Sparse Priors

We continue our quest for a better model for the patch prior probability $P_{\mathcal{F}}(\pi_{\mathbf{x}}\mathcal{F})$ to be used in MAP and Bayesian estimators. This time we consider the family of priors based on the assertion that a patch admits a *sparse representation* in some dictionary.

1 Kurtosis and sparsity

Let us start with a little background in probability theory. Let us consider the Gaussian distribution of a random variable given by the density function

$$f_{\mathcal{X}}(x) \propto e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

and fully characterized by the mean μ and variance σ^2. Since the density function has a negative exponential of the square of x, its tails decay very fast. However, there might be other distributions with much "heavier" tails, that is, containing more probability in the tails.

A common measure for the "tailedness" of a distribution is the notion of *kurtosis*, defined as the normalized central fourth moment,

$$\operatorname{Kurt}\mathcal{X} = \mathbb{E}\left(\frac{\mathcal{X}-\mu}{\sigma}\right)^4 = \frac{\mathbb{E}(\mathcal{X}-\mu)^4}{\sigma^4}.$$

All Gaussian distributions happen to have kurtosis 3. It is customary to define the *kurtosis excess* as $\operatorname{Kurt}\mathcal{X} - 3$. Distributions with positive kurtosis excess are called *super-Gaussian* or *leptokurtic* or, colloquially, as "heavy-tailed". Positive curtosis excess arises in two circumstances: when the probability mass is concentrated around the mean and the data-generating process produces occasional values far from the mean, and when the probability mass is concentrated in the tails of the distribution. On the contrary, when the kurtosis excess is negative, the distribution is called *sub-Gaussian* or *platykurtic* or, informally, as "light-tailed".

The *exponential power distribution* given by the density function

$$f_{\mathcal{X}}(x) \propto e^{-(\alpha|x-\mu|)^p}$$

can be thought of a generalized Gaussian distribution with the mean and variance controlled by the parameters μ and α, respectively, and the shape of the distribution controlled by the power p. Note that $p = 2$ yields exactly the Gaussian distribution with $\operatorname{Kurt}\mathcal{X} - 3 = 0$. For $p < 2$, a super-Gaussian family of distributions is obtained with the notable case of the

Laplace (a.k.a. double-exponential) distribution corresponding to $p = 1$. For $p > 2$, the family is sub-Gaussian converging pointwise to the uniform distribution on $[\mu - \alpha, \mu + \alpha]$ as $p \to \infty$.

Let $\mathbf{x}$ be an n-dimensional vector sampled i.i.d. from a zero-mean super-Gaussian random variable. Since the variable will often realize values around zero, and occasionally large non-zero values, most of the elements of the vector will be nearly zero, with some elements (at uniformly random indices) having a large magnitude. This can be thought of as a probabilistic formalization of the statement *"vector* $\mathbf{x}$ *is sparse"*. In this sense, super-Gaussianity leads to sparsity.

super gaussian, kurt > 3, sparse

2 Bases, frames and dictionaries

We will now need a few important notions in linear algebra and functional analysis. Recall that we defined a patch of a signal f centered at $\mathbf{x}$ as the function f translated by $\mathbf{x}$ and restricted to the domain $\square = \left[-\frac{T}{2}, \frac{T}{2}\right]^d$. For notation convenience, in the following treatment we refer to the patch by the name f itself. Recall that any continuous function on $\square$ could be decomposed into the Fourier series

$$f = \sum_{\mathbf{n} \in \mathbb{Z}^d} c_{\mathbf{n}} \phi_{\mathbf{n}},$$

with

$$\phi_{\mathbf{n}}(\mathbf{x}) = e^{i \frac{\mathbf{x}^{\mathrm{T}} \mathbf{n}}{T}}$$

and the coefficients $c_{\mathbf{n}} = \langle f, \phi_{\mathbf{n}} \rangle_{L^2(\square)}$. An important observation here is that the functions $\{\phi_{\mathbf{n}}\}_{\mathbf{n} \in \mathbb{Z}^d}$ are *linearly-independent* and *span* the space of functions $L^2(\square)$. Such a set is called a *basis* of the said space. The crucial feature of a basis is the uniqueness of the decomposition. In other words, there exists only one set of coefficients $\{c_{\mathbf{n}}\}$ representing f in $\{\phi_{\mathbf{n}}\}_{\mathbf{n}}$; in the particular case discussed above, the basis is furthermore orthonormal, so we have a simple formula for computing the coefficients using orthogonal projections on the basis functions[1].

Let us now be given a collection of *linearly-dependent* functions $\{\phi_{\mathbf{n}}\}_{\mathbf{n}}$ spanning $L^2(\square)$. Such vectors no more form a basis and therefore the benefit of unique representation is lost. However, as we will see in the sequel, the loss of uniqueness can be actually a very powerful tool. A formal generalization of a basis is the notion of a *frame* defined through the following condition: if there exists two constants $0 < A \leq B < \infty$ such that for every f in the space

$$A \|f\|^2 \leq \sum_{\mathbf{n} \in \mathbb{Z}^d} |\langle f, \phi_{\mathbf{n}} \rangle|^2 \leq B \|f\|^2,$$

the set $\{\phi_{\mathbf{n}}\}_{\mathbf{n}}$ is called a *frame*. A proper frame (that is, which is not a basis) is called *overcomplete* or *redundant*. Intuitively, it contains "more" functions that a basis would need. Formally, there exist an infinite set of coefficients $c = \{c_{\mathbf{n}}\}$ representing f w.r.t. $\{\phi_{\mathbf{n}}\}$.

[1] In the general case, projection onto the bi-orthonormal set of functions yields the coefficients.

We will associate with the frame the *synthesis operator* $\Phi : \ell^2 \to L^2(\square)$ producing

$$\Phi c = \sum_{\mathbf{n}\in\mathbb{Z}^d} c_{\mathbf{n}}\phi_{\mathbf{n}}$$

from the sequence of coefficients $\{c_{\mathbf{n}}\}$. The adjoint *analysis operator* $\Phi^* : L^2(\square) \to \ell^2$ mapps f to c.

In the signal processing jargon, it is customary to speak of a frame as of a *dictionary* and of the constituent functions as of *atoms*. Then, a function f can be represented as a superposition of atoms. If the dictionary is overcomplete, such a representation is not unique.

For example, the set of functions

$$\phi_{\mathbf{n}}(\mathbf{x}) = e^{i\frac{\mathbf{x}^{\mathrm{T}}\mathbf{n}}{aT}}$$

with $a > 1$ forms a frame on $\square$. We can think of it as an overcomplete Fourier transform dictionary. Taking the real value yields the overcomplete cosine dictionary,

$$\phi_{\mathbf{n}}(\mathbf{x}) = \cos\left(\frac{\mathbf{x}^{\mathrm{T}}\mathbf{n}}{aT}\right).$$

There exists a plethora of other useful dictionaries such as wavelets, ridgelets, curvelets, countourlets (which all have a beautiful theory) and there also exists a possibility to *learn* the dictionary (optimal in some sense) from examples. The latter approach has been shown advantageous in many applications.

Discrete dictionaries When the patch domain is *sampled*, say, on the lattice $\frac{T}{N}\mathbb{Z}^d$, f is given on a finite set of $(2N+1)^d$ points

$$f[\mathbf{n}] = f\left(\frac{T}{N}\mathbf{n}\right)$$

with $\mathbf{n} \in \{-N, \dots, N\}^d$. Reordering the samples $f[\mathbf{n}]$ into a long $n = (2N+1)^d$-dimensional vector $\mathbf{f}$, we can think of the dictionary representation as

$$\mathbf{f} = \mathbf{\Phi c},$$

where $\mathbf{\Phi}$ is an $n \times k$ matrix whose columns are the sampled versions of the frame functions $\phi_{\mathbf{k}}$ sampled and reodered into n-dimensional vectors, $\boldsymbol{\phi}_{\mathbf{k}}$ that are ordered in some order such that there are $k = an > n$ functions. The overcomplete cosine frame readily becomes the overcomplete *discrete cosine transform* (DCT) dictionary.

3 Sparse patch priors

A crucial empirical observation is that patches of natural images can be approximated by a small number of atoms in many overcomplete dictionaries such as the overcomplete cosine dictionary. We will formalize this by letting

$$\pi_{\mathbf{x}}\mathcal{F} = \Phi c + \mathcal{E} = \sum_{\mathbf{n}\in\mathbb{Z}^d} c_{\mathbf{n}}\phi_{\mathbf{n}} + \mathcal{E}$$

and asserting that the coefficients $c_{\mathbf{n}}$ are i.i.d. with a super-Gaussian distribution, while the residual $\mathcal{E}$ is white Gaussian with the variance $\sigma_{\mathcal{E}}^2$.

This leads to the following negative log likelihood:

$$-\log f_{\pi_{\mathbf{x}}\mathcal{F}|c(\mathbf{x})}\left(\pi_{\mathbf{x}}\mathcal{F} = f|c(\mathbf{x}) = \{c_{\mathbf{n}}(\mathbf{x})\}\right) = -\log f_{\mathcal{S}}\left(f - \Phi c(\mathbf{x})\right) = \sigma_{\mathcal{E}}^2 \left\|f - \Phi c(\mathbf{x})\right\|_{L^2(\square)}^2 + \text{const}$$

with the negative log prior density of the coefficients,

$$-\log f_c\left(c(\mathbf{x})\right) = \sum_{\mathbf{n}\in\mathbb{Z}^d} -\log f_c(c_{\mathbf{n}}(\mathbf{x})).$$

For example, using the exponential power family for f_c leads to

$$-\log f_c\left(c(\mathbf{x})\right) = \alpha^p \sum_{\mathbf{n}\in\mathbb{Z}^d} |c_{\mathbf{n}}(\mathbf{x})|^p + \text{const}.$$

Invoking the Bayes theorem leads to the following posterior density

$$-\log f_{c(\mathbf{x})|\pi_{\mathbf{x}}\mathcal{F}}\left(c(\mathbf{x})|f\right) = \sigma_{\mathcal{E}}^2 \left\|f - \Phi c(\mathbf{x})\right\|_{L^2(\square)}^2 + \alpha^p \sum_{\mathbf{n}\in\mathbb{Z}^d} |c_{\mathbf{n}}(\mathbf{x})|^p + \text{const}.$$

Note that in order to promote sparsity of the coefficients $p < 2$ has to be small. However, for $p < 1$, the negative log density is non-convex, which is a major complication as we eventually would like to add it as a prior term to our optimization problem in a Bayesian or MAP estimatior. A pragmatic choice is $p = 1$, the smallest value of p keeping the term convex (yet making it non-smooth at zero). This choice corresponds to the Laplacian distribution, leading to the following aggregate of the L_2 norm with the ℓ_1 norm:

$$\begin{aligned} -\log f_{c(\mathbf{x})|\pi_{\mathbf{x}}\mathcal{F}}\left(c(\mathbf{x})|f\right) &= \sigma_{\mathcal{E}}^2 \left\|f - \Phi c(\mathbf{x})\right\|_{L^2(\square)}^2 + \alpha \sum_{\mathbf{n}\in\mathbb{Z}^d} |c_{\mathbf{n}}(\mathbf{x})| + \text{const} \\ &= \left\|f - \Phi c(\mathbf{x})\right\|_{L^2(\square)}^2 + \lambda \sum_{\mathbf{n}\in\mathbb{Z}^d} |c_{\mathbf{n}}(\mathbf{x})| + \text{const}, \end{aligned}$$

where $\lambda = \dfrac{\alpha}{\sigma_{\mathcal{E}}^2}$.

Some flavors of sparse priors use dictionaries with non-negative functions and further assert non-negative coefficients, $c_{\mathbf{n}}(\mathbf{x}) \geq 0$. This incorporates the prior information that the signal is non-negative, and is often employed for modelling images and spectral magnitudes of audio signals.

4 MAP estimation with a sparse patch prior

At this point, we have two possibilities. We can rephrase our model as

$$\pi_{\mathbf{x}}\mathcal{Y} = \Phi c + \mathcal{E}$$

estimate the contents of a patch directly as

$$\pi_{\mathbf{x}}\hat{f} = \Phi\hat{c}$$

where

$$\begin{aligned}\hat{c} &= \arg\min_c \log f_{c|\pi_{\mathbf{x}}\mathcal{Y}}\left(c|\pi_{\mathbf{x}}\mathcal{Y} = \pi_{\mathbf{x}}y\right) \\ &= \arg\min_c \|\pi_{\mathbf{x}}y - \Phi c\|^2_{L^2(\square)} + \lambda\sum_{\mathbf{n}\in\mathbb{Z}^d}|c_{\mathbf{n}}|.\end{aligned}$$

This can be done for every $\mathbf{x}$; afterwards, the overlapping estimated patches are aggregated by simple averaging. However, such an avergaing steers us farther away from the assumption of sparse representation, as a sum of sparsely represented functions is less sparsely represented. A cure can be a smarter way to aggregate the patches.

An alternative approach is to estimate the entire signal f by solving

$$\hat{f} = \arg\min_{f,c}\mu\|f-y\|^2_{L^2(\mathbb{R}^d)} + \int_{\mathbb{R}^d}\left(\|\pi_{\mathbf{x}}f - \Phi c(\mathbf{x})\|^2_{L^2(\square)} + \lambda\sum_{\mathbf{n}\in\mathbb{Z}^d}|c_{\mathbf{n}}(\mathbf{x})|\right)d\mathbf{x}$$

simultaneously for f and $\{c_{\mathbf{n}}(\mathbf{x})\}$ at *all* patches.

5 Iterative shrinkage

Regardless of which flavor of the above MAP estimators we choose, both require the minimization of an aggregate of the L_2 and ℓ_1 norms. We will rewrite this problem in the form

$$\begin{aligned}\arg\min_c \frac{1}{2}\|\Phi c - y\|^2_{L^2(\square)} + \lambda\sum_{\mathbf{n}\in\mathbb{Z}^d}|c_{\mathbf{n}}| &= \arg\min_c \frac{1}{2}\langle\Phi c - y, \Phi c - y\rangle + \lambda\|c\|_1 \\ &= \arg\min_c \frac{1}{2}\langle\Phi c, \Phi c\rangle - \langle\Phi c, y\rangle + \frac{1}{2}\langle y, y\rangle + \lambda\|c\|_1 \\ &= \arg\min_c \frac{1}{2}\langle\Phi^*\Phi c, c\rangle - \langle\Phi^* y, c\rangle + \lambda\|c\|_1,\end{aligned}$$

where the inner products are on $L^2(\square)$ and the ℓ_1 norm is on the space of sequences. This problem bears the name of Lasso (short for *least absolute shrinkage and selection operator*) in statistics.

Note that the objective is a function of the sequence c with the first two terms differentiable that we will denote as

$$g(c) \;=\; \frac{1}{2}\langle \Phi^*\Phi c, c\rangle - \langle \Phi^* y, c\rangle,$$

and a non-differentiable third term,

$$h(c) = \lambda \|c\|_1.$$

Let us fix some c and approximate $g(c)$ around c as

$$\begin{aligned} g(u-c) &\approx g(c) + \langle \nabla g(c), (u-c)\rangle_{\ell^2} + \frac{1}{2\eta}\|u-c\|_{\ell^2}^2 \\ &= \frac{1}{2\eta}\,\|u - c + \eta\nabla g(c)\|_{\ell^2}^2 - \frac{\eta}{2}\|\nabla g(c)\|_{\ell^2}^2 + g(c) \\ &= \frac{1}{2\eta}\,\|u - (c - \eta\nabla g(c))\|_{\ell^2}^2 + \text{const.} \end{aligned}$$

Here η controls the curvature of the second-order term in the approximation. Plugging this approximation into the minimization problem yields

$$\begin{aligned} \arg\min_u g(u-c) + h(u) &= \arg\min_u \frac{1}{2\eta}\,\|u - (c - \eta\nabla g(c))\|_{\ell^2}^2 + \lambda\|u\|_1 \\ &= \arg\min_u \frac{1}{2}\,\|u - z\|_{\ell^2}^2 + \eta\lambda\|u\|_1 \\ &= \arg\min_u \sum_{\mathbf{n}\in\mathbb{Z}^d} \frac{1}{2}(u_{\mathbf{n}} - z_{\mathbf{n}})^2 + \eta\lambda|u_{\mathbf{n}}| \\ &= \{\arg\min_u \frac{1}{2}(u - z_{\mathbf{n}})^2 + \eta\lambda|u|\}_{\mathbf{n}\in\mathbb{Z}^d}, \end{aligned}$$

where

$$z = c - \eta\nabla g(c) = c - \eta\,\Phi^*(\Phi c - y).$$

Note that the latter problem is coordinate-separable, so we can consider the following one-dimensional minimization problem:

$$\arg\min_u \frac{1}{2}(u-z)^2 + \lambda|u|$$

Since the problem is non-smooth at $u = 0$, we cannot readily take a derivative w.r.t. to u and compare it to zero; instead, we have to use the sub-differential set

$$\partial|u| = \begin{cases} \operatorname{sign} u & : u \neq 0 \\ \text{[-1,1]} & : u = 0 \end{cases}$$

This yields $u = y - \lambda\alpha$ for $\alpha \in \partial|u|$. Whenever $z \in [-\lambda, \lambda]$, we can set $u = 0$ with $\alpha = \dfrac{z}{\lambda} \in (\partial|u|\big|_{u=0}$. Otherwise, $u \neq 0$ and we have $\alpha = \operatorname{sign} u$ yielding $u = z - \lambda \operatorname{sign} u$. If

$z > \lambda$, the right hand side is positive, hence $u > 0$ and $\operatorname{sign} u = 1$; this yields $u = z - \lambda$. Similarly, for $y < -\lambda$, we have $u < 0$ and hence $u = z + \lambda$. We can therefore summarize the solution to the problem as

$$\arg\min_u \frac{1}{2}(u-z)^2 + \lambda|u| = \begin{cases} 0 & : -\lambda \le z \le \lambda \\ z - \lambda & : z > \lambda \\ z + \lambda & : z < -\lambda \end{cases}$$

This operation on z is called *soft thresholding* or *shrinkage*[2] and will be denoted by $u = \mathcal{S}_\lambda(z)$.

Plugging this result back into our original multi-dimensional problem yields

$$\arg\min_u g(u-c) + h(u) \quad = \quad \{\mathcal{S}_{\eta\lambda}(z_{\mathbf{n}})\}_{\mathbf{n}\in\mathbb{Z}^d} = \mathcal{S}_{\eta\lambda}(c - \eta\,\Phi^*(\Phi c - y)),$$

where in the last passage the shrinkage operator $\mathcal{S}_{\eta\lambda}$ is applied element-wise to the sequence. We can repeat the process iterartively starting at some initial c^0 (upperscript indices denote iteration number),

$$c^{k+1} = \mathcal{S}_{\eta\lambda}(c^k - \eta\,\Phi^*(\Phi c^k - y)),$$

yielding a process known as *iterative shrinkage* (a.k.a. iterative shrinkage and thresholding algorithm or ISTA for short)[3]. Note that the step $c^k - \eta\,\Phi^*(\Phi c^k - y)$ inside the shrinkage operator is a gradient descent step at point c^k with the step size η, and it would be exactly the step if the objective lacked the term $h(c)$. The shrinkage operator accounts for this additional term.

Non-negativity In the case of non-negative dictionaries with non-negative coefficients, we can modify our one-dimensional Lasso problem by adding a non-negativity constraint on u,

$$\arg\min_{u\ge 0} \frac{1}{2}(u-z)^2 + \lambda|u|$$

The solution is obtained in the same manner, except thart now u cannot attain negative values. This leads to the one-sided shrinkage operator

$$\arg\min_{u\ge 0} \frac{1}{2}(u-z)^2 + \lambda|u| = \begin{cases} 0 & : z \le \lambda \\ z - \lambda & : z > \lambda \end{cases} = \mathcal{R}_\lambda(z).$$

This operator is known under the name of *rectified linear unit* (or ReLU for short) in the deep learning literature. We will explore this surprising connection more in the sequel.

[2] This operator can be viewed as the *proximity map* of the function $\lambda|c|$.

[3] This is a member of a larger family of non-smooth optimization algorithms known as proximal methods.