell x_\ell)_{r,\cdot} = \sum_{s=1}^{d_\ell} \psi_{\ell,r,s} \star x_{\ell,s,\cdot}$$

and the bias term $b_\ell \in \mathbb{R}$ is constant (to maintain translation invariance).

The non-linear maps across layers serve two purposes: as before a pointwise non-linearity is applied, and then a sub-sampling helps to reduce the computational complexity of the network. This is very similar to the construction of the fast wavelet transform. Denoting by m_k the amount of down-sampling, where usually $m_k = 1$ (no reduction) or $m_k = 2$ (reduction by a factor two in each direction). One has

$$\lambda_\ell(u) = \left(\tilde{\lambda}_\ell(u_{s,m_k,\cdot}) \right)_{s=1,\dots,d_{\ell+1}}.$$

In the literature, it has been proposed to replace linear sub-sampling by non-linear sub-sampling, for instance the so-called max-pooling (that operate by taking the maximum among groups of m_ℓ successive values), but

it seems that linear sub-sampling is sufficient in practice when used in conjunction with very deep (large L) architectures.

The intuition behind such model is that as one moves deeper through the layers, the neurons are receptive to larger areas in the image domain (although, since the transform is non-linear, precisely giving sense to this statement and defining a proper “receptive field” is non-trivial). Using an increasing number of channels helps to define different classes of “detectors” (for the first layer, they detect simple patterns such as edges and corner, and progressively capture more elaborated shapes).

In practice, the last few layers (2 or 3) of such a CNN architectures are chosen to be fully connected. This is possible because, thanks to the sub-sampling, the dimension of these layers are small.

The parameters of such a model are the filters $\beta = (\psi_{\ell,r,s})_{\ell,s,r}$, and they are trained by minimizing the ERM functional (16.7). The gradient is typically computed by backpropagation. Indeed, when computing the gradient with respect to some filter $\psi_{\ell,r,s}$, the feedforward computational graph has the form (17.12). For simplicity, we re-formulate this computation in the case of a single channel per layer (multiple layer can be understood as replacing convolution by matrix-domain convolution). The forward pass computes all the inner coefficients, by traversing the network from $\ell = 0$ to $\ell = L - 1$,

$$x_{\ell+1} = \lambda_\ell(\psi_\ell \star x_\ell)$$

where $\lambda_\ell(u) = (\tilde{\lambda}_\ell(u_i))_i$ is applied component wise. Then, denoting $\mathcal{E}(\beta) = \mathcal{L}(\beta, y)$ the loss to be minimized with respect to the set of filters $\beta = (\psi_\ell)_\ell$, and denoting $\nabla_\ell \mathcal{E}(\beta) = \frac{\partial \mathcal{E}(\beta)}{\partial \psi_\ell}$ the gradient with respect to ψ_ℓ , one computes all the gradients by traversing the network in reverse order, from $\ell = L - 1$ to $\ell = 0$

$$\nabla_\ell \mathcal{E}(\beta) = [\lambda'_\ell(\psi_\ell \star x_\ell)] \odot [\bar{\psi}_\ell \star \nabla_{\ell+1} \mathcal{E}(\beta)], \quad (17.18)$$

where $\lambda'_\ell(u) = (\tilde{\lambda}'_\ell(u_i))_i$ applies the derivative of $\tilde{\lambda}_\ell$ component wise, and where $\bar{\psi}_\ell = \psi_\ell(-\cdot)$ is the reversed filter. Here, \odot is the pointwise multiplication of vectors. The recursion is initialized as $\nabla \mathcal{E}_L(\beta) = \nabla \mathcal{L}(x_L, y)$, the gradient of the loss itself.

This recursion (17.18) is the celebrated backpropagation algorithm put forward by Yann LeCun. Note that to understand and code these iterations, one does not need to rely on the advanced machinery of reverse mode automatic differentiation exposed in Section 17.2.2. The general automatic differentiation method is however crucial to master because advanced deep-learning architectures are not purely feedforward, and might include recursive connexions. Furthermore, automatic differentiation is useful outside deep learning, and considerably eases prototyping for modern data-sciences with complicated non-linear models.

17.3.4 Scattering Transform

The scattering transform, introduced by Mallat and his collaborators, is a specific instance of deep convolutional network, where the filters $(\psi_{\ell,r,s})_{\ell,s,r}$ are not trained, and are fixed to be wavelet filters. This network can be understood as a non-linear extension of the wavelet transform. In practice, the fact that it is fixed prevent it to be applied to arbitrary data (and is used mostly on signals and images) and it does not lead to state of the art results for natural images. Nevertheless, it allows to derive some regularity properties about the feature extraction map $f(\cdot, \beta)$ computed by the network in term of stability to diffeomorphisms. It can also be used as a set of fixed initial features which can be further enhanced by a trained deep network, as shown by Edouard Oyallon.

Chapter 18

Optimal Transport

The main reference for this chapter is the book “Computational Optimal Transport”¹. We will only recap here some important points.

18.1 Radon Measures

Measures. We will interchangeably the term histogram or probability vector for any element $\mathbf{a} \in \Sigma_n$ that belongs to the probability simplex

$$\Sigma_n \stackrel{\text{def.}}{=} \left\{ \mathbf{a} \in \mathbb{R}_+^n ; \sum_{i=1}^n \mathbf{a}_i = 1 \right\}.$$

A discrete measure with weights \mathbf{a} and locations $x_1, \dots, x_n \in \mathcal{X}$ reads

$$\alpha = \sum_{i=1}^n \mathbf{a}_i \delta_{x_i} \tag{18.1}$$

where δ_x is the Dirac at position x , intuitively a unit of mass which is infinitely concentrated at location x . Such a measure describes a probability measure if, additionally, $\mathbf{a} \in \Sigma_n$, and more generally a positive measure if each of the “weights” described in vector \mathbf{a} is positive itself.

Remark 9 (General measures). A convenient feature of OT is that it can deal with discrete and continuous “objects” within the same framework. Such objects only need to be modelled as measures. This corresponds to the notion of Radon measures $\mathcal{M}(\mathcal{X})$ on the space \mathcal{X} . The formal definition of that set requires that \mathcal{X} is equipped with a distance, usually denoted d , because one can only access a measure by “testing” (integrating) it against continuous functions, denoted $f \in \mathcal{C}(\mathcal{X})$.

Integration of $f \in \mathcal{C}(\mathcal{X})$ against a discrete measure α computes a sum

$$\int_{\mathcal{X}} f(x) d\alpha(x) = \sum_{i=1}^n \mathbf{a}_i f(x_i).$$

More general measures, for instance on $\mathcal{X} = \mathbb{R}^d$ (where $d \in \mathbb{N}^*$ is the dimension), can have a density $d\alpha(x) = \rho_\alpha(x) dx$ w.r.t. the Lebesgue measure, often denoted $\rho_\alpha = \frac{d\alpha}{dx}$, which means that

$$\forall h \in \mathcal{C}(\mathbb{R}^d), \quad \int_{\mathbb{R}^d} h(x) d\alpha(x) = \int_{\mathbb{R}^d} h(x) \rho_\alpha(x) dx.$$

¹optimaltransport.github.io

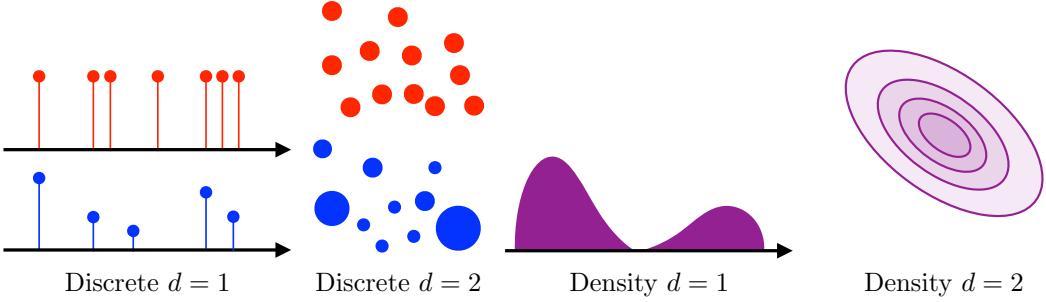


Figure 18.1: Schematic display of discrete distributions $\alpha = \sum_{i=1}^n \mathbf{a}_i \delta_{x_i}$ (red corresponds to empirical uniform distribution $\mathbf{a}_i = 1/n$, and blue to arbitrary distributions) and densities $d\alpha(x) = \rho_\alpha(x)dx$ (in violet), in both 1-D and 2-D. Discrete distributions in 1-D are displayed using vertical segments (with length equal to \mathbf{a}_i) and in 2-D using point clouds (radius equal to \mathbf{a}_i).

An arbitrary measure $\alpha \in \mathcal{M}(\mathcal{X})$ (which needs not to have a density nor be a sum of Diracs) is defined by the fact that it can be integrated against any continuous function $f \in \mathcal{C}(\mathcal{X})$ and obtain $\int_{\mathcal{X}} f(x)d\alpha(x) \in \mathbb{R}$. If \mathcal{X} is not compact, one should also impose that f has compact support or at least a 0 limit at infinity. Measure as thus in some sense “less regular” than functions, but more regular than distributions (which are dual to smooth functions). For instance, the derivative of a Dirac is not a measure. We denote $\mathcal{M}_+(\mathcal{X})$ the set of all positive measures on \mathcal{X} . The set of probability measures is denoted $\mathcal{M}_+^1(\mathcal{X})$, which means that any $\alpha \in \mathcal{M}_+^1(\mathcal{X})$ is positive, and that $\alpha(\mathcal{X}) = \int_{\mathcal{X}} d\alpha = 1$. Figure 18.1 offers a visualization of the different classes of measures, beyond histograms, considered in this work.

Operators on measures. For some continuous map $T : \mathcal{X} \rightarrow \mathcal{Y}$, we define the pushforward operator $T_\sharp : \mathcal{M}(\mathcal{X}) \rightarrow \mathcal{M}(\mathcal{Y})$. For discrete measures (18.1), the pushforward operation consists simply in moving the positions of all the points in the support of the measure

$$T_\sharp \alpha \stackrel{\text{def.}}{=} \sum_i \mathbf{a}_i \delta_{T(x_i)}.$$

For more general measures, for instance for those with a density, the notion of push-forward plays a fundamental to describe spatial modifications of probability measures. The formal definition reads as follow.

Definition 23 (Push-forward). *For $T : \mathcal{X} \rightarrow \mathcal{Y}$, the push forward measure $\beta = T_\sharp \alpha \in \mathcal{M}(\mathcal{Y})$ of some $\alpha \in \mathcal{M}(\mathcal{X})$ reads*

$$\forall h \in \mathcal{C}(\mathcal{Y}), \quad \int_{\mathcal{Y}} h(y)d\beta(y) = \int_{\mathcal{X}} h(T(x))d\alpha(x). \quad (18.2)$$

Equivalently, for any measurable set $B \subset \mathcal{Y}$, one has

$$\beta(B) = \alpha(\{x \in \mathcal{X} ; T(x) \in B\}). \quad (18.3)$$

Note that T_\sharp preserves positivity and total mass, so that if $\alpha \in \mathcal{M}_+^1(\mathcal{X})$ then $T_\sharp \alpha \in \mathcal{M}_+^1(\mathcal{Y})$.

Intuitively, a measurable map $T : \mathcal{X} \rightarrow \mathcal{Y}$, can be interpreted as a function “moving” a single point from a measurable space to another. The more general extension T_\sharp can now “move” an entire probability measure on \mathcal{X} towards a new probability measure on \mathcal{Y} . The operator T_\sharp “pushes forward” each elementary mass of a measure α on \mathcal{X} by applying the map T to obtain then an elementary mass in \mathcal{Y} , to build on aggregate a new measure on \mathcal{Y} written $T_\sharp \alpha$. Note that such a push-forward $T_\sharp : \mathcal{M}_+^1(\mathcal{X}) \rightarrow \mathcal{M}_+^1(\mathcal{Y})$ is a linear operator between measures in the sense that for two measures α_1, α_2 on \mathcal{X} , $T_\sharp(\alpha_1 + \alpha_2) = T_\sharp \alpha_1 + T_\sharp \alpha_2$.

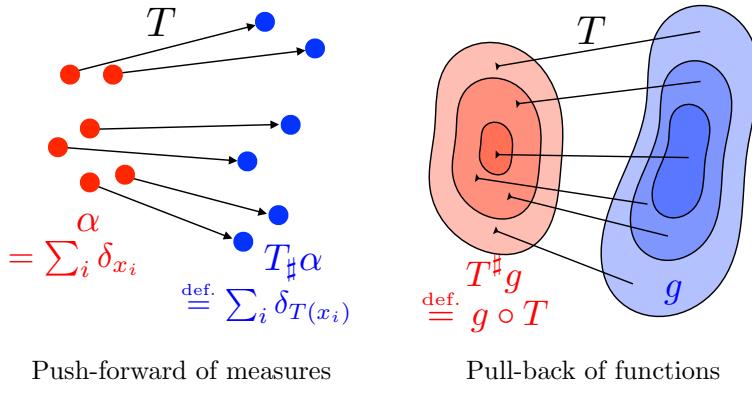


Figure 18.2: Comparison of push-forward $T_{\#}$ and pull-back $T^{\#}$.

Remark 10 (Push-forward for densities). Explicitly doing the change of variable in formula (18.2) for measures with densities $(\rho_{\alpha}, \rho_{\beta})$ on \mathbb{R}^d (assuming T is smooth and a bijection) shows that a push-forward acts on densities linearly as a change of variables in the integration formula, indeed

$$\rho_{\alpha}(x) = |\det(T'(x))| \rho_{\beta}(T(x)) \quad (18.4)$$

where $T'(x) \in \mathbb{R}^{d \times d}$ is the Jacobian matrix of T (the matrix formed by taking the gradient of each coordinate of T). This implies, denoting $y = T(x)$

$$|\det(T'(x))| = \frac{\rho_{\alpha}(x)}{\rho_{\beta}(y)}.$$

Remark 11 (Push-forward vs. pull-back). The push-forward $T_{\#}$ of measures should not be confounded with the pull-back of function $T^{\#} : \mathcal{C}(\mathcal{Y}) \rightarrow \mathcal{C}(\mathcal{X})$ which corresponds to the “warping” of functions. It is the linear map defined, for $g \in \mathcal{C}(\mathcal{Y})$ by $T^{\#}g = g \circ T$. Push-forward and pull-back are actually adjoint one from each others, in the sense that

$$\forall (\alpha, g) \in \mathcal{M}(\mathcal{X}) \times \mathcal{C}(\mathcal{Y}), \quad \int_{\mathcal{Y}} g d(T_{\#}\alpha) = \int_{\mathcal{X}} (T^{\#}g) d\alpha.$$

It is important to realize that even if (α, β) have densities $(\rho_{\alpha}, \rho_{\beta})$, $T_{\#}\alpha$ is not equal to $T^{\#}\rho_{\beta}$, because of the presence of the Jacobian in (18.4). This explains why OT should be used with caution to perform image registration, because it does not operate as an image warping method. Figure 18.2 illustrate the distinction between these push-forward and pull-back operators.

Remark 12 (Measures and random variables). Radon measures can also be viewed as representing the distributions of random variables. A random variable X on \mathcal{X} is actually a map $X : \Omega \rightarrow \mathcal{X}$ from some abstract (often un-specified) probabilized space (Ω, \mathbb{P}) , and its distribution α is the Radon measure $X \in \mathcal{M}_+^1(\mathcal{X})$ such that $\mathbb{P}(X \in A) = \alpha(A) = \int_A d\alpha(x)$. Equivalently, it is the push-forward of \mathbb{P} by X , $\alpha = X_{\#}\mathbb{P}$. Applying another push-forward $\beta = T_{\#}\alpha$ for $T : \mathcal{X} \rightarrow \mathcal{Y}$, following (18.2), is equivalent to defining another random variable $Y = T(X) : \omega \in \Omega \rightarrow T(X(\omega)) \in \mathcal{Y}$, so that β is the distribution of Y . Drawing a random sample y from Y is thus simply achieved by computing $y = T(x)$ where x is drawn from X .

Convergence of random variable. Convergence of random variable (in probability, almost sure, in law), convergence of measures (strong, weak).

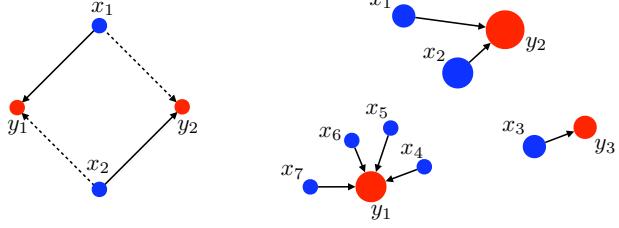


Figure 18.3: (left) blue dots from measure α and red dots from measure β are pairwise equidistant. Hence, either matching $\sigma = (1, 2)$ (full line) or $\sigma = (2, 1)$ (dotted line) is optimal. (right) a Monge map can associate the blue measure α to the red measure β . The weights α_i are displayed proportionally to the area of the disk marked at each location. The mapping here is such that $T(x_1) = T(x_2) = y_2$, $T(x_3) = y_3$, whereas for $4 \leq i \leq 7$ we have $T(x_i) = y_1$.

18.2 Monge Problem

Given a cost matrix $(\mathbf{C}_{i,j})_{i \in \llbracket n \rrbracket, j \in \llbracket m \rrbracket}$, assuming $n = m$, the optimal assignment problem seeks for a bijection σ in the set $\text{Perm}(n)$ of permutations of n elements solving

$$\min_{\sigma \in \text{Perm}(n)} \frac{1}{n} \sum_{i=1}^n \mathbf{C}_{i, \sigma(i)}. \quad (18.5)$$

One could naively evaluate the cost function above using all permutations in the set $\text{Perm}(n)$. However, that set has size $n!$, which is gigantic even for small n . Consider for instance that such a set has more than 10^{100} elements [?] when n is as small as 70. That problem can therefore only be solved if there exist efficient algorithms to optimize that cost function over the set of permutations, which will be the subject of §??.

Remark 13 (Uniqueness). Note that the optimal assignment problem may have several optimal solutions. Suppose for instance that $n = m = 2$ and that the matrix \mathbf{C} is the pairwise distance matrix between the 4 corners of a 2-dimensional square of side length 1, as represented in the left plot in Figure 18.3. In that case only two assignments exist, and they share the same cost.

For discrete measures

$$\alpha = \sum_{i=1}^n \mathbf{a}_i \delta_{x_i} \quad \text{and} \quad \beta = \sum_{j=1}^m \mathbf{b}_j \delta_{y_j} \quad (18.6)$$

the Monge problem [?] seeks for a map that associates to each point x_i a single point y_j , and which must push the mass of α toward the mass of β , which is to say that such a map $T : \{x_1, \dots, x_n\} \rightarrow \{y_1, \dots, y_m\}$ must verify that

$$\forall j \in \llbracket m \rrbracket, \quad \mathbf{b}_j = \sum_{i: T(x_i) = y_j} \mathbf{a}_i \quad (18.7)$$

which we write in compact form as $T_\sharp \alpha = \beta$. This map should minimize some transportation cost, which is parameterized by a function $c(x, y)$ defined for points $(x, y) \in \mathcal{X} \times \mathcal{Y}$

$$\min_T \left\{ \sum_i c(x_i, T(x_i)) ; T_\sharp \alpha = \beta \right\}. \quad (18.8)$$

Such a map between discrete points can be of course encoded, assuming all x 's and y 's are distinct, using indices $\sigma : \llbracket n \rrbracket \rightarrow \llbracket m \rrbracket$ so that $j = \sigma(i)$, and the mass conservation is written as

$$\sum_{i \in \sigma^{-1}(j)} \mathbf{a}_i = \mathbf{b}_j.$$

In the special case when $n = m$ and all weights are uniform, that is $\mathbf{a}_i = \mathbf{b}_j = 1/n$, then the mass conservation constraint implies that T is a bijection, such that $T(x_i) = y_{\sigma(i)}$, and the Monge problem is equivalent to the optimal matching problem (18.5) where the cost matrix is

$$\mathbf{C}_{i,j} \stackrel{\text{def}}{=} c(x_i, y_j).$$

When $n \neq m$, note that, optimality aside, Monge maps may not even exist between an empirical measure to another. This happens when their weight vectors are not compatible, which is always the case when the target measure has more points than the source measure. For instance, the right plot in Figure 18.3 shows an (optimal) Monge map between α and β , but there is no Monge map from β to α .

Monge problem (18.8) is extended to the setting of two arbitrary probability measures (α, β) on two spaces $(\mathcal{X}, \mathcal{Y})$ as finding a map $T : \mathcal{X} \rightarrow \mathcal{Y}$ that minimizes

$$\min_T \left\{ \int_{\mathcal{X}} c(x, T(x)) d\alpha(x) ; T_{\sharp}\alpha = \beta \right\} \quad (18.9)$$

The constraint $T_{\sharp}\alpha = \beta$ means that T pushes forward the mass of α to β , and makes use of the push-forward operator (18.2).

18.3 Kantorovitch Problem

The assignment problem has several limitations in practical settings, also encountered when using the Monge problem. Indeed, because the assignment problem is formulated as a permutation problem, it can only be used to compare two point clouds of the *same* size. A direct generalization to discrete measures with non-uniform weights can be carried out using Monge's formalism of pushforward maps, but that formulation may also be degenerate if there does not exist feasible solutions satisfying the mass conservation constraint (18.7) (see the end of Remark ??). Additionally, the assignment Problem (18.8) is combinatorial, whereas the feasible set for the Monge Problem (18.9), consisting in all push-forward measures that satisfy the mass conservation constraint, is *non-convex*. Both are therefore difficult to solve in their original formulation.

Kantorovitch formulation for discrete measures. The key idea of [?] is to relax the deterministic nature of transportation, namely the fact that a source point x_i can only be assigned to another, or transported to one and one location $T(x_i)$ only. Kantorovich proposes instead that the mass at any point x_i be potentially dispatched across several locations. Kantorovich moves away from the idea that mass transportation should be “deterministic” to consider instead a “probabilistic” (or “fuzzy”) transportation, which allows what is commonly known now as “mass splitting” from a source towards several targets. This flexibility is encoded using, in place of a permutation σ or a map T , a coupling matrix $\mathbf{P} \in \mathbb{R}_+^{n \times m}$, where $\mathbf{P}_{i,j}$ describes the amount of mass flowing from bin i (or point x_i) towards bin j (or point x_j), x_i towards y_j in the formalism of discrete measures (18.6). Admissible couplings admit a far simpler characterization than Monge maps:

$$\mathbf{U}(\mathbf{a}, \mathbf{b}) \stackrel{\text{def}}{=} \left\{ \mathbf{P} \in \mathbb{R}_+^{n \times m} ; \mathbf{P} \mathbf{1}_m = \mathbf{a} \text{ and } \mathbf{P}^T \mathbf{1}_n = \mathbf{b} \right\}, \quad (18.10)$$

where we used the following matrix-vector notation

$$\mathbf{P} \mathbf{1}_m = \left(\sum_j \mathbf{P}_{i,j} \right)_i \in \mathbb{R}^n \quad \text{and} \quad \mathbf{P}^T \mathbf{1}_n = \left(\sum_i \mathbf{P}_{i,j} \right)_j \in \mathbb{R}^m.$$

The set of matrices $\mathbf{U}(\mathbf{a}, \mathbf{b})$ is bounded, defined by $n + m$ equality constraints, and therefore a convex polytope (the convex hull of a finite set of matrices).

Additionally, whereas the Monge formulation (as illustrated in the right plot of Figure 18.3) was intrinsically asymmetric, Kantorovich's relaxed formulation is always symmetric, in the sense that a coupling \mathbf{P} is in $\mathbf{U}(\mathbf{a}, \mathbf{b})$ if and only if \mathbf{P}^T is in $\mathbf{U}(\mathbf{b}, \mathbf{a})$.

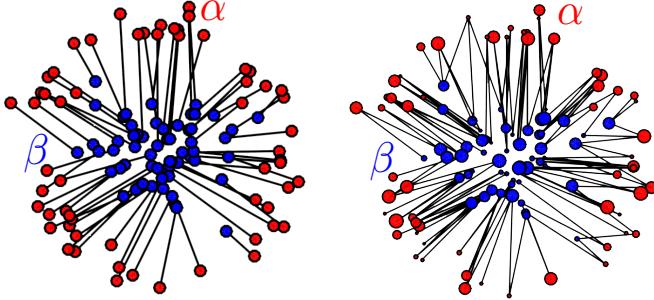


Figure 18.4: Comparison of optimal matching and generic couplings. A black segment between x_i and y_j indicates a non-zero element in the displayed optimal coupling $\mathbf{P}_{i,j}$ solving (18.11). Left: optimal matching, corresponding to the setting of Proposition (54) (empirical measures with the same number $n = m$ of points). Right: these two weighted point clouds cannot be matched; instead a Kantorovich coupling can be used to associate two arbitrary discrete measures.

Kantorovich's optimal transport problem now reads

$$L_{\mathbf{C}}(\mathbf{a}, \mathbf{b}) \stackrel{\text{def.}}{=} \min_{\mathbf{P} \in \mathbf{U}(\mathbf{a}, \mathbf{b})} \langle \mathbf{C}, \mathbf{P} \rangle \stackrel{\text{def.}}{=} \sum_{i,j} \mathbf{C}_{i,j} \mathbf{P}_{i,j}. \quad (18.11)$$

This is a linear program (see Chapter ??), and as is usually the case with such programs, its solutions are not necessarily unique.

Permutation Matrices as Couplings For a permutation $\sigma \in \text{Perm}(n)$, we write \mathbf{P}_σ for the corresponding permutation matrix,

$$\forall (i, j) \in \llbracket n \rrbracket^2, \quad (\mathbf{P}_\sigma)_{i,j} = \begin{cases} 1/n & \text{if } j = \sigma_i, \\ 0 & \text{otherwise.} \end{cases} \quad (18.12)$$

One can check that in that case

$$\langle \mathbf{C}, \mathbf{P}_\sigma \rangle = \frac{1}{n} \sum_{i=1}^n \mathbf{C}_{i,\sigma_i},$$

which shows that the assignment problem (18.5) can be recast as a Kantorovich problem (18.11) where the couplings \mathbf{P} are restricted to be exactly permutation matrices:

$$\min_{\sigma \in \text{Perm}(n)} \frac{1}{n} \sum_{i=1}^n \mathbf{C}_{i,\sigma(i)} = \min_{\sigma \in \text{Perm}(n)} \langle \mathbf{C}, \mathbf{P}_\sigma \rangle.$$

Next, one can easily check that the set of permutation matrices is strictly included in the so-called Birkhoff polytope $\mathbf{U}(\mathbb{1}_n/n, \mathbb{1}_n, n)$. Indeed, for any permutation σ we have $\mathbf{P}_\sigma \mathbb{1} = \mathbb{1}_n$ and $\mathbf{P}_\sigma^T \mathbb{1} = \mathbb{1}_n$, whereas $\mathbb{1}_n \mathbb{1}_n^T / n^2$ is a valid coupling but not a permutation matrix. Therefore, one has naturally that

$$\min_{\sigma \in \text{Perm}(n)} \langle \mathbf{C}, \mathbf{P}_\sigma \rangle \leq L_{\mathbf{C}}(\mathbb{1}_n/n, \mathbb{1}_n/n).$$

The following proposition shows that these problems result in fact in the same optimum, namely that one can always find a permutation matrix that minimizes Kantorovich's problem (18.11) between two uniform measures $\mathbf{a} = \mathbf{b} = \mathbb{1}_n/n$, which shows that the Kantorovich relaxation is *tight* when considered on assignment problems. Figure 18.4 shows on the left a 2-D example of optimal matching corresponding to this special case.

Proposition 54 (Kantorovich for matching). *If $m = n$ and $\mathbf{a} = \mathbf{b} = \mathbb{1}_n/n$, then there exists an optimal solution for Problem (18.11) \mathbf{P}_{σ^*} , which is a permutation matrix associated to an optimal permutation $\sigma^* \in \text{Perm}(n)$ for Problem (18.5).*

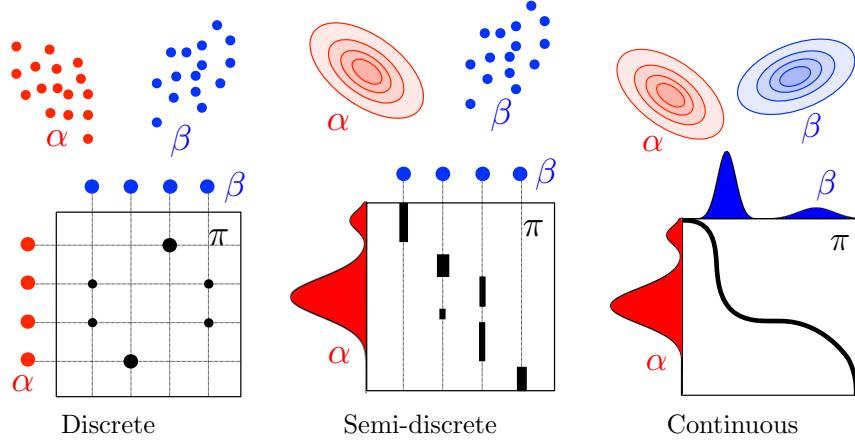


Figure 18.5: Schematic view of input measures (α, β) and couplings $\mathcal{U}(\alpha, \beta)$ encountered in the three main scenarios for Kantorovich OT. Chapter ?? is dedicated to the semi-discrete setup.

Proof. Birkhoff's theorem states that the set of extremal points of $\mathbf{U}(\mathbb{1}_n/n, \mathbb{1}_n/n)$ is equal to the set of permutation matrices. A fundamental theorem of linear programming [?, Theorem 2.7] states that the minimum of a linear objective in a non-empty polyhedron, if finite, is reached at an extremal point of the polyhedron. \square

Kantorovitch formulation for arbitrary measures. The definition of \mathcal{L}_c in (??) can be extended to arbitrary measures by considering couplings $\pi \in \mathcal{M}_+^1(\mathcal{X} \times \mathcal{Y})$ which are joint distributions over the product space. The discrete case is a special situation where one imposes this product measure to be of the form $\pi = \sum_{i,j} \mathbf{P}_{i,j} \delta_{(x_i, y_j)}$. In the general case, the mass conservation constraint (18.10) should be rewritten as a marginal constraint on joint probability distributions

$$\mathcal{U}(\alpha, \beta) \stackrel{\text{def.}}{=} \{ \pi \in \mathcal{M}_+^1(\mathcal{X} \times \mathcal{Y}) ; P_{\mathcal{X}\sharp}\pi = \alpha \text{ and } P_{\mathcal{Y}\sharp}\pi = \beta \}. \quad (18.13)$$

Here $P_{\mathcal{X}\sharp}$ and $P_{\mathcal{Y}\sharp}$ are the push-forward (see Definition 23) by the projections $P_{\mathcal{X}}(x, y) = x$ and $P_{\mathcal{Y}}(x, y) = y$. Figure 18.5 shows a schematic visualization of the coupling constraints for different class of problem (discrete measures and densities). Using (18.3), these marginal constraints are equivalent to imposing that $\pi(A \times \mathcal{Y}) = \alpha(A)$ and $\pi(\mathcal{X} \times B) = \beta(B)$ for sets $A \subset \mathcal{X}$ and $B \subset \mathcal{Y}$.

The Kantorovich problem (18.11) is then generalized as

$$\mathcal{L}_c(\alpha, \beta) \stackrel{\text{def.}}{=} \min_{\pi \in \mathcal{U}(\alpha, \beta)} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) d\pi(x, y). \quad (18.14)$$

This is an infinite-dimensional linear program over a space of measures. Figure 18.6 shows examples of discrete and continuous optimal coupling solving (18.14). Figure 18.7 shows other examples of optimal 1-D couplings, involving discrete and continuous marginals.

On compact domain $(\mathcal{X}, \mathcal{Y})$, (18.14) always has a solution, because using the weak-* topology (so called weak topology of measures), the set of measure is compact, and a linear function with a continuous $c(x, y)$ is weak-* continuous. And the set of constraint is non empty, taking $\alpha \otimes \beta$. On non compact domain, needs to impose moment condition on α and β .

Wasserstein distances. An important feature of OT is that it defines a distance between histograms and probability measures as soon as the cost matrix satisfies certain suitable properties. Indeed, OT can be understood as a canonical way to lift a ground distance between points to a distance between histogram or measures.

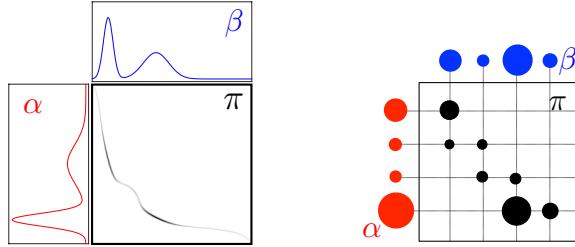


Figure 18.6: Left: “continuous” coupling π solving (18.13) between two 1-D measure with density. The coupling is localized along the graph of the Monge map $(x, T(x))$ (displayed in black). Right: “discrete” coupling T solving (18.11) between two discrete measures of the form (18.6). The non-zero entries $T_{i,j}$ are display with a black disk at position (i, j) with radius proportional to $T_{i,j}$.

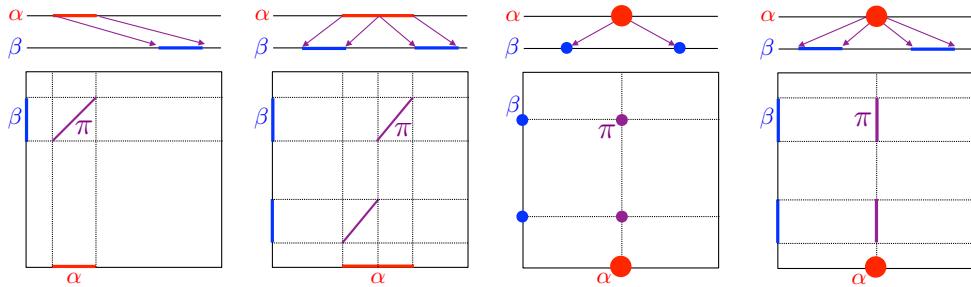


Figure 18.7: Four simple examples of optimal couplings between 1-D distributions, represented as maps above (arrows) and couplings below. Inspired by [?].

We first consider the case where, using a term first introduce by [?], the “ground metric” matrix \mathbf{C} is fixed, representing substitution costs between bins, and shared across several histograms we would like to compare. The following proposition states that OT provides a meaningful distance between histograms supported on these bins.

Proposition 55. *We suppose $n = m$, and that for some $p \geq 1$, $\mathbf{C} = \mathbf{D}^p = (\mathbf{D}_{i,j}^p)_{i,j} \in \mathbb{R}^{n \times n}$ where $\mathbf{D} \in \mathbb{R}_+^{n \times n}$ is a distance on $\llbracket n \rrbracket$, i.e.*

1. $\mathbf{D} \in \mathbb{R}_+^{n \times n}$ is symmetric;
2. $\mathbf{D}_{i,j} = 0$ if and only if $i = j$;
3. $\forall (i, j, k) \in \llbracket n \rrbracket^3, \mathbf{D}_{i,k} \leq \mathbf{D}_{i,j} + \mathbf{D}_{j,k}$.

Then

$$W_p(\mathbf{a}, \mathbf{b}) \stackrel{\text{def.}}{=} L_{\mathbf{D}^p}(\mathbf{a}, \mathbf{b})^{1/p} \quad (18.15)$$

(note that W_p depends on \mathbf{D}) defines the p -Wasserstein distance on Σ_n , i.e. W_p is symmetric, positive, $W_p(\mathbf{a}, \mathbf{b}) = 0$ if and only if $\mathbf{a} = \mathbf{b}$, and it satisfies the triangle inequality

$$\forall \mathbf{a}, \mathbf{a}', \mathbf{b} \in \Sigma_n, \quad W_p(\mathbf{a}, \mathbf{b}) \leq W_p(\mathbf{a}, \mathbf{a}') + W_p(\mathbf{a}', \mathbf{b}).$$

Proof. Symmetry and definiteness of the distance are easy to prove: since $\mathbf{C} = \mathbf{D}^p$ has a null diagonal, $W_p(\mathbf{a}, \mathbf{a}) = 0$, with corresponding optimal transport matrix $\mathbf{P}^* = \text{diag}(\mathbf{a})$; by the positivity of all off-diagonal elements of \mathbf{D}^p , $W_p(\mathbf{a}, \mathbf{b}) > 0$ whenever $\mathbf{a} \neq \mathbf{b}$ (because in this case, an admissible coupling necessarily has a non-zero element outside the diagonal); by symmetry of \mathbf{D}^p , $W_p(\mathbf{a}, \mathbf{b}) = 0$ is itself a symmetric function.

To prove the triangle inequality of Wasserstein distances for arbitrary measures, [?, Theorem 7.3] uses the gluing lemma, which stresses the existence of couplings with a prescribed structure. In the discrete setting,

the explicit construction of this glued coupling is simple. Let $\mathbf{a}, \mathbf{b}, \mathbf{c} \in \Sigma_n$. Let \mathbf{P} and \mathbf{Q} be two optimal solutions of the transport problems between \mathbf{a} and \mathbf{b} , and \mathbf{b} and \mathbf{c} respectively. We define $\bar{\mathbf{b}}_j \stackrel{\text{def.}}{=} \mathbf{b}_j$ if $\mathbf{b}_j > 0$ and set otherwise $\bar{\mathbf{b}}_j = 1$ (or actually any other value). We then define

$$\mathbf{S} \stackrel{\text{def.}}{=} \mathbf{P} \operatorname{diag}(1/\bar{\mathbf{b}}) \mathbf{Q} \in \mathbb{R}_+^{n \times n}.$$

We remark that $\mathbf{S} \in U(\mathbf{a}, \mathbf{c})$ because

$$\mathbf{S}\mathbf{1}_n = \mathbf{P} \operatorname{diag}(1/\bar{\mathbf{b}}) \mathbf{Q}\mathbf{1}_n = \mathbf{P}(\mathbf{b}/\bar{\mathbf{b}}) = \mathbf{P}\mathbf{1}_{\operatorname{Supp}(\mathbf{b})} = \mathbf{a}$$

where we denoted $\mathbf{1}_{\operatorname{Supp}(\mathbf{b})}$ the indicator of the support of \mathbf{b} , and we use the fact that $\mathbf{P}\mathbf{1}_{\operatorname{Supp}(\mathbf{b})} = \mathbf{P}\mathbf{1} = \mathbf{b}$ because necessarily $\mathbf{P}_{i,j} = 0$ for $j \notin \operatorname{Supp}(\mathbf{b})$. Similarly one verifies that $\mathbf{S}^\top \mathbf{1}_n = \mathbf{c}$.

The triangle inequality follows from

$$\begin{aligned} W_p(\mathbf{a}, \mathbf{c}) &= \left(\min_{\mathbf{P} \in U(\mathbf{a}, \mathbf{c})} \langle \mathbf{P}, \mathbf{D}^p \rangle \right)^{1/p} \leq \langle \mathbf{S}, \mathbf{D}^p \rangle^{1/p} \\ &= \left(\sum_{ik} \mathbf{D}_{ik}^p \sum_j \frac{\mathbf{P}_{ij} \mathbf{Q}_{jk}}{\bar{\mathbf{b}}_j} \right)^{1/p} \leq \left(\sum_{ijk} (\mathbf{D}_{ij} + \mathbf{D}_{jk})^p \frac{\mathbf{P}_{ij} \mathbf{Q}_{jk}}{\bar{\mathbf{b}}_j} \right)^{1/p} \\ &\leq \left(\sum_{ijk} \mathbf{D}_{ij}^p \frac{\mathbf{P}_{ij} \mathbf{Q}_{jk}}{\bar{\mathbf{b}}_j} \right)^{1/p} + \left(\sum_{ijk} \mathbf{D}_{jk}^p \frac{\mathbf{P}_{ij} \mathbf{Q}_{jk}}{\bar{\mathbf{b}}_j} \right)^{1/p} \\ &= \left(\sum_{ij} \mathbf{D}_{ij}^p \mathbf{P}_{ij} \sum_k \frac{\mathbf{Q}_{jk}}{\bar{\mathbf{b}}_j} \right)^{1/p} + \left(\sum_{jk} \mathbf{D}_{jk}^p \mathbf{Q}_{jk} \sum_i \frac{\mathbf{P}_{ij}}{\bar{\mathbf{b}}_j} \right)^{1/p} \\ &= \left(\sum_{ij} \mathbf{D}_{ij}^p \mathbf{P}_{ij} \right)^{1/p} + \left(\sum_{jk} \mathbf{D}_{jk}^p \mathbf{Q}_{jk} \right)^{1/p} \\ &= W_p(\mathbf{a}, \mathbf{b}) + W_p(\mathbf{b}, \mathbf{c}). \end{aligned}$$

The first inequality is due to the suboptimality of \mathbf{S} , the second is the usual triangle inequality for elements in \mathbf{D} , and the third comes from Minkowski's inequality. \square

Proposition 55 generalizes from histogram to arbitrary measures that need not be discrete.

Proposition 56. *We assume $\mathcal{X} = \mathcal{Y}$, and that for some $p \geq 1$, $c(x, y) = d(x, y)^p$ where d is a distance on \mathcal{X} , i.e.*

- (i) $d(x, y) = d(y, x) \geq 0$;
- (ii) $d(x, y) = 0$ if and only if $x = y$;
- (iii) $\forall (x, y, z) \in \mathcal{X}^3, d(x, z) \leq d(x, y) + d(y, z)$.

Then

$$W_p(\alpha, \beta) \stackrel{\text{def.}}{=} \mathcal{L}_{d^p}(\alpha, \beta)^{1/p} \quad (18.16)$$

(note that W_p depends on d) defines the p -Wasserstein distance on \mathcal{X} , i.e. W_p is symmetric, positive, $W_p(\alpha, \beta) = 0$ if and only if $\alpha = \beta$, and it satisfies the triangle inequality

$$\forall (\alpha, \beta, \gamma) \in \mathcal{M}_+^1(\mathcal{X})^3, \quad W_p(\alpha, \gamma) \leq W_p(\alpha, \beta) + W_p(\beta, \gamma).$$

Proof. The proof follows the same approach as that for Proposition 55 and relies on the existence of a coupling between (α, γ) obtained by “guying” optimal couplings between (α, β) and (β, γ) . \square

The Wasserstein distance \mathcal{W}_p has many important properties, the most important one being that it is a weak distance, *i.e.* it allows to compare singular distributions (for instance discrete ones) and to quantify spatial shift between the supports of the distributions. In particular, “classical” distances (or divergences) are not even defined between discrete distributions (the L^2 norm can only be applied to continuous measures with a density with respect to a base measure, and the discrete ℓ^2 norm requires the positions (x_i, y_j) to be fixed to work). In sharp contrast, one has that for any $p > 0$, $\mathcal{W}_p^p(\delta_x, \delta_y) = d(x, y)$. Indeed, it suffices to notice that $\mathcal{U}(\delta_x, \delta_y) = \{\delta_{x,y}\}$ and therefore the Kantorovich problem having only one feasible solution, $\mathcal{W}_p^p(\delta_x, \delta_y)$ is necessarily $(d(x, y)^p)^{1/p} = d(x, y)$. This shows that $\mathcal{W}_p(\delta_x, \delta_y) \rightarrow 0$ if $x \rightarrow y$. This property corresponds to the fact that \mathcal{W}_p is a way to quantify the weak convergence as we now define.

Definition 24 (Weak convergence). $(\alpha_k)_k$ converges weakly to α in $\mathcal{M}_+^1(\mathcal{X})$ (denoted $\alpha_k \rightharpoonup \alpha$) if and only if for any continuous function $g \in \mathcal{C}(\mathcal{X})$, $\int_{\mathcal{X}} g d\alpha_k \rightarrow \int_{\mathcal{X}} g d\alpha$. This notion of weak convergence corresponds to the convergence in law of random vectors.

This convergence can be shown to be equivalent to $\mathcal{W}_p(\alpha_k, \alpha) \rightarrow 0$ [?, Theorem 6.8] (together with a convergence of the moments up to order p for unbounded metric spaces).

Note that there exists alternative distances which also metrize weak convergence. The simplest one are Hilbertian norms, defined as

$$\|\alpha\|_k^2 \stackrel{\text{def.}}{=} \mathbb{E}_{\alpha \otimes \alpha}(k) = \int_{\mathcal{X} \times \mathcal{X}} k(x, y) d\alpha(x) d\alpha(y)$$

for a suitable choice of kernel $k : \mathcal{X}^2 \rightarrow \mathbb{R}$. The most famous of such kernel is the Gaussian one $k(x, y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}}$ for some choice of bandwidth $\sigma > 0$.

This convergence should not be confounded with the strong convergence of measures, which is metrized by the TV norm $\|\alpha\|_{TV} \stackrel{\text{def.}}{=} |\alpha|(\mathcal{X})$, which is the total mass of the absolute value of the measure.

Algorithms Since (??) is a linear program, it is possible to use any classical linear program solver, such as interior point methods or simplex. In practice, the network simplex is an efficient option, and it used pivoting rule adapted to the OT constraint set. In the case of the assignment problem, $\mathbf{a} = \mathbf{b} = \mathbf{1}_n/n$, there exists faster combinatorial optimization scheme, the most famous ones being the Hungarian algorithm and the auction algorithm, which have roughly $O(n^3)$ complexity. Section 18.5 details an approximate algorithm, which is typically faster, and amenable to parallelisation, but do not compute exactly the solution to the OT problem.

18.4 Duality

The Kantorovich problem (18.11) is a constrained convex minimization problem, and as such, it can be naturally paired with a so-called dual problem, which is a constrained concave maximization problem. The following fundamental proposition, which is a special case of Fenchel-Rockafellar duality theory, explains the relationship between the primal and dual problems.

Proposition 57. One has

$$L_C(\mathbf{a}, \mathbf{b}) = \max_{(\mathbf{f}, \mathbf{g}) \in \mathbf{R}(\mathbf{a}, \mathbf{b})} \langle \mathbf{f}, \mathbf{a} \rangle + \langle \mathbf{g}, \mathbf{b} \rangle \quad (18.17)$$

where the set of admissible potentials is

$$\mathbf{R}(\mathbf{a}, \mathbf{b}) \stackrel{\text{def.}}{=} \{(\mathbf{f}, \mathbf{g}) \in \mathbb{R}^n \times \mathbb{R}^m ; \forall (i, j) \in [\![n]\!] \times [\![m]\!], \mathbf{f} \oplus \mathbf{g} \leqslant \mathbf{C}\} \quad (18.18)$$

Proof. This result is a direct consequence of the more general result on the strong duality for linear programs [?, p.148, Theo.4.4]. The easier part of that result, namely that the right-hand side of Equation (18.17)

is a lower bound on $L_C(\mathbf{a}, \mathbf{b})$ is discussed in ???. For the sake of completeness, let us derive this dual problem with the use of Lagrangian duality. The Lagrangian associate to (18.11) reads

$$\min_{\mathbf{P} \geq 0} \max_{(\mathbf{f}, \mathbf{g}) \in \mathbb{R}^n \times \mathbb{R}^m} \langle \mathbf{C}, \mathbf{P} \rangle + \langle \mathbf{a} - \mathbf{P}\mathbf{1}_m, \mathbf{f} \rangle + \langle \mathbf{b} - \mathbf{P}^\top \mathbf{1}_n, \mathbf{g} \rangle. \quad (18.19)$$

For linear program, one can always exchange the min and the max and get the same value of the linear program, and one thus consider

$$\max_{(\mathbf{f}, \mathbf{g}) \in \mathbb{R}^n \times \mathbb{R}^m} \langle \mathbf{a}, \mathbf{f} \rangle + \langle \mathbf{b}, \mathbf{g} \rangle + \min_{\mathbf{P} \geq 0} \langle \mathbf{C} - \mathbf{f}\mathbf{1}_m^\top - \mathbf{1}_n\mathbf{g}^\top, \mathbf{P} \rangle.$$

We conclude by remarking that

$$\min_{\mathbf{P} \geq 0} \langle \mathbf{Q}, \mathbf{P} \rangle = \begin{cases} 0 & \text{if } \mathbf{Q} \geq 0 \\ -\infty & \text{otherwise} \end{cases}$$

so that the constraint reads $\mathbf{C} - \mathbf{f}\mathbf{1}_m^\top - \mathbf{1}_n\mathbf{g}^\top = \mathbf{C} - \mathbf{f} \oplus \mathbf{g} \geq 0$. \square

The primal-dual optimality relation for the Lagrangian (18.19) allows to locate the support of the optimal transport plan

$$\text{Supp}(\mathbf{P}) \subset \{(i, j) \in [\![n]\!] \times [\![m]\!]; \mathbf{f}_i + \mathbf{g}_j = \mathbf{C}_{i,j}\}. \quad (18.20)$$

To extend this primal-dual construction to arbitrary measures, it is important to realize that measures are naturally paired in duality with continuous functions (a measure can only be accessed through integration against continuous functions). The duality is formalized in the following proposition, which boils down to Proposition 57 when dealing with discrete measures.

Proposition 58. *One has*

$$\mathcal{L}_c(\alpha, \beta) = \max_{(f, g) \in \mathcal{R}(c)} \int_{\mathcal{X}} f(x) d\alpha(x) + \int_{\mathcal{Y}} g(y) d\beta(y), \quad (18.21)$$

where the set of admissible dual potentials is

$$\mathcal{R}(c) \stackrel{\text{def.}}{=} \{(f, g) \in \mathcal{C}(\mathcal{X}) \times \mathcal{C}(\mathcal{Y}); \forall (x, y), f(x) + g(y) \leq c(x, y)\}. \quad (18.22)$$

Here, (f, g) is a pair of continuous functions, and are often called “Kantorovich potentials”.

The discrete case (18.17) corresponds to the dual vectors being samples of the continuous potentials, i.e. $(\mathbf{f}_i, \mathbf{g}_j) = (f(x_i), g(y_j))$. The primal-dual optimality conditions allow to track the support of optimal plan, and (18.20) is generalized as

$$\text{Supp}(\pi) \subset \{(x, y) \in \mathcal{X} \times \mathcal{Y}; f(x) + g(y) = c(x, y)\}. \quad (18.23)$$

Note that in contrast to the primal problem (18.14), showing the existence of solutions to (18.21) is non-trivial, because the constraint set $\mathcal{R}(c)$ is not compact and the function to minimize non-coercive. Using the machinery of c -transform detailed in Section ??, one can however show that optimal (f, g) are necessarily Lipschitz regular, which enable to replace the constraint by a compact one.

Benier’s Theorem and Monge-Ampère PDE The following celebrated theorem of [?] ensures that in \mathbb{R}^d for $p = 2$, if at least one of the two inputs measures has a density, then Kantorovitch and Monge problems are equivalent.

Theorem 53 (Brenier). *In the case $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$ and $c(x, y) = \|x - y\|^2$, if at least one of the two inputs measures (denoted α) has a density ρ_α with respect to the Lebesgue measure, then the optimal π in the Kantorovich formulation (18.14) is unique, and is supported on the graph $(x, T(x))$ of a “Monge map” $T: \mathbb{R}^d \rightarrow \mathbb{R}^d$. This means that $\pi = (\text{Id}, T)_\sharp \mu$, i.e.*

$$\forall h \in \mathcal{C}(\mathcal{X} \times \mathcal{Y}), \quad \int_{\mathcal{X} \times \mathcal{Y}} h(x, y) d\pi(x, y) = \int_{\mathcal{X}} h(x, T(x)) d\mu(x). \quad (18.24)$$

Furthermore, this map T is uniquely defined as the gradient of a convex function φ , $T(x) = \nabla\varphi(x)$, where φ is the unique (up to an additive constant) convex function such that $(\nabla\varphi)_\sharp\mu = \nu$. This convex function is related to the dual potential f solving (18.21) as $\varphi(x) = \frac{\|x\|^2}{2} - f(x)$.

Proof. We sketch the main ingredients of the proof, more details can be found for instance in [?]. We remark that $\int c d\pi = C_{\alpha,\beta} - 2 \int \langle x, y \rangle d\pi(x, y)$ where the constant is $C_{\alpha,\beta} = \int \|x\|^2 d\alpha(x) + \int \|y\|^2 d\beta(y)$. Instead of solving (18.14), one can thus consider the following problem

$$\max_{\pi \in \mathcal{U}(\alpha, \beta)} \int_{\mathcal{X} \times \mathcal{Y}} \langle x, y \rangle d\pi(x, y),$$

whose dual reads

$$\min_{(\varphi, \psi)} \left\{ \int_{\mathcal{X}} \varphi d\alpha + \int_{\mathcal{Y}} \psi d\beta ; \forall (x, y), \quad \varphi(x) + \psi(y) \geq \langle x, y \rangle \right\}. \quad (18.25)$$

The relation between these variables and those of (18.22) is $(\varphi, \psi) = (\frac{\|x\|^2}{2} - f, \frac{\|y\|^2}{2} - g)$. One can replace the constraint by

$$\forall y, \quad \psi(y) \geq \varphi^*(y) \stackrel{\text{def.}}{=} \sup_x \langle x, y \rangle - \varphi(x). \quad (18.26)$$

Here φ^* is the Legendre transform of φ and is a convex function as a supremum of linear forms (see also ??). Since the objective appearing in (18.27) is linear and the integrating measures positive, one can minimize explicitly with respect to φ and set $\psi = \varphi^*$ in order to consider the unconstraint problem

$$\min_{\varphi} \int_{\mathcal{X}} \varphi d\alpha + \int_{\mathcal{Y}} \varphi^* d\beta, \quad (18.27)$$

see also Section ?? for a generalization of this idea to generic costs $c(x, y)$. By iterating this argument twice, one can replace φ by φ^{**} , which is a convex function, and thus impose in (18.27) that φ is convex. Condition (18.23) shows that an optimal π is supported on $\{(x, y) ; \varphi(x) + \varphi^*(y) = \langle x, y \rangle\}$ which shows that such an y is optimal for the minimization (18.26) of the Legendre transform, whose optimality condition reads $y \in \partial\varphi(x)$. Since φ is convex, it is differentiable almost everywhere, and since α has a density, it is also differentiable α -almost everywhere. This shows that for each x , the associated y is uniquely defined α -almost everywhere as $y = \nabla\varphi(x)$, and shows that necessarily $\pi = (\text{Id}, \nabla\varphi)_\sharp\alpha$. \square

This results shows that in the setting of \mathcal{W}_2 with non-singular densities, the Monge problem (18.9) and its Kantorovich relaxation (18.14) are equal (the relaxation is tight). This is the continuous analog of Proposition 54 for the assignment case (54), which states that the minimum of the optimal transport problem is achieved, when the marginals are equal and uniform, at a permutation matrix (a discrete map). Brenier's theorem, stating that an optimal transport map must be the gradient of a convex function, should be examined under the light that a convex function is the natural generalization of the notion of increasing functions in dimension more than one. Optimal transport can thus plays an important role to define quantile functions in arbitrary dimensions, which in turn is useful for applications to quantile regression problems [?].

Note also that this theorem can be extended in many directions. The condition that α has a density can be weakened to the condition that it does not give mass to “small sets” having Hausdorff dimension smaller than $d - 1$ (e.g. hypersurfaces). One can also consider costs of the form $c(x, y) = h(x - y)$ where h is a strictly convex function.

For measures with densities, using (18.4), one obtains that φ is the unique (up to the addition of a constant) convex function which solves the following Monge-Ampère-type equation

$$\det(\partial^2\varphi(x))\rho_\beta(\nabla\varphi(x)) = \rho_\alpha(x) \quad (18.28)$$

where $\partial^2\varphi(x) \in \mathbb{R}^{d \times d}$ is the hessian of φ . The Monge-Ampère operator $\det(\partial^2\varphi(x))$ can be understood as a non-linear degenerate Laplacian. In the limit of small displacements, $\varphi = \text{Id} + \varepsilon\varphi$, one indeed recovers the Laplacian Δ as a linearization since for smooth maps

$$\det(\partial^2\varphi(x)) = 1 + \varepsilon\Delta\varphi(x) + o(\varepsilon).$$

The convexity constraint forces $\det(\partial^2\varphi(x)) \geq 0$ and is necessary for this equation to have a solution.

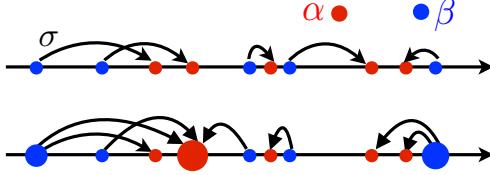


Figure 18.8: 1-D optimal couplings: each arrow $x_i \rightarrow y_j$ indicate a non-zero $\mathbf{P}_{i,j}$ in the optimal coupling. Top: empirical measures with same number of points (optimal matching). Bottom: generic case. This corresponds to monotone rearrangements, if $x_i \leq x_{i'}$ are such that $\mathbf{P}_{i,j} \neq 0, \mathbf{P}_{i',j'} \neq 0$, then necessarily $y_j \leq y_{j'}$.

Special cases In general, computing OT distances is numerically involved. We review special favorable cases where the resolution of the OT problem is easy.

Remark 14 (Binary Cost Matrix and 1-Norm). One can easily check that when the cost matrix \mathbf{C} is zero on the diagonal and 1 elsewhere, namely when $\mathbf{C} = \mathbf{1}_{n \times n} - I_n$, the OT distance between \mathbf{a} and \mathbf{b} is equal to the 1-norm of their difference, $L_{\mathbf{C}}(\mathbf{a}, \mathbf{b}) = \|\mathbf{a} - \mathbf{b}\|_1$. One can also easily check that this result extends to discrete and discrete measures in the case where $c(x, y)$ is 0 if $x = y$ and 1 when $x \neq y$. The OT distance between two discrete measures α and β is equal to their total variation distance.

Remark 15 (1-D case – Empirical measures). Here $\mathcal{X} = \mathbb{R}$. Assuming $\alpha = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$ and $\beta = \frac{1}{n} \sum_{j=1}^n \delta_{y_j}$, and assuming (without loss of generality) that the points are ordered, i.e. $x_1 \leq x_2 \leq \dots \leq x_n$ and $y_1 \leq y_2 \leq \dots \leq y_n$, then one has the simple formula

$$\mathcal{W}_p(\alpha, \beta)^p = \sum_{i=1}^p |x_i - y_i|^p, \quad (18.29)$$

i.e. locally (if one assumes distinct points), $\mathcal{W}_p(\alpha, \beta)$ is the ℓ^p norm between two vectors of ordered values of α and β . That statement is only valid locally, in the sense that the order (and those vector representations) might change whenever some of the values change. That formula is a simple consequence of the more general remark given below. Figure 18.8, top row, illustrates the 1-D transportation map between empirical measures with the same number of points. The bottom row shows how this monotone map generalizes to arbitrary discrete measures. It is possible to leverage this 1-D computation to also compute efficiently OT on the circle, see [?]. Note that in the case of concave cost of the distance, for instance when $p < 1$, the behaviour of the optimal transport plan is very different, see [?], which describes an efficient solver in this case.

Remark 16 (1-D case – Generic case). For a measure α on \mathbb{R} , we introduce the cumulative function

$$\forall x \in \mathbb{R}, \quad \mathcal{C}_{\alpha}(x) \stackrel{\text{def.}}{=} \int_{-\infty}^x d\alpha, \quad (18.30)$$

which is a function $\mathcal{C}_{\alpha} : \mathbb{R} \rightarrow [0, 1]$, and its pseudo-inverse $\mathcal{C}_{\alpha}^{-1} : [0, 1] \rightarrow \mathbb{R} \cup \{-\infty\}$

$$\forall r \in [0, 1], \quad \mathcal{C}_{\alpha}^{-1}(r) = \min_x \{x \in \mathbb{R} \cup \{-\infty\} ; \mathcal{C}_{\alpha}(x) \geq r\}.$$

That function is also called the generalized quantile function of α . For any $p \geq 1$, one has

$$\mathcal{W}_p(\alpha, \beta)^p = \|\mathcal{C}_{\alpha}^{-1} - \mathcal{C}_{\beta}^{-1}\|_{L^p([0,1])}^p = \int_0^1 |\mathcal{C}_{\alpha}^{-1}(r) - \mathcal{C}_{\beta}^{-1}(r)|^p dr. \quad (18.31)$$

This means that through the map $\alpha \mapsto \mathcal{C}_{\alpha}^{-1}$, the Wasserstein distance is isometric to a linear space equipped with the L^p norm, or, equivalently, that the Wasserstein distance for measures on the real line is a Hilbertian metric. This makes the geometry of 1-D optimal transport very simple, but also very different from its

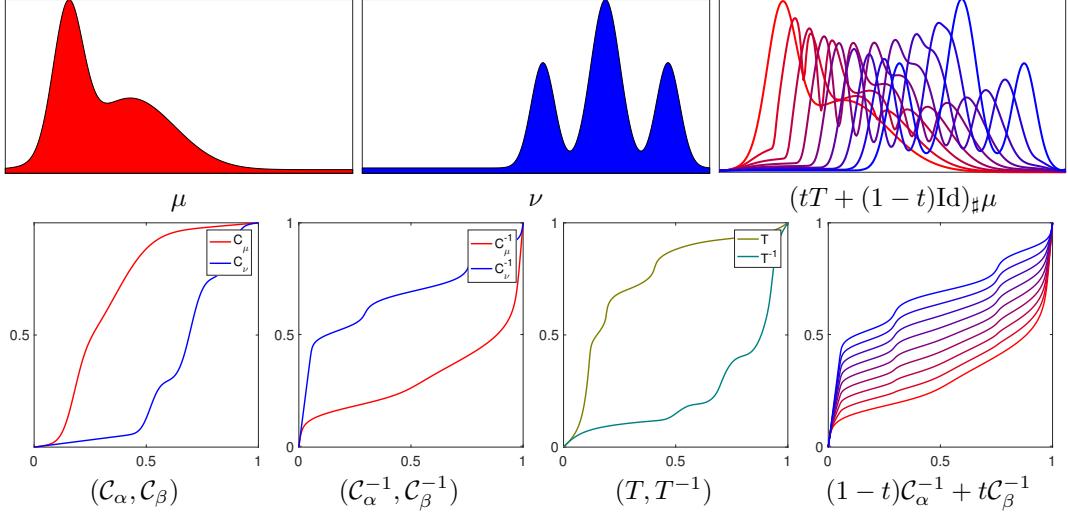


Figure 18.9: Computation of OT and displacement interpolation between two 1-D measures, using cumulant function as detailed in (18.34).

geometry in higher dimensions, which is not Hilbertian as discussed in Proposition ?? and more generally in §??. For $p = 1$, one even has the simpler formula

$$\mathcal{W}_1(\alpha, \beta) = \|\mathcal{C}_\alpha - \mathcal{C}_\beta\|_{L^1(\mathbb{R})} = \int_{\mathbb{R}} |\mathcal{C}_\alpha(x) - \mathcal{C}_\beta(x)| dx \quad (18.32)$$

$$= \int_{\mathbb{R}} \left| \int_{-\infty}^x d(\alpha - \beta) \right| dx. \quad (18.33)$$

which shows that \mathcal{W}_1 is a norm (see §?? for the generalization to arbitrary dimensions). An optimal Monge map T such that $T_\sharp \alpha = \beta$ is then defined by

$$T = \mathcal{C}_\beta^{-1} \circ \mathcal{C}_\alpha. \quad (18.34)$$

Figure 18.9 illustrates the computation of 1-D OT through cumulative functions. It also displays displacement interpolations, computed as detailed in (??), see also Remark ???. For a detailed survey of the properties of optimal transport in 1-D, we refer the reader to [?, Chapter 2].

Remark 17 (Distance between Gaussians). If $\alpha = \mathcal{N}(\mathbf{m}_\alpha, \Sigma_\alpha)$ and $\beta = \mathcal{N}(\mathbf{m}_\beta, \Sigma_\beta)$ are two Gaussians in \mathbb{R}^d , then one can show that the following map

$$T : x \mapsto \mathbf{m}_\beta + A(x - \mathbf{m}_\alpha), \quad (18.35)$$

where

$$A = \Sigma_\alpha^{-\frac{1}{2}} \left(\Sigma_\alpha^{\frac{1}{2}} \Sigma_\beta \Sigma_\alpha^{\frac{1}{2}} \right)^{\frac{1}{2}} \Sigma_\alpha^{-\frac{1}{2}} = A^T,$$

is such that $T_\sharp \rho_\alpha = \rho_\beta$. Indeed, one simply has to notice that the change of variables formula (18.4) is satisfied since

$$\begin{aligned} \rho_\beta(T(x)) &= \det(2\pi\Sigma_\beta)^{-\frac{1}{2}} \exp(-\langle T(x) - \mathbf{m}_\beta, \Sigma_\beta^{-1}(T(x) - \mathbf{m}_\beta) \rangle) \\ &= \det(2\pi\Sigma_\beta)^{-\frac{1}{2}} \exp(-\langle x - \mathbf{m}_\alpha, A^T \Sigma_\beta^{-1} A(x - \mathbf{m}_\alpha) \rangle) \\ &= \det(2\pi\Sigma_\beta)^{-\frac{1}{2}} \exp(-\langle x - \mathbf{m}_\alpha, \Sigma_\alpha^{-1}(x - \mathbf{m}_\alpha) \rangle), \end{aligned}$$

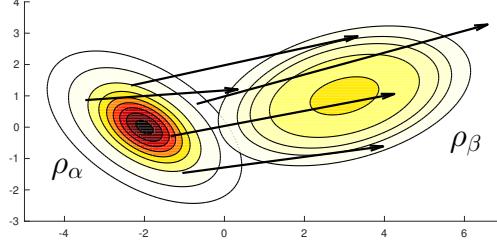


Figure 18.10: Two Gaussians ρ_α and ρ_β , represented using the contour plots of their densities, with respective mean and variance matrices $\mathbf{m}_\alpha = (-2, 0)$, $\Sigma_\alpha = \frac{1}{2} (1 - \frac{1}{2}; -\frac{1}{2} 1)$ and $\mathbf{m}_\beta = (3, 1)$, $\Sigma_\beta = (2, \frac{1}{2}; \frac{1}{2}, 1)$. The arrows originate at random points x taken on the plane and end at the corresponding mappings of those points $T(x) = \mathbf{m}_\beta + A(x - \mathbf{m}_\alpha)$.

and since T is a linear map we have that

$$|\det T'(x)| = \det A = \left(\frac{\det \Sigma_\beta}{\det \Sigma_\alpha} \right)^{\frac{1}{2}}$$

and we therefore recover $\rho_\alpha = |\det T'| \rho_\beta$ meaning $T_\sharp \alpha = \beta$. Notice now that T is the gradient of the convex function $\psi : x \mapsto \frac{1}{2} \langle x - \mathbf{m}_\alpha, A(x - \mathbf{m}_\alpha) \rangle + \langle \mathbf{m}_\beta, x \rangle$ to conclude, using Brenier's theorem [?] (see Remark ??) that T is optimal. Both that map T and the corresponding potential ψ are illustrated in Figures 18.10 and ??.

With additional calculations involving first and second order moments of ρ_α , we obtain that the transport cost of that map is

$$\mathcal{W}_2^2(\alpha, \beta) = \|\mathbf{m}_\alpha - \mathbf{m}_\beta\|^2 + \mathcal{B}(\Sigma_\alpha, \Sigma_\beta)^2 \quad (18.36)$$

where \mathcal{B} is the so-called Bures' metric [?] between positive definite matrices (see also [?, ?]),

$$\mathcal{B}(\Sigma_\alpha, \Sigma_\beta)^2 \stackrel{\text{def.}}{=} \text{tr} \left(\Sigma_\alpha + \Sigma_\beta - 2(\Sigma_\alpha^{1/2} \Sigma_\beta \Sigma_\alpha^{1/2})^{1/2} \right), \quad (18.37)$$

where $\Sigma^{1/2}$ is the matrix square root. One can show that \mathcal{B} is a distance on covariance matrices, and that \mathcal{B}^2 is convex with respect to both its arguments. In the case where $\Sigma_\alpha = \text{diag}(r_i)_i$ and $\Sigma_\beta = \text{diag}(s_i)_i$ are diagonals, the Bures metric is the Hellinger distance

$$\mathcal{B}(\Sigma_\alpha, \Sigma_\beta) = \|\sqrt{r} - \sqrt{s}\|_2.$$

For 1-D Gaussians, \mathcal{W}_2 is thus the Euclidean distance on the 2-D plane $(\mathbf{m}, \sqrt{\Sigma})$, as illustrated in Figure 18.11. For a detailed treatment of the Wasserstein geometry of Gaussian distributions, we refer to [?].

18.5 Sinkhorn

This section introduces a family of numerical scheme to approximate solutions to Kantorovich formulation of optimal transport and its many generalizations. It operates by adding an entropic regularization penalty to the original problem. This regularization has several important advantages, but a few stand out particularly: The minimization of the regularized problem can be solved using a simple alternate minimization scheme; that scheme translates into iterations that are simple matrix products, making them particularly suited to execution of GPU; the resulting approximate distance is smooth with respect to input histogram weights and positions of the Diracs.

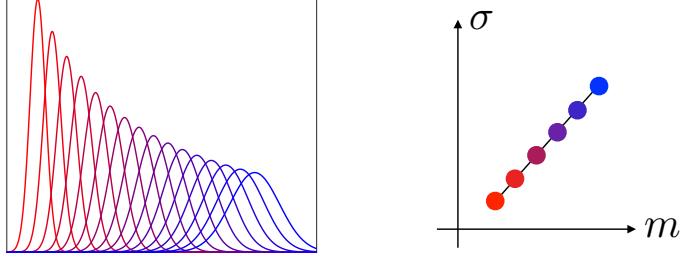


Figure 18.11: Computation of displacement interpolation between two 1-D Gaussians. Denoting $\mathcal{G}_{m,\sigma}(x) \stackrel{\text{def.}}{=} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}}$ the Gaussian density, it thus shows the interpolation $\mathcal{G}_{(1-t)m_0+tm_1,(1-t)\sigma_0+t\sigma_1}$.

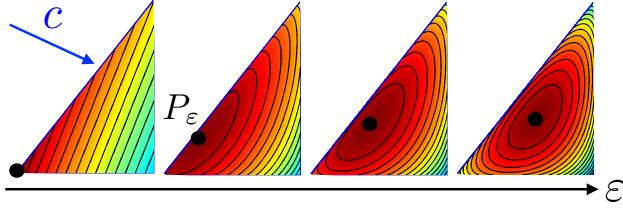


Figure 18.12: Impact of ε on the optimization of a linear function on the simplex, solving $\mathbf{P}_\varepsilon = \operatorname{argmin}_{\mathbf{P} \in \Sigma_3} \langle \mathbf{C}, \mathbf{P} \rangle - \varepsilon \mathbf{H}(\mathbf{P})$ for a varying ε .

Entropic Regularization. The discrete entropy of a coupling matrix is defined as

$$\mathbf{H}(\mathbf{P}) \stackrel{\text{def.}}{=} - \sum_{i,j} \mathbf{P}_{i,j} (\log(\mathbf{P}_{i,j}) - 1), \quad (18.38)$$

with an analogous definition for vectors, with the convention that $\mathbf{H}(\mathbf{a}) = -\infty$ if one of the entries \mathbf{a}_j is 0 or negative. The function \mathbf{H} is 1-strongly concave, because its hessian is $\partial^2 \mathbf{H}(\mathbf{P}) = -\operatorname{diag}(1/\mathbf{P}_{i,j})$ and $\mathbf{P}_{i,j} \leq 1$. The idea of the entropic regularization of optimal transport is to use $-\mathbf{H}$ as a regularizing function to obtain approximate solutions to the original transport problem (18.11):

$$L_C^\varepsilon(\mathbf{a}, \mathbf{b}) \stackrel{\text{def.}}{=} \min_{\mathbf{P} \in U(\mathbf{a}, \mathbf{b})} \langle \mathbf{P}, \mathbf{C} \rangle - \varepsilon \mathbf{H}(\mathbf{P}). \quad (18.39)$$

Since the objective is a ε -strongly convex function, problem 18.39 has a unique optimal solution. The idea to regularize the optimal transport problem by an entropic term can be traced back to modeling ideas in transportation theory [?]: Actual traffic patterns in a network do not agree with those predicted by the solution of the optimal transport problem. Indeed, the former are more diffuse than the latter, which tend to rely on a few routes as a result of the sparsity of optimal couplings to the solution of 18.11. To balance for that, researchers in transportation proposed a model, called the “gravity” model [?], that is able to form a more “blurred” traffic prediction.

Figure 18.12 illustrates the effect of the entropy to regularize a linear program over the simples Σ_3 (which can thus be visualized as a triangle in 2-D). Note how the entropy pushes the original LP solution away from the boundary of the triangle. The optimal \mathbf{P}_ε progressively moves toward an “entropic center” of the triangle. This is further detailed in the proposition below. The convergence of the solution of that regularized problem towards an optimal solution of the original linear program has been studied by [?].

Proposition 59 (Convergence with ε). *The unique solution \mathbf{P}_ε of (18.39) converges to the optimal solution with maximal entropy within the set of all optimal solutions of the Kantorovich problem, namely*

$$\mathbf{P}_\varepsilon \xrightarrow{\varepsilon \rightarrow 0} \operatorname{argmin}_{\mathbf{P}} \{-\mathbf{H}(\mathbf{P}) ; \mathbf{P} \in U(\mathbf{a}, \mathbf{b}), \langle \mathbf{P}, \mathbf{C} \rangle = L_C(\mathbf{a}, \mathbf{b})\} \quad (18.40)$$

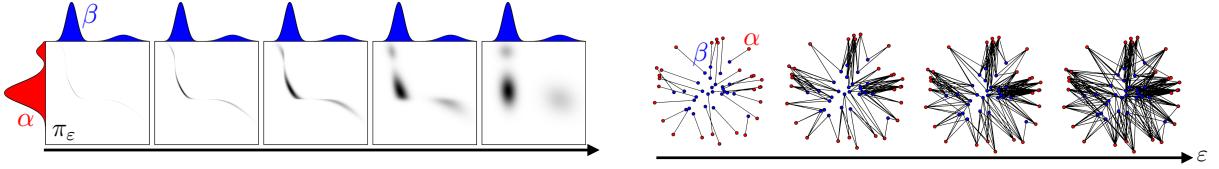


Figure 18.13: Impact of ε on coupling between densities and discrete distributions, illustrating Proposition 59. Left: between two 1-D densities. Right: between two 2-D discrete empirical densities with same number $n = m$ of points (only entries of the optimal $(\mathbf{P}_{i,j})_{i,j}$ above a small threshold are displayed as segments between x_i and y_j).

so that in particular

$$L_C^\varepsilon(\mathbf{a}, \mathbf{b}) \xrightarrow{\varepsilon \rightarrow 0} L_C(\mathbf{a}, \mathbf{b}).$$

One has

$$\mathbf{P}_\varepsilon \xrightarrow{\varepsilon \rightarrow \infty} \mathbf{ab}^T = (\mathbf{a}_i \mathbf{b}_j)_{i,j}. \quad (18.41)$$

Proof. We consider a sequence $(\varepsilon_\ell)_\ell$ such that $\varepsilon_\ell \rightarrow 0$ and $\varepsilon_\ell > 0$. We denote \mathbf{P}_ℓ the solution of (18.39) for $\varepsilon = \varepsilon_\ell$. Since $\mathbf{U}(\mathbf{a}, \mathbf{b})$ is bounded, we can extract a sequence (that we do not relabel for sake of simplicity) such that $\mathbf{P}_\ell \rightarrow \mathbf{P}^*$. Since $\mathbf{U}(\mathbf{a}, \mathbf{b})$ is closed, $\mathbf{P}^* \in \mathbf{U}(\mathbf{a}, \mathbf{b})$. We consider any \mathbf{P} such that $\langle \mathbf{C}, \mathbf{P} \rangle = L_C(\mathbf{a}, \mathbf{b})$. By optimality of \mathbf{P} and \mathbf{P}_ℓ for their respective optimization problems (for $\varepsilon = 0$ and $\varepsilon = \varepsilon_\ell$), one has

$$0 \leq \langle \mathbf{C}, \mathbf{P}_\ell \rangle - \langle \mathbf{C}, \mathbf{P} \rangle \leq \varepsilon_\ell (\mathbf{H}(\mathbf{P}_\ell) - \mathbf{H}(\mathbf{P})). \quad (18.42)$$

Since \mathbf{H} is continuous, taking the limit $\ell \rightarrow +\infty$ in this expression shows that $\langle \mathbf{C}, \mathbf{P}^* \rangle = \langle \mathbf{C}, \mathbf{P} \rangle$ so that \mathbf{P}^* is a feasible point of (18.40). Furthermore, dividing by ε_ℓ in (18.42) and taking the limit shows that $\mathbf{H}(\mathbf{P}) \leq \mathbf{H}(\mathbf{P}^*)$, which shows that \mathbf{P}^* is a solution of (18.40). Since the solution \mathbf{P}_0^* to this program is unique by strict convexity of $-\mathbf{H}$, one has $\mathbf{P}^* = \mathbf{P}_0^*$, and the whole sequence is converging. \square

Formula (18.40) states that for low regularization, the solution converges to the maximum entropy optimal transport coupling. In sharp contrast, (18.41) shows that for large regularization, the solution converges to the coupling with maximal entropy between two prescribed marginals \mathbf{a}, \mathbf{b} , namely the joint probability between two independent random variables with prescribed distributions. A refined analysis of this convergence is performed in [?], including a first order expansion in ε (resp. $1/\varepsilon$) near $\varepsilon = 0$ (resp $\varepsilon = +\infty$). Figure 18.13 shows visually the effect of these two convergence. A key insight is that, as ε increases, the optimal coupling becomes less and less sparse (in the sense of having entries larger than a prescribed thresholds), which in turn as the effect of both accelerating computational algorithms (as we study in §18.5) but also leading to faster statistical convergence (as exposed in §??).

Defining the Kullback-Leibler divergence between couplings as

$$\mathbf{KL}(\mathbf{P}|\mathbf{K}) \stackrel{\text{def.}}{=} \sum_{i,j} \mathbf{P}_{i,j} \log \left(\frac{\mathbf{P}_{i,j}}{\mathbf{K}_{i,j}} \right) - \mathbf{P}_{i,j} + \mathbf{K}_{i,j}, \quad (18.43)$$

the unique solution \mathbf{P}_ε of (18.39) is a projection onto $\mathbf{U}(\mathbf{a}, \mathbf{b})$ of the Gibbs kernel associated to the cost matrix \mathbf{C} as

$$\mathbf{K}_{i,j} \stackrel{\text{def.}}{=} e^{-\frac{\mathbf{C}_{i,j}}{\varepsilon}}$$

Indeed one has that using the definition above

$$\mathbf{P}_\varepsilon = \text{Proj}_{\mathbf{U}(\mathbf{a}, \mathbf{b})}^{\mathbf{KL}}(\mathbf{K}) \stackrel{\text{def.}}{=} \underset{\mathbf{P} \in \mathbf{U}(\mathbf{a}, \mathbf{b})}{\text{argmin}} \mathbf{KL}(\mathbf{P}|\mathbf{K}). \quad (18.44)$$

Remark 18 (General formulation). One can consider arbitrary measures by replacing the discrete entropy by the relative entropy with respect to the product measure $d\alpha \otimes d\beta(x, y) \stackrel{\text{def.}}{=} d\alpha(x)d\beta(y)$, and propose a regularized counterpart to (18.14) using

$$\mathcal{L}_c^\varepsilon(\alpha, \beta) \stackrel{\text{def.}}{=} \min_{\pi \in \mathcal{U}(\alpha, \beta)} \int_{X \times Y} c(x, y) d\pi(x, y) + \varepsilon \text{KL}(\pi | \alpha \otimes \beta) \quad (18.45)$$

where the relative entropy is a generalization of the discrete Kullback-Leibler divergence (18.43)

$$\begin{aligned} \text{KL}(\pi | \xi) &\stackrel{\text{def.}}{=} \int_{\mathcal{X} \times \mathcal{Y}} \log \left(\frac{d\pi}{d\xi}(x, y) \right) d\pi(x, y) + \\ &\quad \int_{\mathcal{X} \times \mathcal{Y}} (d\xi(x, y) - d\pi(x, y)), \end{aligned} \quad (18.46)$$

and by convention $\text{KL}(\pi | \xi) = +\infty$ if π does not have a density $\frac{d\pi}{d\xi}$ with respect to ξ . It is important to realize that the reference measure $\alpha \otimes \beta$ chosen in (18.45) to define the entropic regularizing term $\text{KL}(\cdot | \alpha \otimes \beta)$ plays no specific role, only its support matters.

Formula (18.45) can be re-factored as a projection problem

$$\min_{\pi \in \mathcal{U}(\alpha, \beta)} \text{KL}(\pi | \mathcal{K}) \quad (18.47)$$

where \mathcal{K} is the Gibbs distributions $d\mathcal{K}(x, y) \stackrel{\text{def.}}{=} e^{-\frac{c(x, y)}{\varepsilon}} d\mu(x) d\nu(y)$. This problem is often referred to as the “static Schrödinger problem” [?, ?], since it was initially considered by Schrödinger in statistical physics [?]. As $\varepsilon \rightarrow 0$, the unique solution to (18.47) converges to the maximum entropy solution to (18.14), see [?, ?]. §?? details an alternate “dynamic” formulation of the Schrödinger problem over the space of paths connecting the points of two measures.

Sinkhorn’s Algorithm The following proposition shows that the solution of (18.39) has a specific form, which can be parameterized using $n + m$ variables. That parameterization is therefore essentially dual, in the sense that a coupling \mathbf{P} in $\mathbf{U}(\mathbf{a}, \mathbf{b})$ has nm variables but $n + m$ constraints.

Proposition 60. *The solution to (18.39) is unique and has the form*

$$\forall (i, j) \in [\![n]\!] \times [\![m]\!], \quad \mathbf{P}_{i,j} = \mathbf{u}_i \mathbf{K}_{i,j} \mathbf{v}_j \quad (18.48)$$

for two (unknown) scaling variable $(\mathbf{u}, \mathbf{v}) \in \mathbb{R}_+^n \times \mathbb{R}_+^m$.

Proof. Introducing two dual variables $\mathbf{f} \in \mathbb{R}^n$, $\mathbf{g} \in \mathbb{R}^m$ for each marginal constraint, the Lagrangian of (18.39) reads

$$\mathcal{E}(\mathbf{P}, \mathbf{f}, \mathbf{g}) = \langle \mathbf{P}, \mathbf{C} \rangle - \varepsilon \mathbf{H}(\mathbf{P}) - \langle \mathbf{f}, \mathbf{P} \mathbf{1}_m - \mathbf{a} \rangle - \langle \mathbf{g}, \mathbf{P}^\top \mathbf{1}_n - \mathbf{b} \rangle.$$

Considering first order conditions, we have

$$\frac{\partial \mathcal{E}(\mathbf{P}, \mathbf{f}, \mathbf{g})}{\partial \mathbf{P}_{i,j}} = \mathbf{C}_{i,j} - \varepsilon \log(\mathbf{P}_{i,j}) - \mathbf{f}_i - \mathbf{g}_j.$$

which results, for an optimal \mathbf{P} coupling to the regularized problem, in the expression $\mathbf{P}_{i,j} = e^{\mathbf{f}_i/\varepsilon} e^{-\mathbf{C}_{i,j}/\varepsilon} e^{\mathbf{g}_j/\varepsilon}$ which can be rewritten in the form provided in the proposition using non-negative vectors \mathbf{u} and \mathbf{v} . \square

The factorization of the optimal solution exhibited in Equation (18.48) can be conveniently rewritten in matrix form as $\mathbf{P} = \text{diag}(\mathbf{u}) \mathbf{K} \text{diag}(\mathbf{v})$. \mathbf{u}, \mathbf{v} must therefore satisfy the following non-linear equations which correspond to the mass conservation constraints inherent to $\mathbf{U}(\mathbf{a}, \mathbf{b})$,

$$\text{diag}(\mathbf{u}) \mathbf{K} \text{diag}(\mathbf{v}) \mathbf{1}_m = \mathbf{a}, \quad \text{and} \quad \text{diag}(\mathbf{v}) \mathbf{K}^\top \text{diag}(\mathbf{u}) \mathbf{1}_n = \mathbf{b}, \quad (18.49)$$

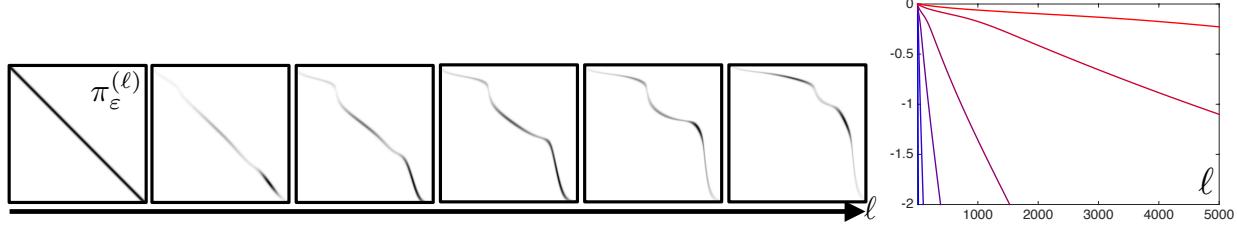


Figure 18.14: Left: evolution of the coupling $\pi_\varepsilon^{(\ell)} = \text{diag}(\mathbf{U}^{(\ell)})\mathbf{K}\text{diag}(\mathbf{V}^{(\ell)})$ computed at iteration ℓ of Sinkhorn's iterations, for 1-D densities. Right: impact of ε the convergence rate of Sinkhorn, as measured in term of marginal constraint violation $\log(\|\pi_\varepsilon^{(\ell)} \mathbf{1}_m - \mathbf{b}\|_1)$.

These two equations can be further simplified, since $\text{diag}(\mathbf{v})\mathbf{1}_m$ is simply \mathbf{v} , and the multiplication of $\text{diag}(\mathbf{u})$ times $\mathbf{K}\mathbf{v}$ is

$$\mathbf{u} \odot (\mathbf{K}\mathbf{v}) = \mathbf{a} \quad \text{and} \quad \mathbf{v} \odot (\mathbf{K}^T \mathbf{u}) = \mathbf{b} \quad (18.50)$$

where \odot corresponds to entry-wise multiplication of vectors. That problem is known in the numerical analysis community as the matrix scaling problem (see [?] and references therein). An intuitive way to try to solve these equations is to solve them iteratively, by modifying first \mathbf{u} so that it satisfies the left-hand side of Equation (18.50) and then \mathbf{v} to satisfy its right-hand side. These two updates define Sinkhorn's algorithm:

$$\mathbf{u}^{(\ell+1)} \stackrel{\text{def.}}{=} \frac{\mathbf{a}}{\mathbf{K}\mathbf{v}^{(\ell)}} \quad \text{and} \quad \mathbf{v}^{(\ell+1)} \stackrel{\text{def.}}{=} \frac{\mathbf{b}}{\mathbf{K}^T \mathbf{u}^{(\ell+1)}}, \quad (18.51)$$

initialized with an arbitrary positive vector $\mathbf{v}^{(0)} = \mathbf{1}_m$. The division operator used above between two vectors is to be understood entry-wise. Note that a different initialization will likely lead to a different solution for \mathbf{u}, \mathbf{v} , since \mathbf{u}, \mathbf{v} are only defined up to a multiplicative constant (if \mathbf{u}, \mathbf{v} satisfy (18.49) then so do $\lambda\mathbf{u}, \mathbf{v}/\lambda$ for any $\lambda > 0$). It turns out however that these iterations converge (see Remark 19 for a justification using iterative projections, and Remark 21 for a strict contraction result) and all result in the same optimal coupling $\text{diag}(\mathbf{u})\mathbf{K}\text{diag}(\mathbf{v})$. Figure 18.14, top row, shows the evolution of the coupling $\text{diag}(\mathbf{U}^{(\ell)})\mathbf{K}\text{diag}(\mathbf{V}^{(\ell)})$ computed by Sinkhorn iterations. It evolves from the Gibbs kernel \mathbf{K} towards the optimal coupling solving (18.39) by progressively shifting the mass away from the diagonal.

Remark 19 (Relation with iterative projections). Denoting

$$\mathcal{C}_{\mathbf{a}}^1 \stackrel{\text{def.}}{=} \{\mathbf{P} ; \mathbf{P}\mathbf{1}_m = \mathbf{a}\} \quad \text{and} \quad \mathcal{C}_{\mathbf{b}}^2 \stackrel{\text{def.}}{=} \left\{ \mathbf{P} ; \mathbf{P}^T \mathbf{1}_m = \mathbf{b} \right\}$$

the rows and columns constraints, one has $\mathbf{U}(\mathbf{a}, \mathbf{b}) = \mathcal{C}_{\mathbf{a}}^1 \cap \mathcal{C}_{\mathbf{b}}^2$. One can use Bregman iterative projections [?]

$$\mathbf{P}^{(\ell+1)} \stackrel{\text{def.}}{=} \text{Proj}_{\mathcal{C}_{\mathbf{a}}^1}^{\mathbf{KL}}(\mathbf{P}^{(\ell)}) \quad \text{and} \quad \mathbf{P}^{(\ell+2)} \stackrel{\text{def.}}{=} \text{Proj}_{\mathcal{C}_{\mathbf{b}}^2}^{\mathbf{KL}}(\mathbf{P}^{(\ell+1)}). \quad (18.52)$$

Since the sets $\mathcal{C}_{\mathbf{a}}^1$ and $\mathcal{C}_{\mathbf{b}}^2$ are affine, these iterations are known to converge to the solution of (18.44), see [?]. These iterate are equivalent to Sinkhorn iterations (18.51) since defining

$$\mathbf{P}^{(2\ell)} \stackrel{\text{def.}}{=} \text{diag}(\mathbf{u}^{(\ell)})\mathbf{K}\text{diag}(\mathbf{v}^{(\ell)}),$$

one has

$$\begin{aligned} \mathbf{P}^{(2\ell+1)} &\stackrel{\text{def.}}{=} \text{diag}(\mathbf{u}^{(\ell+1)})\mathbf{K}\text{diag}(\mathbf{v}^{(\ell)}) \\ \text{and } \mathbf{P}^{(2\ell+2)} &\stackrel{\text{def.}}{=} \text{diag}(\mathbf{u}^{(\ell+1)})\mathbf{K}\text{diag}(\mathbf{v}^{(\ell+1)}) \end{aligned}$$

In practice however one should prefer using (18.51) which only requires manipulating scaling vectors and multiplication against a Gibbs kernel, which can often be accelerated (see below Remarks ?? and ??).

Remark 20 (Hilbert metric). As initially explained by [?], the global convergence analysis of Sinkhorn is greatly simplified using Hilbert projective metric on $\mathbb{R}_{+,*}^n$ (positive vectors), defined as

$$\forall (\mathbf{u}, \mathbf{u}') \in (\mathbb{R}_{+,*}^n)^2, \quad d_{\mathcal{H}}(\mathbf{u}, \mathbf{u}') \stackrel{\text{def.}}{=} \log \max_{i,i'} \frac{\mathbf{u}_i \mathbf{u}'_{i'}}{\mathbf{u}'_{i'} \mathbf{u}_i}.$$

This can be shown to be a distance on the projective cone $\mathbb{R}_{+,*}^n / \sim$, where $\mathbf{u} \sim \mathbf{u}'$ means that $\exists s > 0, \mathbf{u} = s\mathbf{u}'$ (the vector are equal up to rescaling, hence the naming “projective”). This means that $d_{\mathcal{H}}$ satisfies the triangular inequality and $d_{\mathcal{H}}(\mathbf{u}, \mathbf{u}') = 0$ if and only if $\mathbf{u} \sim \mathbf{u}'$. This is a projective version of Hilbert’s original distance on bounded open convex sets [?]. The projective cone $\mathbb{R}_{+,*}^n / \sim$ is a complete metric space for this distance. It was introduced independently by [?] and [?] to provide a quantitative proof of Perron-Frobenius theorem, which, as explained in Remark ?? is linked to a local linearization of Sinkhorn’s iterates. They proved the following fundamental theorem, which shows that a positive matrix is a strict contraction on the cone of positive vectors.

Theorem 54. Let $\mathbf{K} \in \mathbb{R}_{+,*}^{n \times m}$, then for $(\mathbf{v}, \mathbf{v}') \in (\mathbb{R}_{+,*}^m)^2$

$$d_{\mathcal{H}}(\mathbf{K}\mathbf{v}, \mathbf{K}\mathbf{v}') \leq \lambda(\mathbf{K})d_{\mathcal{H}}(\mathbf{v}, \mathbf{v}') \text{ where } \begin{cases} \lambda(\mathbf{K}) \stackrel{\text{def.}}{=} \frac{\sqrt{\eta(\mathbf{K})}-1}{\sqrt{\eta(\mathbf{K})}+1} < 1 \\ \eta(\mathbf{K}) \stackrel{\text{def.}}{=} \max_{i,j,k,\ell} \frac{\mathbf{K}_{i,k}\mathbf{K}_{j,\ell}}{\mathbf{K}_{j,k}\mathbf{K}_{i,\ell}}. \end{cases}$$

Remark 21 (Global convergence). The following theorem, proved by [?], makes use of this Theorem 54 to show the linear convergence of Sinkhorn’s iterations.

Theorem 55. One has $(\mathbf{u}^{(\ell)}, \mathbf{v}^{(\ell)}) \rightarrow (\mathbf{u}^*, \mathbf{v}^*)$ and

$$d_{\mathcal{H}}(\mathbf{u}^{(\ell)}, \mathbf{u}^*) = O(\lambda(\mathbf{K})^{2\ell}), \quad d_{\mathcal{H}}(\mathbf{v}^{(\ell)}, \mathbf{v}^*) = O(\lambda(\mathbf{K})^{2\ell}). \quad (18.53)$$

One also has

$$\begin{aligned} d_{\mathcal{H}}(\mathbf{u}^{(\ell)}, \mathbf{u}^*) &\leq \frac{d_{\mathcal{H}}(\mathbf{P}^{(\ell)}\mathbf{1}_m, \mathbf{a})}{1 - \lambda(\mathbf{K})} \\ d_{\mathcal{H}}(\mathbf{v}^{(\ell)}, \mathbf{v}^*) &\leq \frac{d_{\mathcal{H}}(\mathbf{P}^{(\ell),\top}\mathbf{1}_n, \mathbf{b})}{1 - \lambda(\mathbf{K})} \end{aligned} \quad (18.54)$$

where we denoted $\mathbf{P}^{(\ell)} \stackrel{\text{def.}}{=} \text{diag}(\mathbf{u}^{(\ell)})\mathbf{K} \text{diag}(\mathbf{v}^{(\ell)})$. Lastly, one has

$$\|\log(\mathbf{P}^{(\ell)}) - \log(\mathbf{P}^*)\|_{\infty} \leq d_{\mathcal{H}}(\mathbf{u}^{(\ell)}, \mathbf{u}^*) + d_{\mathcal{H}}(\mathbf{v}^{(\ell)}, \mathbf{v}^*) \quad (18.55)$$

where \mathbf{P}^* is the unique solution of (18.39).

Proof. One notice that for any $(\mathbf{v}, \mathbf{v}') \in (\mathbb{R}_{+,*}^m)^2$, one has

$$d_{\mathcal{H}}(\mathbf{v}, \mathbf{v}') = d_{\mathcal{H}}(\mathbf{v}/\mathbf{v}', \mathbf{1}_m) = d_{\mathcal{H}}(\mathbf{1}_m/\mathbf{v}, \mathbf{1}_m/\mathbf{v}').$$

This shows that

$$\begin{aligned} d_{\mathcal{H}}(\mathbf{u}^{(\ell+1)}, \mathbf{u}^*) &= d_{\mathcal{H}}\left(\frac{\mathbf{a}}{\mathbf{K}\mathbf{v}^{(\ell)}}, \frac{\mathbf{a}}{\mathbf{K}\mathbf{v}^*}\right) \\ &= d_{\mathcal{H}}(\mathbf{K}\mathbf{v}^{(\ell)}, \mathbf{K}\mathbf{v}^*) \leq \lambda(\mathbf{K})d_{\mathcal{H}}(\mathbf{v}^{(\ell)}, \mathbf{v}^*). \end{aligned}$$

where we used Theorem 54. This shows (18.53). One also has, using the triangular inequality

$$\begin{aligned} d_{\mathcal{H}}(\mathbf{u}^{(\ell)}, \mathbf{u}^*) &\leq d_{\mathcal{H}}(\mathbf{u}^{(\ell+1)}, \mathbf{u}^{(\ell)}) + d_{\mathcal{H}}(\mathbf{u}^{(\ell+1)}, \mathbf{u}^*) \\ &\leq d_{\mathcal{H}}\left(\frac{\mathbf{a}}{\mathbf{K}\mathbf{v}^{(\ell)}}, \mathbf{u}^{(\ell)}\right) + \lambda(\mathbf{K})d_{\mathcal{H}}(\mathbf{u}^{(\ell)}, \mathbf{u}^*) \\ &= d_{\mathcal{H}}\left(\mathbf{a}, \mathbf{u}^{(\ell)} \odot (\mathbf{K}\mathbf{v}^{(\ell)})\right) + \lambda(\mathbf{K})d_{\mathcal{H}}(\mathbf{u}^{(\ell)}, \mathbf{u}^*), \end{aligned}$$

which gives the first part of (18.54) since $\mathbf{u}^{(\ell)} \odot (\mathbf{K}\mathbf{v}^{(\ell)}) = \mathbf{P}^{(\ell)}\mathbf{1}_m$ (the second one being similar). The proof of (18.55) follows from [?, Lemma 3] \square

The bound (18.54) shows that some error measures on the marginal constraints violation, for instance $\|\mathbf{P}^{(\ell)}\mathbf{1}_m - \mathbf{a}\|_1$ and $\|\mathbf{P}^{(\ell)^\top}\mathbf{1}_n - \mathbf{b}\|_1$, are useful stopping criteria to monitor the convergence.

Figure 18.14, bottom row, highlights this linear rate on the constraint violation, and shows how this rate degrades as $\varepsilon \rightarrow 0$. These results are proved in [?] and are tightly connected to nonlinear Perron-Frobenius Theory [?]. Perron-Frobenius theory corresponds to the linearization of the iterations, see (??). This convergence analysis is extended in [?], who shows that each iteration of Sinkhorn increases the permanent of the scaled coupling matrix.

Regularized Dual and Log-domain Computations The following proposition details the dual problem associated to (18.39).

Proposition 61. *One has*

$$L_C^\varepsilon(\mathbf{a}, \mathbf{b}) = \max_{\mathbf{f} \in \mathbb{R}^n, \mathbf{g} \in \mathbb{R}^m} \langle \mathbf{f}, \mathbf{a} \rangle + \langle \mathbf{g}, \mathbf{b} \rangle - \varepsilon \langle e^{\mathbf{f}/\varepsilon}, \mathbf{K} e^{\mathbf{g}/\varepsilon} \rangle. \quad (18.56)$$

The optimal (\mathbf{f}, \mathbf{g}) are linked to scalings (\mathbf{u}, \mathbf{v}) appearing in (18.48) through

$$(\mathbf{u}, \mathbf{v}) = (e^{\mathbf{f}/\varepsilon}, e^{\mathbf{g}/\varepsilon}). \quad (18.57)$$

Proof. We start from the end of the proof of Proposition 60, which links the optimal primal solution \mathbf{P} and dual multipliers \mathbf{f} and \mathbf{g} for the marginal constraints as $\mathbf{P}_{i,j} = e^{\mathbf{f}_i/\varepsilon} e^{-\mathbf{C}_{i,j}/\varepsilon} e^{\mathbf{g}_j/\varepsilon}$. Substituting in the Lagrangian $\mathcal{E}(\mathbf{P}, \mathbf{f}, \mathbf{g})$ of Equation (18.5) the optimal \mathbf{P} as a function of \mathbf{f} and \mathbf{g} , we obtain that the Lagrange dual function equals

$$\mathbf{f}, \mathbf{g} \mapsto \langle e^{\mathbf{f}/\varepsilon}, (\mathbf{K} \odot \mathbf{C}) e^{\mathbf{g}/\varepsilon} \rangle - \varepsilon \mathbf{H}(\text{diag}(e^{\mathbf{f}/\varepsilon}) \mathbf{K} \text{diag}(e^{\mathbf{g}/\varepsilon})). \quad (18.58)$$

The entropy of \mathbf{P} scaled by ε , namely $\varepsilon \langle \mathbf{P}, \log \mathbf{P} - \mathbf{1}_{n \times m} \rangle$ can be stated explicitly as a function of $\mathbf{f}, \mathbf{g}, \mathbf{C}$

$$\begin{aligned} & \langle \text{diag}(e^{\mathbf{f}/\varepsilon}) \mathbf{K} \text{diag}(e^{\mathbf{g}/\varepsilon}), \mathbf{f} \mathbf{1}_m^\top + \mathbf{1}_n \mathbf{g}^\top - \mathbf{C} - \varepsilon \mathbf{1}_{n \times m} \rangle \\ &= -\langle e^{\mathbf{f}/\varepsilon}, (\mathbf{K} \odot \mathbf{C}) e^{\mathbf{g}/\varepsilon} \rangle + \langle \mathbf{f}, \mathbf{a} \rangle + \langle \mathbf{g}, \mathbf{b} \rangle - \varepsilon \langle e^{\mathbf{f}/\varepsilon}, \mathbf{K} e^{\mathbf{g}/\varepsilon} \rangle \end{aligned}$$

therefore, the first term in (18.58) cancels out with the first term in the entropy above. The remaining terms are those displayed in (18.56). \square

Remark 22. Dual for generic measures For generic (non-necessarily discrete) input measures (α, β) , the dual problem (18.56) reads

$$\sup_{f, g \in \mathcal{C}(\mathcal{X}) \times \mathcal{C}(\mathcal{Y})} \int_{\mathcal{X}} f(x) d\alpha(x) + \int_{\mathcal{Y}} g(y) d\beta(y) - \varepsilon \int_{\mathcal{X} \times \mathcal{Y}} e^{\frac{-c(x,y)+f(x)+g(y)}{\varepsilon}} d\alpha(x) d\beta(y)$$

This corresponds to a smoothing of the constraint $\mathcal{R}(c)$ appearing in the original problem (18.21), which is retrieved in the limit $\varepsilon \rightarrow 0$. Proving existence (*i.e.* the sup is actually a max) of these Kantorovich potentials (f, g) in the case of entropic transport is less easy than for classical OT (because one cannot use c -transform and potentials are not automatically Lipschitz). Proof of existence can be done using the convergence of Sinkhorn iterations, see [?] for more details.

Remark 23 (Sinkhorn as a Block Coordinate Ascent on the Dual Problem). A simple approach to solve the unconstrained maximization problem (18.56) is to use an exact *block coordinate ascent* strategy, namely to update alternatively \mathbf{f} and \mathbf{g} to cancel their gradients with respect to the objective of (18.56). Indeed, one can easily notice that, writing $Q(\mathbf{f}, \mathbf{g})$ for the objective of (18.56) that

$$\nabla|_{\mathbf{f}} Q(\mathbf{f}, \mathbf{g}) = \mathbf{a} - e^{\mathbf{f}/\varepsilon} \odot (\mathbf{K} e^{\mathbf{g}/\varepsilon}), \quad (18.59)$$

$$\nabla|_{\mathbf{g}} Q(\mathbf{f}, \mathbf{g}) = \mathbf{b} - e^{\mathbf{g}/\varepsilon} \odot (\mathbf{K}^\top e^{\mathbf{f}/\varepsilon}). \quad (18.60)$$

Block coordinate ascent can therefore be implemented in a closed form by applying successively the following updates, starting from any arbitrary $\mathbf{g}^{(0)}$, for $l \geq 0$,

$$\mathbf{f}^{(\ell+1)} = \varepsilon \log \mathbf{a} - \varepsilon \log \left(\mathbf{K} e^{\mathbf{g}^{(\ell)}/\varepsilon} \right), \quad (18.61)$$

$$\mathbf{g}^{(\ell+1)} = \varepsilon \log \mathbf{b} - \varepsilon \log \left(\mathbf{K}^T e^{\mathbf{f}^{(\ell+1)}/\varepsilon} \right). \quad (18.62)$$

Such iterations are mathematically equivalent to the Sinkhorn iterations (18.51) when considering the primal-dual relations highlighted in (18.57). Indeed, we recover that at any iteration

$$(\mathbf{f}^{(\ell)}, \mathbf{g}^{(\ell)}) = \varepsilon (\log(\mathbf{u}^{(\ell)}), \log(\mathbf{v}^{(\ell)})).$$

Remark 24 (Soft-min rewriting). Iterations (18.61) and (18.62) can be given an alternative interpretation, using the following notation. Given a vector \mathbf{z} of real numbers we write $\min_\varepsilon \mathbf{z}$ for the *soft-minimum* of its coordinates, namely

$$\min_\varepsilon \mathbf{z} = -\varepsilon \log \sum_i e^{-\mathbf{z}_i/\varepsilon}.$$

Note that $\min_\varepsilon(\mathbf{z})$ converges to $\min \mathbf{z}$ for any vector \mathbf{z} as $\varepsilon \rightarrow 0$. Indeed, \min_ε can be interpreted as a differentiable approximation of the \min function. Using these notations, Equations (18.61) and (18.62) can be rewritten

$$(\mathbf{f}^{(\ell+1)})_i = \min_\varepsilon (\mathbf{C}_{ij} - \mathbf{g}_j^{(\ell)})_j + \varepsilon \log \mathbf{a}_i, \quad (18.63)$$

$$(\mathbf{g}^{(\ell+1)})_j = \min_\varepsilon (\mathbf{C}_{ij} - \mathbf{f}_i^{(\ell)})_i + \varepsilon \log \mathbf{b}_j. \quad (18.64)$$

Here the term $\min_\varepsilon (\mathbf{C}_{ij} - \mathbf{g}_j^{(\ell)})_j$ denotes the soft-minimum of all values of the j -th column of matrix $(\mathbf{C} - \mathbf{1}_n(\mathbf{g}^{(\ell)})^\top)$. To simplify notations, we introduce an operator that takes a matrix as input and outputs now a column vector of the soft-minimum values of its columns or rows. Namely, for any matrix $A \in \mathbb{R}^{n \times m}$, we define

$$\begin{aligned} \text{Min}_\varepsilon^{\text{row}}(\mathbf{A}) &\stackrel{\text{def.}}{=} \left(\min_\varepsilon (\mathbf{A}_{i,j})_j \right)_i \in \mathbb{R}^n, \\ \text{Min}_\varepsilon^{\text{col}}(\mathbf{A}) &\stackrel{\text{def.}}{=} \left(\min_\varepsilon (\mathbf{A}_{i,j})_i \right)_j \in \mathbb{R}^m. \end{aligned}$$

Note that these operations are equivalent to the entropic c -transform introduced in §?? (see in particular ??). Using these notations, Sinkhorn's iterates read

$$\mathbf{f}^{(\ell+1)} = \text{Min}_\varepsilon^{\text{row}}(\mathbf{C} - \mathbf{1}_n \mathbf{g}^{(\ell)^\top}) + \varepsilon \log \mathbf{a}, \quad (18.65)$$

$$\mathbf{g}^{(\ell+1)} = \text{Min}_\varepsilon^{\text{col}}(\mathbf{C} - \mathbf{f}^{(\ell)} \mathbf{1}_m^\top) + \varepsilon \log \mathbf{b}. \quad (18.66)$$

Note that as $\varepsilon \rightarrow 0$, \min_ε converges to \min , but the iterations do not converge anymore in the limit $\varepsilon = 0$, because alternate minimization does not converge for constrained problems (which is the case for the unregularized dual (18.17)).

Remark 25 (Log-domain Sinkhorn). While mathematically equivalent to the Sinkhorn updates (18.51), iterations (18.63) and (18.64) suggest to use the *log-sum-exp* stabilization trick to avoid underflow for small values of ε . Writing $\underline{z} = \min \mathbf{z}$, that trick suggests to evaluate $\min_\varepsilon \mathbf{z}$ as

$$\min_\varepsilon \mathbf{z} = \underline{z} - \varepsilon \log \sum_i e^{-(\mathbf{z}_i - \underline{z})/\varepsilon}. \quad (18.67)$$

Instead of subtracting \underline{z} to stabilize the log domain iterations as in (18.67), one can actually subtract the previously computed scalings. This leads to the following stabilized iteration

$$\mathbf{f}^{(\ell+1)} = \text{Min}_\varepsilon^{\text{row}}(\mathbf{S}(\mathbf{f}^{(\ell)}, \mathbf{g}^{(\ell)})) - \mathbf{f}^{(\ell)} + \varepsilon \log(\mathbf{a}) \quad (18.68)$$

$$\mathbf{g}^{(\ell+1)} = \text{Min}_\varepsilon^{\text{col}}(\mathbf{S}(\mathbf{f}^{(\ell+1)}, \mathbf{g}^{(\ell)})) - \mathbf{g}^{(\ell)} + \varepsilon \log(\mathbf{b}), \quad (18.69)$$

where we defined

$$\mathbf{S}(\mathbf{f}, \mathbf{g}) = (\mathbf{C}_{i,j} - \mathbf{f}_i - \mathbf{g}_j)_{i,j}.$$

In contrast to the original iterations (18.51), these log-domain iterations (18.68) and (18.69) are stable for arbitrary $\varepsilon > 0$, because the quantity $\mathbf{S}(\mathbf{f}, \mathbf{g})$ stays bounded during the iterations. The downside is that it requires nm computations of \exp at each step. Computing a $\text{Min}_\varepsilon^{\text{row}}$ or $\text{Min}_\varepsilon^{\text{col}}$ is typically substantially slower than matrix multiplications, and requires computing line by line soft-minima of matrices \mathbf{S} . There is therefore no efficient way to parallelize the application of Sinkhorn maps for several marginals simultaneously. In Euclidean domain of small dimension, it is possible to develop efficient multiscale solvers with a decaying ε strategy to significantly speed up the computation using sparse grids [?].

18.6 Extensions

Wasserstein Barycenters. Given input histogram $\{\mathbf{b}_s\}_{s=1}^S$, where $b_s \in \Sigma_{n_s}$, and weights $\lambda \in \Sigma_S$, a Wasserstein barycenter is computed by minimizing

$$\min_{\mathbf{a} \in \Sigma_n} \sum_{s=1}^S \lambda_s \mathbf{L}_{\mathbf{C}_s}(\mathbf{a}, \mathbf{b}_s) \quad (18.70)$$

where the cost matrices $\mathbf{C}_s \in \mathbb{R}^{n \times n_s}$ need to be specified. A typical setup is “Eulerian”, so that all the barycenters are defined on the same grid, $n_s = n$, $\mathbf{C}_s = \mathbf{C} = \mathbf{D}^p$ is set to be a distance matrix, so that one solves

$$\min_{\mathbf{a} \in \Sigma_n} \sum_{s=1}^S \lambda_s W_p^p(\mathbf{a}, \mathbf{b}_s).$$

This barycenter problem (18.70) was originally introduced by [?] following earlier ideas of [?]. They proved in particular uniqueness of the barycenter for $c(x, y) = \|x - y\|^2$ over $\mathcal{X} = \mathbb{R}^d$, if one of the input measure has a density with respect to the Lebesgue measure (and more generally under the same hypothesis as the one guaranteeing the existence of a Monge map, see Remark ??).

The barycenter problem for histograms (18.70) is in fact a linear program, since one can look for the S couplings $(\mathbf{P}_s)_s$ between each input and the barycenter itself

$$\min_{\mathbf{a} \in \Sigma_n, (\mathbf{P}_s \in \mathbb{R}^{n \times n_s})_s} \left\{ \sum_{s=1}^S \lambda_s \langle \mathbf{P}_s, \mathbf{C}_s \rangle ; \forall s, \mathbf{P}_s^\top \mathbf{1}_{n_s} = \mathbf{a}, \mathbf{P}_s^\top \mathbf{1}_n = \mathbf{b}_s \right\}.$$

Although this problem is an LP, its scale forbids the use generic solvers for medium scale problems. One can therefore resort to using first order methods such as subgradient descent on the dual [?].

Remark 26. Barycenter of arbitrary measures Given a set of input measure $(\beta_s)_s$ defined on some space \mathcal{X} , the barycenter problem becomes

$$\min_{\alpha \in \mathcal{M}_+^1(\mathcal{X})} \sum_{s=1}^S \lambda_s \mathcal{L}_c(\alpha, \beta_s). \quad (18.71)$$

In the case where $\mathcal{X} = \mathbb{R}^d$ and $c(x, y) = \|x - y\|^2$, [?] shows that if one of the input measures has a density, then this barycenter is unique. Problem (18.71) can be viewed as a generalization of the problem of computing barycenters of points $(x_s)_{s=1}^S \in \mathcal{X}^S$ to arbitrary measures. Indeed, if $\beta_s = \delta_{x_s}$ is a single Dirac mass, then a solution to (18.71) is δ_{x^*} where x^* is a Fréchet mean solving (??). Note that for $c(x, y) = \|x - y\|^2$, the mean of the barycenter α^* is necessarily the barycenter of the mean, *i.e.*

$$\int_{\mathcal{X}} x d\alpha^*(x) = \sum_s \lambda_s \int_{\mathcal{X}} x d\alpha_s(x),$$

and the support of α^* is located in the convex hull of the supports of the $(\alpha_s)_s$. The consistency of the approximation of the infinite dimensional optimization (18.71) when approximating the input distribution using discrete ones (and thus solving (18.70) in place) is studied in [?]. Let us also note that it is possible to re-cast (18.71) as a multi-marginal OT problem, see Remark ??.

One can use entropic smoothing and approximate the solution of (18.70) using

$$\min_{\mathbf{a} \in \Sigma_n} \sum_{s=1}^S \lambda_s L_{\mathbf{C}_s}^\varepsilon(\mathbf{a}, \mathbf{b}_s) \quad (18.72)$$

for some $\varepsilon > 0$. This is a smooth convex minimization problem, which can be tackled using gradient descent [?]. An alternative is to use descent method (typically quasi-Newton) on the semi-dual [?], which is useful to integrate additional regularizations on the barycenter (e.g. to impose some smoothness). A simple but effective approach, as remarked in [?] is to rewrite (18.72) as a (weighted) KL projection problem

$$\min_{(\mathbf{P}_s)_s} \left\{ \sum_s \lambda_s \mathbf{KL}(\mathbf{P}_s | \mathbf{K}_s) ; \forall s, \mathbf{P}_s^T \mathbf{1}_m = \mathbf{b}_s, \mathbf{P}_1 \mathbf{1}_1 = \dots = \mathbf{P}_S \mathbf{1}_S \right\} \quad (18.73)$$

where we denoted $\mathbf{K}_s \stackrel{\text{def.}}{=} e^{-\mathbf{C}_s/\varepsilon}$. Here, the barycenter \mathbf{a} is implicitly encoded in the row marginals of all the couplings $\mathbf{P}_s \in \mathbb{R}^{n \times n_s}$ as $\mathbf{a} = \mathbf{P}_1 \mathbf{1}_1 = \dots = \mathbf{P}_S \mathbf{1}_S$. As detailed in [?], one can generalize Sinkhorn to this problem, which also corresponds to iterative projection. This can also be seen as a special case of the generalized Sinkhorn detailed in §???. The optimal couplings $(\mathbf{P}_s)_s$ solving (18.73) are computed in scaling form as

$$\mathbf{P}_s = \text{diag}(\mathbf{u}_s) \mathbf{K} \text{diag}(\mathbf{v}_s), \quad (18.74)$$

and the scalings are sequentially updated as

$$\forall s \in \llbracket 1, S \rrbracket, \quad \mathbf{v}_s^{(\ell+1)} \stackrel{\text{def.}}{=} \frac{\mathbf{b}_s}{\mathbf{K}_s^T \mathbf{u}_s^{(\ell)}}, \quad (18.75)$$

$$\forall s \in \llbracket 1, S \rrbracket, \quad \mathbf{u}_s^{(\ell+1)} \stackrel{\text{def.}}{=} \frac{\mathbf{a}^{(\ell+1)}}{\mathbf{K}_s \mathbf{v}_s^{(\ell+1)}}, \quad (18.76)$$

$$\text{where } \mathbf{a}^{(\ell+1)} \stackrel{\text{def.}}{=} \prod_s (\mathbf{K}_s \mathbf{v}_s^{(\ell+1)})^{\lambda_s}. \quad (18.77)$$

An alternative way to derive these iterations is to perform alternate minimization on the variables of a dual problem, which detailed in the following proposition.

Proposition 62. *The optimal $(\mathbf{u}_s, \mathbf{v}_s)$ appearing in (18.74) can be written as $(\mathbf{u}_s, \mathbf{v}_s) = (e^{\mathbf{f}_s/\varepsilon}, e^{\mathbf{g}_s/\varepsilon})$ where $(\mathbf{f}_s, \mathbf{g}_s)_s$ are the solutions of the following program (whose value matches the one of (18.72))*

$$\max_{(\mathbf{f}_s, \mathbf{g}_s)_s} \left\{ \sum_s \lambda_s \left(\langle \mathbf{g}_s, \mathbf{b}_s \rangle - \varepsilon \langle \mathbf{K}_s e^{\mathbf{g}_s/\varepsilon}, e^{\mathbf{f}_s/\varepsilon} \rangle \right) ; \sum_s \lambda_s \mathbf{f}_s = 0 \right\}. \quad (18.78)$$

Proof. Introducing Lagrange multipliers in (18.73) leads to

$$\begin{aligned} \min_{(\mathbf{P}_s)_s, \mathbf{a}} \max_{(\mathbf{f}_s, \mathbf{g}_s)_s} & \sum_s \lambda_s \left(\varepsilon \mathbf{KL}(\mathbf{P}_s | \mathbf{K}_s) + \langle \mathbf{a} - \mathbf{P}_s \mathbf{1}_m, \mathbf{f}_s \rangle \right. \\ & \left. + \langle \mathbf{b}_s - \mathbf{P}_s^T \mathbf{1}_m, \mathbf{g}_s \rangle \right). \end{aligned}$$

Strong duality holds, so that one can exchange the min and the max, and gets

$$\begin{aligned} \max_{(\mathbf{f}_s, \mathbf{g}_s)_s} & \sum_s \lambda_s \left(\langle \mathbf{g}_s, \mathbf{b}_s \rangle + \min_{\mathbf{P}_s} \varepsilon \mathbf{KL}(\mathbf{P}_s | \mathbf{K}_s) - \langle \mathbf{P}_s, \mathbf{f}_s \oplus \mathbf{g}_s \rangle \right) \\ & + \min_{\mathbf{a}} \langle \sum_s \lambda_s \mathbf{f}_s, \mathbf{a} \rangle. \end{aligned}$$

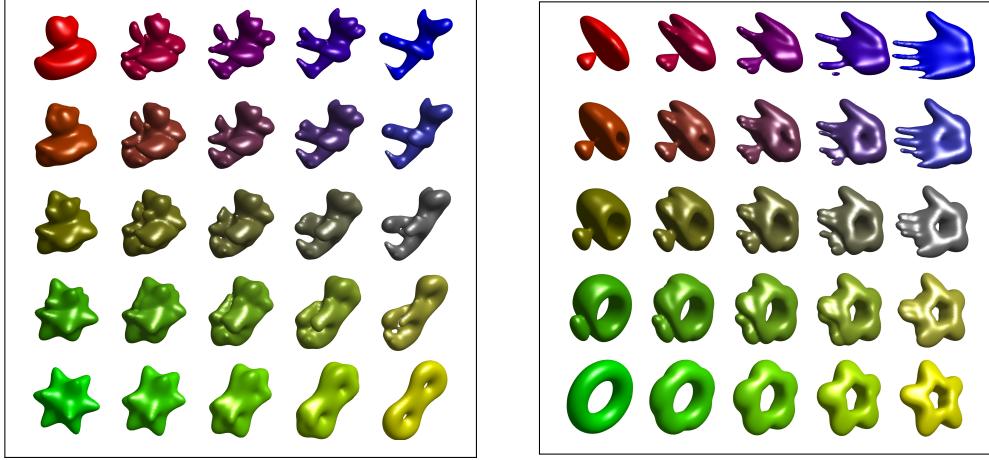


Figure 18.15: Barycenters between 4 input 3-D shapes using entropic regularization (18.72). The weights $(\lambda_s)_s$ are bilinear with respect to the four corners of the square. Shapes are represented as measures that are uniform within the boundaries of the shape and null outside.

The explicit minimization on \mathbf{a} gives the constraint $\sum_s \lambda_s \mathbf{f}_s = 0$ together with

$$\max_{(\mathbf{f}_s, \mathbf{g}_s)_s} \sum_s \lambda_s \langle \mathbf{g}_s, \mathbf{b}_s \rangle - \varepsilon \mathbf{KL}^* \left(\frac{\mathbf{f}_s \oplus \mathbf{g}_s}{\varepsilon} \mid \mathbf{K}_s \right)$$

where $\mathbf{KL}^*(\cdot \mid \mathbf{K}_s)$ is the Legendre transform (??) of the function $\mathbf{KL}^*(\cdot \mid \mathbf{K}_s)$. This Legendre transform reads

$$\mathbf{KL}^*(\mathbf{U} \mid \mathbf{K}) = \sum_{i,j} \mathbf{K}_{i,j} (e^{\mathbf{U}_{i,j}} - 1), \quad (18.79)$$

which shows the desired formula. To show (18.79), since this function is separable, one needs to compute

$$\forall (u, k) \in \mathbb{R}_+^2, \quad \mathbf{KL}^*(u \mid k) \stackrel{\text{def.}}{=} \max_r ur - (r \log(r/k) - r + k)$$

whose optimality condition reads $u = \log(r/k)$, i.e. $r = ke^u$, hence the result. \square

Minimizing (18.78) with respect to each \mathbf{g}_s , while keeping all the other variable fixed, is obtained in closed form by (18.75). Minimizing (18.78) with respect to all the $(\mathbf{f}_s)_s$ requires to solve for \mathbf{a} using (18.77) and leads to the expression (18.76).

Figures ?? and ?? show applications to 2-D and 3-D shapes interpolation. Figure ?? shows a computation of barycenters on a surface, where the ground cost is the square of the geodesic distance. For this figure, the computations are performed using the geodesic in heat approximation detailed in Remark ???. We refer to [?] for more details and other applications to computer graphics and imaging sciences.

Wasserstein Loss. In statistics, text processing or imaging, one must usually compare a probability distribution β arising from measurements to a model, namely a parameterized family of distributions $\{\alpha_\theta, \theta \in \Theta\}$ where Θ is a subset of an Euclidean space. Such a comparison is done through a “loss” or a “fidelity” term, which, in this section, is the Wasserstein distance. In the simplest scenario, the computation of a suitable parameter θ is obtained by minimizing directly

$$\min_{\theta \in \Theta} \mathcal{E}(\theta) \stackrel{\text{def.}}{=} \mathcal{L}_c(\alpha_\theta, \beta). \quad (18.80)$$

Of course, one can consider more complicated problems: for instance, the barycenter problem described in §?? consists in a sum of such terms. However, most of these more advanced problems can be usually

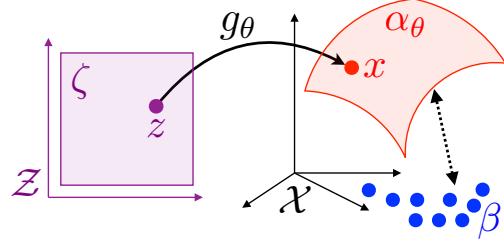


Figure 18.16: Schematic display of the density fitting problem 18.81.

solved by adapting tools defined for basic case: either using the chain rule to compute explicitly derivatives, or using automatic differentiation.

The Wasserstein distance between two histograms or two densities is convex with respect to these inputs, as shown by (18.17) and (18.21) respectively. Therefore, when the parameter θ is itself a histogram, namely $\Theta = \Sigma_n$ and $\alpha_\theta = \theta$, or more generally when θ describes K weights in the simplex, $\Theta = \Sigma_K$, and $\alpha_\theta = \sum_{i=1}^K \theta_i \alpha_i$ is a convex combination of known atoms $\alpha_1, \dots, \alpha_K$ in Σ_N , Problem (18.80) remains convex (the first case corresponds to the barycenter problem, the second to one iteration of the dictionary learning problem with a Wasserstein loss [?]). However, for more general parameterizations $\theta \mapsto \alpha_\theta$, Problem (18.80) is in general not convex.

A practical problem of paramount importance in statistic and machine learning is density fitting. Given some discrete samples $(x_i)_{i=1}^n \subset \mathcal{X}$ from some unknown distribution, the goal is to fit a parametric model $\theta \mapsto \alpha_\theta \in \mathcal{M}(\mathcal{X})$ to the observed empirical input measure β

$$\min_{\theta \in \Theta} \mathcal{L}(\alpha_\theta, \beta) \quad \text{where} \quad \beta = \frac{1}{n} \sum_i \delta_{x_i}, \quad (18.81)$$

where \mathcal{L} is some “loss” function between a discrete and a “continuous” (arbitrary) distribution (see Figure 18.16).

In the case where α_θ as a density $\rho_\theta \stackrel{\text{def.}}{=} \rho_{\alpha_\theta}$ with respect to the Lebesgue measure (or any other fixed reference measure), the maximum likelihood estimator (MLE) is obtained by solving

$$\min_{\theta} \mathcal{L}_{\text{MLE}}(\alpha_\theta, \beta) \stackrel{\text{def.}}{=} - \sum_i \log(\rho_\theta(x_i)).$$

This corresponds to using an empirical counterpart of a Kullback-Leibler loss since, assuming the x_i are i.i.d. samples of some $\bar{\beta}$, then

$$\mathcal{L}_{\text{MLE}}(\alpha, \beta) \xrightarrow{n \rightarrow +\infty} \text{KL}(\alpha | \bar{\beta})$$

This MLE approach is known to lead to optimal estimation procedures in many cases (see for instance [?]). However, it fails to work when estimating singular distributions, typically when the α_θ does not have a density (so that $\mathcal{L}_{\text{MLE}}(\alpha_\theta, \beta) = +\infty$) or when $(x_i)_i$ are samples from some singular $\bar{\beta}$ (so that the α_θ should share the same support as β for $\text{KL}(\alpha | \bar{\beta})$ to be finite, but this support is usually unknown). Another issue is that in several cases of practical interest, the density ρ_θ is inaccessible (or too hard to compute).

A typical setup where both problems (singular and unknown densities) occur is for so-called generative models, where the parametric measure is written as a push-forward of a fixed reference measure $\zeta \in \mathcal{M}(\mathcal{Z})$

$$\alpha_\theta = h_{\theta, \sharp} \zeta \quad \text{where} \quad h_\theta : \mathcal{Z} \rightarrow \mathcal{X}$$

where the push-forward operator is introduced in Definition 23. The space \mathcal{Z} is usually low-dimensional, so that the support of α_θ is localized along a low-dimensional “manifold” and the resulting density is highly singular (it does not have a density with respect to Lebesgue measure). Furthermore, computing this density is usually intractable, while generating i.i.d. samples from α_θ is achieved by computing $x_i = h_\theta(z_i)$ where $(z_i)_i$ are i.i.d. samples from ζ .

In order to cope with such difficult scenario, one has to use weak metrics in place of the MLE functional \mathcal{L}_{MLE} , which needs to be written in dual form as

$$\mathcal{L}(\alpha, \beta) \stackrel{\text{def.}}{=} \max_{(f,g) \in \mathcal{C}(\mathcal{X})^2} \left\{ \int_{\mathcal{X}} f(x) d\alpha(x) + \int_{\mathcal{X}} g(x) d\beta(x) ; (f,g) \in \mathcal{R} \right\}. \quad (18.82)$$

Dual norms exposed in §?? correspond to imposing $\mathcal{R} = \{(f, -f) ; f \in B\}$, while optimal transport (18.21) sets $\mathcal{R} = \mathcal{R}(c)$ as defined in (18.22).

For a fixed θ , evaluating the energy to be minimized in (18.81) using such a loss function corresponds to solving a semi-discrete optimal transport, which is the focus of Chapter ???. Minimizing the energy with respect to θ is much more involved, and is typically highly non-convex.

The class of estimators obtained using $\mathcal{L} = \mathcal{L}_c$, often called “Minimum Kantorovitch Estimators” (MKE), was initially introduced in [?], see also [?].

Gromov-Wasserstein. Optimal transport needs a ground cost \mathbf{C} to compare histograms (\mathbf{a}, \mathbf{b}) , it can thus not be used if the histograms are not defined on the same underlying space, or if one cannot pre-register these spaces to define a ground cost. To address this issue, one can instead only assume a weaker assumption, namely that one has at its disposal two matrices $\mathbf{D} \in \mathbb{R}^{n \times n}$ and $\mathbf{D}' \in \mathbb{R}^{m \times m}$ that represent some relationship between the points on which the histograms are defined. A typical scenario is when these matrices are (power of) distance matrices. The Gromov-Wasserstein problem reads

$$\text{GW}((\mathbf{a}, \mathbf{D}), (\mathbf{b}, \mathbf{D}'))^2 \stackrel{\text{def.}}{=} \min_{\mathbf{P} \in \mathbf{U}(\mathbf{a}, \mathbf{b})} \mathcal{E}_{\mathbf{D}, \mathbf{D}'}(\mathbf{P}) \stackrel{\text{def.}}{=} \sum_{i,j,i',j'} |\mathbf{D}_{i,i'} - \mathbf{D}'_{j,j'}|^2 \mathbf{P}_{i,j} \mathbf{P}_{i',j'}. \quad (18.83)$$

This is a non-convex problem, which can be recast as a Quadratic Assignment Problem (QAP) [?] and is in full generality NP-hard to solve for arbitrary inputs. It is in fact equivalent to a graph matching problem [?] for a particular cost.

One can show that GW satisfies the triangular inequality, and in fact it defines a distance between metric spaces equipped with a probability distribution (here assumed to be discrete in definition (18.83)) up to isometries preserving the measures. This distance was introduced and studied in details by Memoli in [?]. An in-depth mathematical exposition (in particular, its geodesic structure and gradient flows) is given in [?]. See also [?] for applications in computer vision. This distance is also tightly connected with the Gromov-Hausdorff distance [?] between metric spaces, which have been used for shape matching [?, ?].

Remark 27. Gromov-Wasserstein distance The general setting corresponds to computing couplings between metric measure spaces $(\mathcal{X}, d_{\mathcal{X}}, \alpha_{\mathcal{X}})$ and $(\mathcal{Y}, d_{\mathcal{Y}}, \alpha_{\mathcal{Y}})$ where $(d_{\mathcal{X}}, d_{\mathcal{Y}})$ are distances and $(\alpha_{\mathcal{X}}, \alpha_{\mathcal{Y}})$ are measures on their respective spaces. One defines

$$\mathcal{GW}((\alpha_{\mathcal{X}}, d_{\mathcal{X}}), (\alpha_{\mathcal{Y}}, d_{\mathcal{Y}}))^2 \stackrel{\text{def.}}{=} \min_{\pi \in \mathbf{U}(\alpha_{\mathcal{X}}, \alpha_{\mathcal{Y}})} \int_{\mathcal{X}^2 \times \mathcal{Y}^2} |d_{\mathcal{X}}(x, x') - d_{\mathcal{Y}}(y, y')|^2 d\pi(x, y) d\pi(x', y'). \quad (18.84)$$

\mathcal{GW} defines a distance between metric measure spaces up to isometries, where one says that $(\alpha_{\mathcal{X}}, d_{\mathcal{X}})$ and $(\alpha_{\mathcal{Y}}, d_{\mathcal{Y}})$ are isometric if there exists $\varphi : \mathcal{X} \rightarrow \mathcal{Y}$ such that $\varphi_{\sharp} \alpha_{\mathcal{X}} = \alpha_{\mathcal{Y}}$ and $d_{\mathcal{Y}}(\varphi(x), \varphi(x')) = d_{\mathcal{X}}(x, x')$.

Remark 28. Gromov-Wasserstein geodesics The space of metric spaces (up to isometries) endowed with this \mathcal{GW} distance (18.84) has a geodesic structure. [?] shows that the geodesic between $(\mathcal{X}_0, d_{\mathcal{X}_0}, \alpha_0)$ and $(\mathcal{X}_1, d_{\mathcal{X}_1}, \alpha_1)$ can be chosen to be $t \in [0, 1] \mapsto (\mathcal{X}_0 \times \mathcal{X}_1, d_t, \pi^*)$ where π^* is a solution of (18.84) and for all $((x_0, x_1), (x'_0, x'_1)) \in (\mathcal{X}_0 \times \mathcal{X}_1)^2$,

$$d_t((x_0, x_1), (x'_0, x'_1)) \stackrel{\text{def.}}{=} (1-t)d_{\mathcal{X}_0}(x_0, x'_0) + td_{\mathcal{X}_1}(x_1, x'_1).$$

This formula allows one to define and analyze gradient flows which minimize functionals involving metric spaces, see [?]. It is however difficult to handle numerically, because it involves computations over the product space $\mathcal{X}_0 \times \mathcal{X}_1$. A heuristic approach is used in [?] to define geodesics and barycenters of metric measure spaces while imposing the cardinality of the involved spaces and making use of the entropic smoothing (18.85) detailed below.

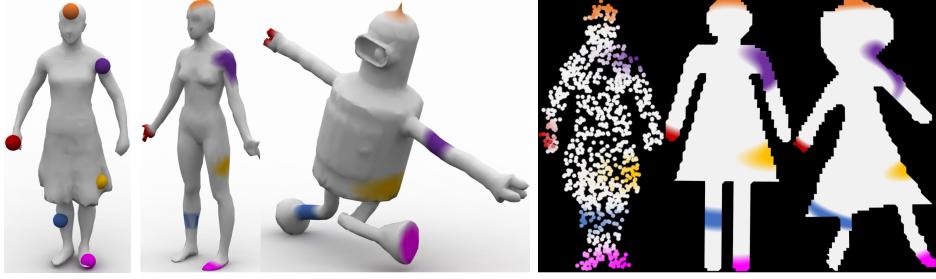


Figure 18.17: Example of fuzzy correspondences computed by solving GW problem (18.85) with Sinkhorn iterations (18.86). Extracted from [?].

To approximate the computation of GW, and to help convergence of minimization schemes to better minima, one can consider the entropic regularized variant

$$\min_{\mathbf{P} \in \mathbf{U}(\mathbf{a}, \mathbf{b})} \mathcal{E}_{\mathbf{D}, \mathbf{D}'}(\mathbf{P}) - \varepsilon \mathbf{H}(\mathbf{P}). \quad (18.85)$$

As proposed initially in [?, ?], and later revisited in [?] for applications in graphics, one can use iteratively Sinkhorn's algorithm to progressively compute a stationary point of (18.85). Indeed, successive linearizations of the objective function lead to consider the succession of updates

$$\mathbf{P}^{(\ell+1)} \stackrel{\text{def.}}{=} \min_{\mathbf{P} \in \mathbf{U}(\mathbf{a}, \mathbf{b})} \langle \mathbf{P}, \mathbf{C}^{(\ell)} \rangle - \varepsilon H(\mathbf{P}) \quad \text{where} \quad (18.86)$$

$$\mathbf{C}^{(\ell)} \stackrel{\text{def.}}{=} \nabla \mathcal{E}_{\mathbf{D}, \mathbf{D}'}(\mathbf{P}^{(\ell)}) = -\mathbf{D}'^T \mathbf{P}^{(\ell)} \mathbf{D},$$

which can be interpreted as a mirror-descent scheme [?]. Each update can thus be solved using Sinkhorn iterations (18.51) with cost $\mathbf{C}^{(\ell)}$. Figure (18.17) illustrates the use of this entropic Gromov-Wasserstein to compute soft maps between domains.

Bibliography

- [1] P. Alliez and C. Gotsman. Recent advances in compression of 3d meshes. In N. A. Dodgson, M. S. Floater, and M. A. Sabin, editors, *Advances in multiresolution for geometric modelling*, pages 3–26. Springer Verlag, 2005.
- [2] P. Alliez, G. Ucelli, C. Gotsman, and M. Attene. Recent advances in remeshing of surfaces. In *AIM@SHAPE report*. 2005.
- [3] Amir Beck. *Introduction to Nonlinear Optimization: Theory, Algorithms, and Applications with MATLAB*. SIAM, 2014.
- [4] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.
- [5] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.
- [6] E. Candès and D. Donoho. New tight frames of curvelets and optimal representations of objects with piecewise C^2 singularities. *Commun. on Pure and Appl. Math.*, 57(2):219–266, 2004.
- [7] E. J. Candès, L. Demanet, D. L. Donoho, and L. Ying. Fast discrete curvelet transforms. *SIAM Multiscale Modeling and Simulation*, 5:861–899, 2005.
- [8] A. Chambolle. An algorithm for total variation minimization and applications. *J. Math. Imaging Vis.*, 20:89–97, 2004.
- [9] Antonin Chambolle, Vicent Caselles, Daniel Cremers, Matteo Novaga, and Thomas Pock. An introduction to total variation for image analysis. *Theoretical foundations and numerical methods for sparse recovery*, 9(263-340):227, 2010.
- [10] Antonin Chambolle and Thomas Pock. An introduction to continuous optimization for imaging. *Acta Numerica*, 25:161–319, 2016.
- [11] S.S. Chen, D.L. Donoho, and M.A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1999.
- [12] F. R. K. Chung. Spectral graph theory. *Regional Conference Series in Mathematics, American Mathematical Society*, 92:1–212, 1997.
- [13] Philippe G Ciarlet. Introduction à l’analyse numérique matricielle et à l’optimisation. 1982.
- [14] P. L. Combettes and V. R. Wajs. Signal recovery by proximal forward-backward splitting. *SIAM Multiscale Modeling and Simulation*, 4(4), 2005.
- [15] P. Schroeder et al. D. Zorin. Subdivision surfaces in character animation. In *Course notes at SIGGRAPH 2000*, July 2000.

- [16] I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Commun. on Pure and Appl. Math.*, 57:1413–1541, 2004.
- [17] I. Daubechies and W. Sweldens. Factoring wavelet transforms into lifting steps. *J. Fourier Anal. Appl.*, 4(3):245–267, 1998.
- [18] D. Donoho and I. Johnstone. Ideal spatial adaptation via wavelet shrinkage. *Biometrika*, 81:425–455, Dec 1994.
- [19] Heinz Werner Engl, Martin Hanke, and Andreas Neubauer. *Regularization of inverse problems*, volume 375. Springer Science & Business Media, 1996.
- [20] M. Figueiredo and R. Nowak. An EM Algorithm for Wavelet-Based Image Restoration. *IEEE Trans. Image Proc.*, 12(8):906–916, 2003.
- [21] M. S. Floater and K. Hormann. Surface parameterization: a tutorial and survey. In N. A. Dodgson, M. S. Floater, and M. A. Sabin, editors, *Advances in multiresolution for geometric modelling*, pages 157–186. Springer Verlag, 2005.
- [22] Simon Foucart and Holger Rauhut. *A mathematical introduction to compressive sensing*, volume 1. Birkhäuser Basel, 2013.
- [23] I. Guskov, W. Sweldens, and P. Schröder. Multiresolution signal processing for meshes. In Alyn Rockwood, editor, *Proceedings of the Conference on Computer Graphics (Siggraph99)*, pages 325–334. ACM Press, August 8–13 1999.
- [24] A. Khodakovsky, P. Schröder, and W. Sweldens. Progressive geometry compression. In *Proceedings of the Computer Graphics Conference 2000 (SIGGRAPH-00)*, pages 271–278, New York, July 23–28 2000. ACM Press.
- [25] L. Kobbelt. $\sqrt{3}$ subdivision. In Sheila Hoffmeyer, editor, *Proc. of SIGGRAPH'00*, pages 103–112, New York, July 23–28 2000. ACM Press.
- [26] M. Lounsbery, T. D. DeRose, and J. Warren. Multiresolution analysis for surfaces of arbitrary topological type. *ACM Trans. Graph.*, 16(1):34–73, 1997.
- [27] S. Mallat. *A Wavelet Tour of Signal Processing, 3rd edition*. Academic Press, San Diego, 2009.
- [28] Stephane Mallat. *A wavelet tour of signal processing: the sparse way*. Academic press, 2008.
- [29] D. Mumford and J. Shah. Optimal approximation by piecewise smooth functions and associated variational problems. *Commun. on Pure and Appl. Math.*, 42:577–685, 1989.
- [30] Neal Parikh, Stephen Boyd, et al. Proximal algorithms. *Foundations and Trends® in Optimization*, 1(3):127–239, 2014.
- [31] Gabriel Peyré. *L’algèbre discrète de la transformée de Fourier*. Ellipses, 2004.
- [32] J. Portilla, V. Strela, M.J. Wainwright, and Simoncelli E.P. Image denoising using scale mixtures of Gaussians in the wavelet domain. *IEEE Trans. Image Proc.*, 12(11):1338–1351, November 2003.
- [33] E. Praun and H. Hoppe. Spherical parametrization and remeshing. *ACM Transactions on Graphics*, 22(3):340–349, July 2003.
- [34] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Phys. D*, 60(1-4):259–268, 1992.
- [35] Otmar Scherzer, Markus Grasmair, Harald Grossauer, Markus Haltmeier, Frank Lenzen, and L Sirovich. *Variational methods in imaging*. Springer, 2009.

- [36] P. Schröder and W. Sweldens. Spherical Wavelets: Efficiently Representing Functions on the Sphere. In *Proc. of SIGGRAPH 95*, pages 161–172, 1995.
- [37] P. Schröder and W. Sweldens. Spherical wavelets: Texture processing. In P. Hanrahan and W. Purgathofer, editors, *Rendering Techniques '95*. Springer Verlag, Wien, New York, August 1995.
- [38] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.
- [39] A. Sheffer, E. Praun, and K. Rose. Mesh parameterization methods and their applications. *Found. Trends. Comput. Graph. Vis.*, 2(2):105–171, 2006.
- [40] Jean-Luc Starck, Fionn Murtagh, and Jalal Fadili. *Sparse image and signal processing: Wavelets and related geometric multiscale analysis*. Cambridge university press, 2015.
- [41] W. Sweldens. The lifting scheme: A custom-design construction of biorthogonal wavelets. *Applied and Computation Harmonic Analysis*, 3(2):186–200, 1996.
- [42] W. Sweldens. The lifting scheme: A construction of second generation wavelets. *SIAM J. Math. Anal.*, 29(2):511–546, 1997.