Linear Inverse Problems (1/2)

Mathematical Foundations of Signal Processing

Dr. Matthieu Simeoni

November 9, 2020

EPFL

Table of contents

Introduction

Most real-life approximation problems can be formulated as inverse problems:

Inverse Problem

Consider an unknown *signal* $f \in \mathscr{L}^2\left(\mathbb{R}^d\right)$ and assume that the latter is *probed* by some sensing device, resulting in a data vector $\boldsymbol{y} = [y_1, \dots, y_L] \in \mathbb{R}^L$ of L measurements. Recovering f from the data vector $\boldsymbol{y}$ is called an inverse problem.

We make the following assumptions:

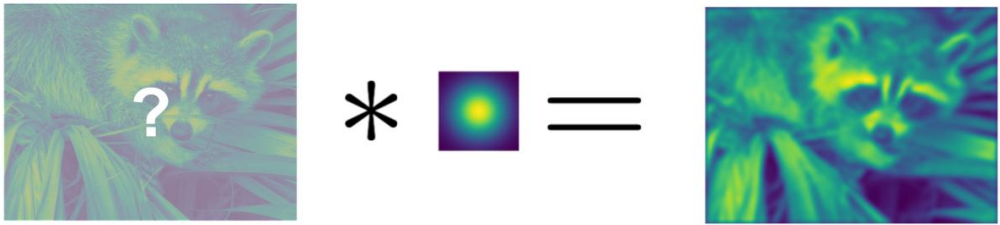
1. To account for sensing *inaccuracies*, the data vector $\boldsymbol{y}$ is assumed to be the outcome of a random vector $\boldsymbol{Y} = [Y_1, \dots, Y_L] : \Omega \to \mathbb{R}^L$, fluctuating according to some noise distribution. The entries of $\mathbb{E}[\boldsymbol{Y}] = \tilde{\boldsymbol{y}}$ are called the ideal measurements –these are the measurements that would be obtained in the absence of noise.
2. The measurements are assumed unbiased and linear, i.e. $\mathbb{E}[\boldsymbol{Y}] = \Phi^* f = \left[\langle f, \varphi_1 \rangle, \dots, \langle f, \varphi_L \rangle\right]$, for some sampling functionals $\{\varphi_1, \dots, \varphi_L\} \subset \mathscr{L}^2\left(\mathbb{R}^d\right)$, modelling the acquisition system.

Common Sampling Functionals

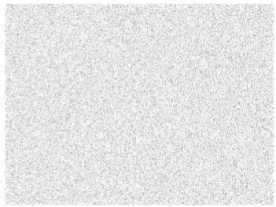
Common Sampling Functionals

- **Spatial Sampling:**
  $\tilde{y}_i = f(\boldsymbol{x}_i) = \int_{\mathbb{R}^d} f(\boldsymbol{x})\delta(\boldsymbol{x}-\boldsymbol{x}_i)\,d\boldsymbol{x} \quad \rightarrow \quad \varphi_i(\boldsymbol{x}) = \delta(\boldsymbol{x}-\boldsymbol{x}_i), \quad \boldsymbol{x}_i \in \mathbb{R}^d.$
- **Fourier Sampling:**
  $\tilde{y}_{i1} = \int_{\mathbb{R}^d} f(\boldsymbol{x}) \cos\left(\langle \boldsymbol{x}, \boldsymbol{\omega}_i \rangle\right) d\boldsymbol{x} \quad \rightarrow \quad \varphi_{i1}(\boldsymbol{x}) = \cos\left(\langle \boldsymbol{x}, \boldsymbol{\omega}_i \rangle\right), \quad \boldsymbol{\omega}_i \in \mathbb{R}^d.$
  $\tilde{y}_{i2} = \int_{\mathbb{R}^d} f(\boldsymbol{x}) \sin\left(\langle \boldsymbol{x}, \boldsymbol{\omega}_i \rangle\right) d\boldsymbol{x} \quad \rightarrow \quad \varphi_{i2}(\boldsymbol{x}) = \sin\left(\langle \boldsymbol{x}, \boldsymbol{\omega}_i \rangle\right), \quad \boldsymbol{\omega}_i \in \mathbb{R}^d.$
- **Radon Sampling:**
  $\tilde{y}_i = \check{f}(p_i, \boldsymbol{\xi}_i) = \int_{\mathbb{R}^d} f(\boldsymbol{x})\delta\left(p_i - \langle \boldsymbol{x}, \boldsymbol{\xi}_i \rangle\right) d\boldsymbol{x} \quad \rightarrow \quad \varphi_i(\boldsymbol{x}) = \delta\left(p_i - \langle \boldsymbol{x}, \boldsymbol{\xi}_i \rangle\right), \quad p_i > 0, \boldsymbol{\xi}_i \in \mathbb{S}^{N-1}.$
- **Filtering:**
  $\tilde{y}_i = \{f * \varphi\}(\boldsymbol{x}_i) = \int_{\mathbb{R}^d} \varphi(\boldsymbol{x}_i - \boldsymbol{x}) f(\boldsymbol{x})\,d\boldsymbol{x} \quad \rightarrow \quad \varphi_i(\boldsymbol{x}) = \varphi(\boldsymbol{x}_i - \boldsymbol{x}), \quad \boldsymbol{x}_i \in \mathbb{R}^d, \varphi : \mathbb{R}^d \rightarrow \mathbb{R}.$
- **Mean-Pooling:**
  $\tilde{y}_i = \frac{1}{|\Omega_i|}\int_{\Omega_i} f(\boldsymbol{x})\,d\boldsymbol{x} \quad \rightarrow \quad \varphi_i(\boldsymbol{x}) = \frac{1}{|\Omega_i|}\chi_{\Omega_i}(\boldsymbol{x}) := \begin{cases} |\Omega_i|^{-1} & \text{if } \boldsymbol{x} \in \Omega_i \\ 0 & \text{otherwise} \end{cases}, \Omega_i \subset \mathbb{R}^d.$

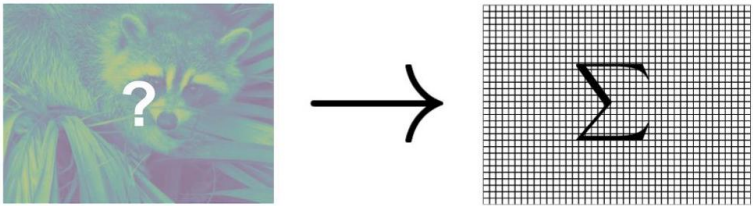
Example: Deblurring

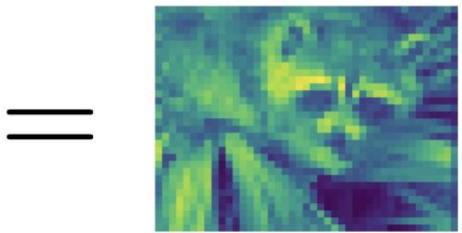


Example: Inpainting



Example: Unpooling



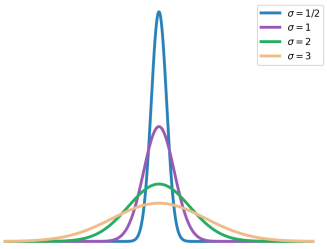


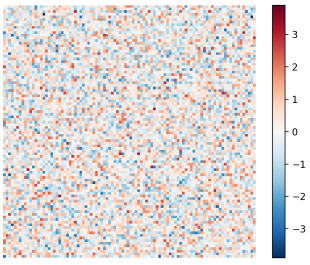
Gaussian White Noise

Assume that sensor inaccuracies are independent and result from the *sum of many independent perturbations*. Then, from the central limit theorem, sensor inaccuracies can be modelled as independent realisations of an *additive* Gaussian white noise:

$$Y_i = \tilde{y}_i + N_i, \quad \text{where} \quad N_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma^2\right), \quad p_N(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\sigma^2}\right), \quad x \in \mathbb{R},$$

where p_N is the noise probability density function. Notice that we have indeed $\mathbb{E}[Y_i] = \tilde{y}_i$ for each $i = 1, \ldots, L$.



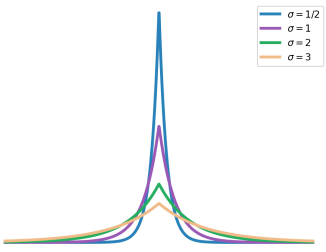


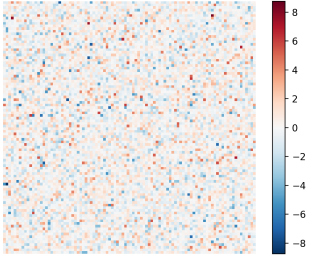
Laplacian/Salt-and-pepper White Noise

Assume that sensor inaccuracies are independent and present strong outliers (for example due to *malfunctioning* sensors). Then, sensor inaccuracies can be modelled as independent realisations of an *additive* Laplacian white noise, also called salt-and-pepper noise:

$$Y_i = \tilde{y}_i + N_i, \quad \text{where} \quad N_i \overset{\text{i.i.d.}}{\sim} \text{Laplace}\,(0, \sigma), \quad p_N(x) = \frac{1}{2\sigma} \exp\left(-\frac{|x|}{\sigma}\right), \quad x \in \mathbb{R},$$

where p_N is the noise probability density function. Notice that we have indeed $\mathbb{E}[Y_i] = \tilde{y}_i$ for each $i = 1, \ldots, L$.



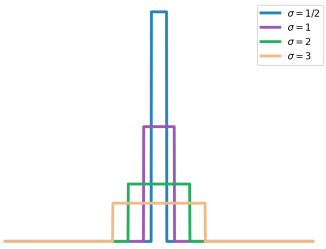


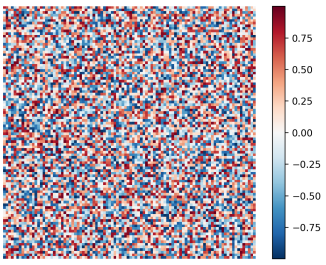
Uniform/Quantisation White Noise

Assume that sensor inaccuracies are independent and primarily caused by quantisation artefacts –i.e. round-off errors incurred by storing digits with finite precision. Then, sensor inaccuracies can be modelled as independent realisations of an *additive* uniform white noise, also called quantisation noise:

$$Y_i = \tilde{y}_i + N_i, \quad \text{where} \quad N_i \overset{\text{i.i.d.}}{\sim} U\left(-\frac{\sigma}{2}, \frac{\sigma}{2}\right), \quad p_N(x) = \begin{cases} 1/\sigma & \text{if } x \in [-\sigma/2, \sigma/2] \\ 0 & \text{if } x \notin [-\sigma/2, \sigma/2]. \end{cases},$$

where p_N is the noise probability density function. Notice that we have indeed $\mathbb{E}[Y_i] = \tilde{y}_i$ for each $i = 1, \ldots, L$.



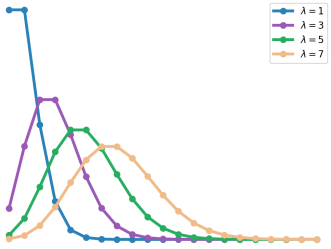


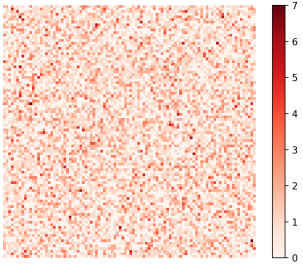
Poisson/Shot Noise

Assume that the measurements are independent and originate from a counting process –i.e. $\boldsymbol{Y}: \Omega \to \in \mathbb{N}^L$. Then, sensor inaccuracies can be modelled as independent realisations of a non additive Poisson noise, also called shot noise:

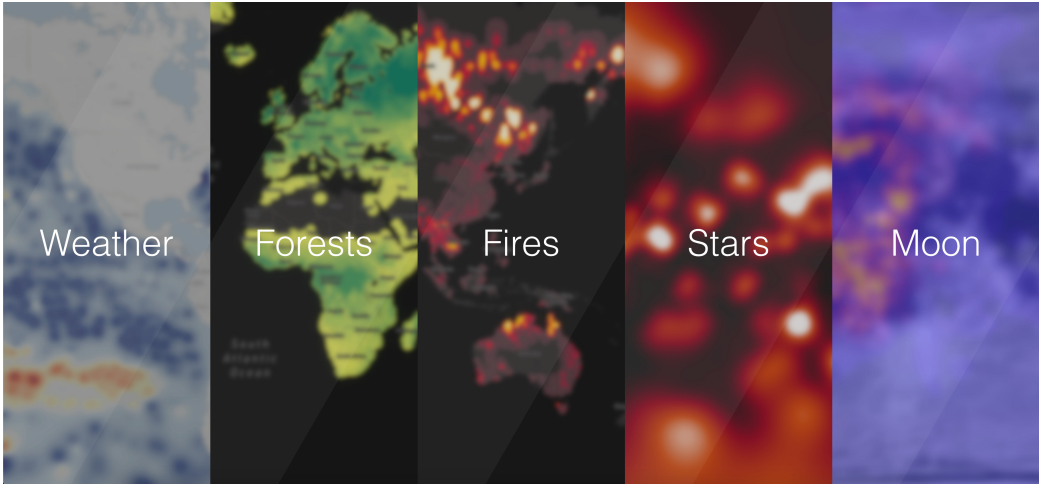
$$Y_i \overset{\text{ind}}{\sim} \text{Poisson}\left(\tilde{y}_i\right), \quad p_{Y_i}(k) = \frac{\tilde{y}_i^k e^{-\tilde{y}_i}}{k!}, \quad \forall k \in \mathbb{N},$$

where p_{Y_i} is the probability density function for the ith measurement. Using properties from the Poisson distribution, we can indeed show that $\mathbb{E}[Y_i] = \tilde{y}_i$ for each $i = 1, \ldots, L$.





Real-Life Examples: Meteorology, Forestry, Astronomy...



https://matthieumeo.github.io/

Pixelisation

Since the number of measurements is finite, it is reasonable to constrain the signal f to be finite-dimensional:[1]

$$f = \sum_{n=1}^{N} \alpha_n \psi_n = \Psi \boldsymbol{\alpha}, \qquad \boldsymbol{\alpha} = [\alpha_1, \dots, \alpha_N] \in \mathbb{R}^N \tag{1}$$

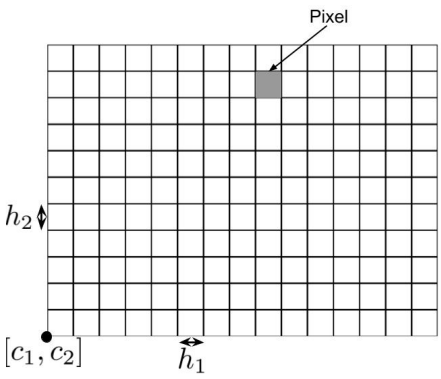
for some suitable basis functions $\{\psi_n,\, n = 1, \dots, N\} \subset \mathscr{L}^2(\mathbb{R}^d)$. Typically, the basis functions are chosen as indicator functions of regular rectangular tiles of $\mathbb{R}^d$ called pixels. For example:

$$\psi_n(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \boldsymbol{x} \in [c_1 + (n-1)h_1, c_1 + nh_1] \times \cdots \times \left[c_d + (n-1)h_d, c_d + nh_d\right], \\ 0 & \text{otherwise,} \end{cases}$$

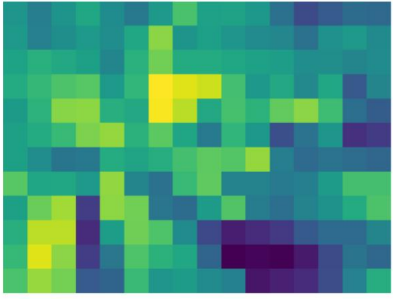
where $\boldsymbol{c} = [c_1, \dots, c_d]$ are the coordinates of the lower-left corner of the first pixel, and $\{h_1, \dots, h_d\}$ are the sizes of the pixels across each dimension. The parametric signal f in (1) is then a piecewise constant signal than can be **stored/manipulated/displayed** *efficiently* via multi-dimensional array (hence the popularity of pixel-based discretisation schemes).

[1] Infinite-dimensional signals may indeed have an infinite number of degrees of freedom, which cannot hope to estimate from a finite number of measurements only.

Pixelisation



$$f(x,y) = \sum_{n=1}^{N} \alpha_n \psi_n(x,y)$$



Discrete Inverse Problems

Assuming the parametric model (1) induces a discrete inverse problem:

$$\textit{Find } \boldsymbol{\alpha} \in \mathbb{R}^N \textit{ from the noisy measurements } \boldsymbol{y} \rightsquigarrow \boldsymbol{Y} \textit{ where } \mathbb{E}[\boldsymbol{Y}] = \Phi^* \Psi \boldsymbol{\alpha} = \boldsymbol{G}\boldsymbol{\alpha}.$$

The operator $\boldsymbol{G} : \mathbb{R}^N \to \mathbb{R}^L$ is a rectangular matrix given by:[2]

$$\mathbb{R}^{L \times N} \ni \boldsymbol{G} = \begin{bmatrix} \langle \psi_1, \varphi_1 \rangle & \cdots & \langle \psi_N, \varphi_1 \rangle \\ \vdots & \ddots & \vdots \\ \langle \psi_1, \varphi_L \rangle & \cdots & \langle \psi_1, \varphi_L \rangle \end{bmatrix} = \begin{bmatrix} \int_{\Omega_1} \varphi_1(\boldsymbol{x}) d\boldsymbol{x} & \cdots & \int_{\Omega_N} \varphi_1(\boldsymbol{x}) d\boldsymbol{x} \\ \vdots & \ddots & \vdots \\ \int_{\Omega_1} \varphi_L(\boldsymbol{x}) d\boldsymbol{x} & \cdots & \int_{\Omega_N} \varphi_L(\boldsymbol{x}) d\boldsymbol{x} \end{bmatrix}$$

$$\simeq \eta \begin{bmatrix} \varphi_1(\boldsymbol{\xi}_1) & \cdots & \varphi_1(\boldsymbol{\xi}_N) \\ \vdots & \ddots & \vdots \\ \varphi_L(\boldsymbol{\xi}_1) & \cdots & \varphi_L(\boldsymbol{\xi}_N) \end{bmatrix},$$

where $\eta = \prod_{k=1}^{d} h_k$, and $\{\Omega_n\}_n \in \mathscr{P}(\mathbb{R}^d)$ and $\{\boldsymbol{\xi}_n\}_n \subset \mathbb{R}^d$ are the *supports* and *centroids* of each pixel, respectively.

[2] The last approximate equality results from the midpoint rule.

Inverse Problems are Ill-Posed

To solve the inverse problem one can approximate the mean $\mathbb{E}[Y]$ by its *one-sample empirical estimate* $\boldsymbol{y}$ and solve the linear problem:

$$\boldsymbol{y} = \boldsymbol{G}\boldsymbol{\alpha}. \tag{2}$$

Unfortunately, (2) is in general ill-posed:

1. **There may exist no solutions to (2).** If $N < L$ indeed (or more generally if $\boldsymbol{G}$ is not surjective), $\mathscr{R}(\boldsymbol{G}) \subsetneq \mathbb{R}^N$. Therefore the noisy data vector $\boldsymbol{y}$ is not guaranteed to belong to $\mathscr{R}(\boldsymbol{G})$.
2. **There may exist more than one solution to (2).** If $N > L$ indeed (or more generally if $\boldsymbol{G}$ is not injective), $\mathscr{N}(\boldsymbol{G}) \neq \{\boldsymbol{0}\}$. Therefore, if $\boldsymbol{\alpha}^\star$ is a solution to (2), then $\boldsymbol{\alpha}^\star + \boldsymbol{\beta}$ is also a solution $\forall \boldsymbol{\beta} \in \mathscr{N}(\boldsymbol{G})$:

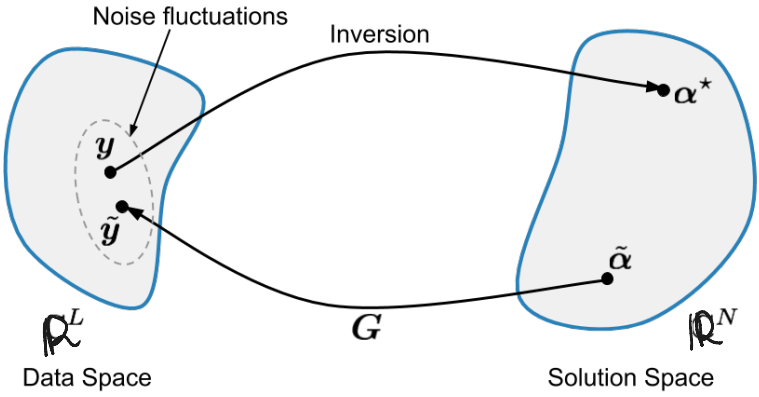
$$\boldsymbol{G}(\boldsymbol{\alpha}^\star + \boldsymbol{\beta}) = \boldsymbol{G}\boldsymbol{\alpha}^\star + \boldsymbol{G}\boldsymbol{\beta} = \boldsymbol{G}\boldsymbol{\alpha}^\star = \boldsymbol{y}.$$

3. **Solutions to (2) may be numerically unstable.** If $\boldsymbol{G}$ is surjective for example, then $\boldsymbol{G}^\dagger = \boldsymbol{G}^T(\boldsymbol{G}\boldsymbol{G}^T)^{-1}$ is a right-inverse of $\boldsymbol{G}$ and $\boldsymbol{\alpha}^\star(\boldsymbol{y}) = \boldsymbol{G}^T(\boldsymbol{G}\boldsymbol{G}^T)^{-1}\boldsymbol{y}$ is a solution to (2). We have then

$$\|\boldsymbol{\alpha}^\star(\boldsymbol{y})\|_2 \leq \|\boldsymbol{G}\|_2 \|(\boldsymbol{G}^T\boldsymbol{G})^{-1}\|_2 = \underbrace{\frac{\sqrt{\lambda_{max}(\boldsymbol{G}^T\boldsymbol{G})}}{\lambda_{min}(\boldsymbol{G}^T\boldsymbol{G})}}_{\text{Can be very large!}} \|\boldsymbol{y}\|_2, \qquad \forall \boldsymbol{y} \in \mathbb{R}^L.$$

The reconstruction linear map $\boldsymbol{y} \mapsto \boldsymbol{\alpha}^\star(\boldsymbol{y})$ can hence be virtually unbounded making it *unstable*.

Inverse Problems are Unstable



Small perturbations on the data affect greatly the solution!

Regularising Inverse Problems

The linear system (2) is not only ill-posed but also non sensible: matching exactly the measurements is not desirable since the latter are in practice corrupted by instrumental noise.

A more sensible approach consists instead in solving the inverse problem by means of a penalised optimisation problem, confronting the physical evidence to the analyst's a priori beliefs about the solution (e.g. smoothness, sparsity) via a data-fidelity and regularisation term, respectively:

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^N} F(\boldsymbol{y}, \boldsymbol{G}\boldsymbol{\alpha}) \quad + \quad \lambda \mathscr{R}(\boldsymbol{\alpha}). \tag{3}$$

The various quantities involved in (3) can be interpreted as follows:

- $F : \mathbb{R}^L \times \mathbb{R}^L \to \mathbb{R}_+ \cup \{+\infty\}$ is a cost/data-fidelity functional, measuring the discrepancy between the observed and predicted measurements $\boldsymbol{y}$ and $\boldsymbol{G}\boldsymbol{\alpha}$ respectively.
- $\mathscr{R} : \mathbb{R}^N \to \mathbb{R}_+ \cup \{+\infty\}$ is a regularisation/penalty functional favouring simple and well-behaved solutions (typically with a finite number of degrees of freedom).
- $\lambda > 0$ is a regularisation/penalty parameter which controls the amount of regularisation by putting the regularisation functional and the cost functional on a similar scale.

Existence of Solutions

Theorem: (Existence of Solutions to (3))

Consider the following set of assumptions:

1. For all $\boldsymbol{y} \in \mathbb{R}^L$, the univariate cost trace functionals

$$F(\boldsymbol{y}, \cdot) : \begin{cases} \mathbb{R}^L \to \mathbb{R}_+ \cup \{+\infty\} \\ \boldsymbol{z} \mapsto F(\boldsymbol{y}, \boldsymbol{z}) \end{cases}$$

   and the regularisation functional $\mathscr{R} : \mathbb{R}^N \to \mathbb{R}_+ \cup \{+\infty\}$ are proper, convex and lower semi-continuous (see Slide 20 for a definition).

2. The objective functional of (3) is coercive, i.e. $\lim_{\|\boldsymbol{\alpha}\|_2 \to +\infty} F(\boldsymbol{y}, \boldsymbol{G}\boldsymbol{\alpha}) + \lambda \mathscr{R}(\boldsymbol{\alpha}) = +\infty$.

Then, the solution set $\mathcal{V} = \operatorname{argmin}_{\boldsymbol{\alpha} \in \mathbb{R}^N} F(\boldsymbol{y}, \boldsymbol{G}\boldsymbol{\alpha}) + \lambda \mathscr{R}(\boldsymbol{\alpha})$, is non empty, convex and compact.[3]

The proof of this theorem can be deduced from [1, Proposition 8] (for reference only do not check it!).

[3] In finite dimension, a compact set is a closed and bounded set.

Proper, Convex, Lower Semi-Continuous Functional

Definition: (Proper Convex Functional)

A function $F : \mathbb{R}^N \to \mathbb{R} \cup \{-\infty, +\infty\}$ is called convex if

$$\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^N,\ \forall \theta \in [0,1]: \qquad F\big(\theta \boldsymbol{x} + (1-\theta)\boldsymbol{y}\big) \leq \theta F(\boldsymbol{x}) + (1-\theta) F(\boldsymbol{y}), \tag{4}$$

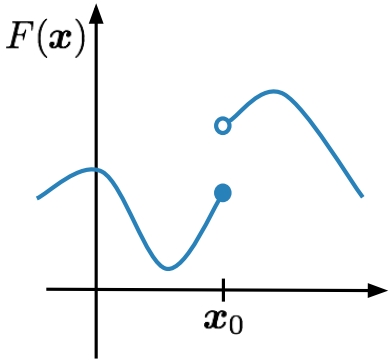
and strictly convex if the inequality in (4) is strict. If moreover, $F(\boldsymbol{x}) > -\infty$ for all $\boldsymbol{x} \in \mathbb{R}^N$ and $D = \{\boldsymbol{x} \in \mathbb{R}^N : F(\boldsymbol{x}) < +\infty\} \neq \emptyset$, then F is called a proper (strictly) convex function.[4]

Definition: (Lower Semi-Continuity)

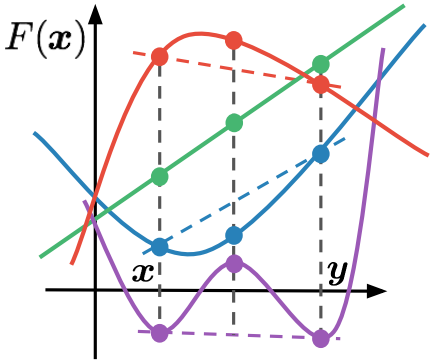
A function $F : \mathbb{R}^N \to \mathbb{R} \cup \{-\infty, +\infty\}$ is said lower semi-continuous (lwsc) at $\boldsymbol{x}_0 \in \mathbb{R}^N$ if for every $y < F(\boldsymbol{x}_0)$ there exists a neighborhood $U \subset \mathbb{R}^N$ of $\boldsymbol{x}_0$ such that $F(\boldsymbol{x}) \geq y$, $\forall \boldsymbol{x} \in U$.

[4] In short, a convex function is proper if its domain is nonempty and it never attains $-\infty$.

Proper, Convex, Lower Semi-Continuous Functional



Example of Lower Semi-Continuous Function



Example of Strictly Convex, Convex, Concave and Non Convex Functions

Unicity of Solutions

Theorem: (Unicity of Solutions)

Assume that F and $\mathscr{R}$ are as in Slide 19 and that the objective functional $\mathscr{J}(\boldsymbol{\alpha}) := F(\boldsymbol{y}, \boldsymbol{G}\boldsymbol{\alpha}) + \lambda \mathscr{R}(\boldsymbol{\alpha})$ is strictly convex. Then (3) admits a unique solution.

Proof: Assume that there exists at least two distinct solutions $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2 \in \mathscr{V}$. Then, by the strict convexity of $\mathscr{J}$, we have $\forall \theta \in [0,1]$: $\mathscr{J}(\theta \boldsymbol{\alpha}_1 + (1-\theta)\boldsymbol{\alpha}_2) < \theta \mathscr{J}(\boldsymbol{\alpha}_1) + (1-\theta)\mathscr{J}(\boldsymbol{\alpha}_2)$, and hence $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2$ do not minimise $\mathscr{H}$ which is a contradiction. □

Sufficient conditions for the strict convexity of $\mathscr{J}$ are: $F(\boldsymbol{y}, \cdot)$ is strictly convex and $\boldsymbol{G}$ is injective, or $\mathscr{R}$ is strictly convex. When $\mathscr{J}$ is not strictly convex we can still retain a weaker form of unicity:

Theorem: (Unicity of Predicted Measurements)

Assume that F and $\mathscr{R}$ are as in Slide 19 and that $F(\boldsymbol{y}, \cdot)$ is strictly convex. Then there exists a unique $\boldsymbol{y}^\star \in \mathbb{R}^L$ such that $\boldsymbol{G}\boldsymbol{\alpha}^\star = \boldsymbol{y}^\star, \quad \forall \boldsymbol{\alpha}^\star \in \mathscr{V} = \operatorname{argmin}_{\boldsymbol{\alpha} \in \mathbb{R}^N} \mathscr{J}(\boldsymbol{\alpha})$, i.e. every solution yield the same predicted measurements.

Proof (Unicity of Predicted Measurements)

$\exists \alpha_1, \alpha_2 \in V \quad G\alpha_1 = y_1 \quad G\alpha_2 = y_2 \quad y_1 \neq y_2$

$J(\alpha) = F(y_1, G\alpha) + \lambda R(\alpha) = F_y(G\alpha) + \lambda R(\alpha)$

$F_y(\cdot)$ is strictly cvx. Consider $\alpha_3 = \theta\alpha_1 + (1-\theta)\alpha_2 \in V$ convex

$J(\alpha_3) = F_y(\theta\alpha_1 + (1-\theta)\alpha_2) + \lambda \underbrace{R(\theta\alpha_1 + (1-\theta)\alpha_2)}_{\leq \theta R(\alpha_1) + (1-\theta) R(\alpha_2)}$

$< \theta F_y(\alpha_1) + (1-\theta) F_y(\alpha_2) + \lambda(\theta R(\alpha_1) + (1-\theta) R(\alpha_2))$

$\uparrow$ strictly cvx of F

$= \theta \underbrace{J(\alpha_1)}_{J^*} + (1-\theta) \underbrace{J(\alpha_2)}_{J^*} = J^*$ ← min value of J

Choosing the Cost Functional (Noiseless Case)

In a noiseless setup, one has *full trust* in the measurements. It is therefore natural to require that any solution of (3) be consistent with the data at hand, i.e. $\boldsymbol{y} = \boldsymbol{G}\boldsymbol{\alpha} \; \forall \boldsymbol{\alpha} \in \mathcal{V}$. This can be achieved by choosing the cost functional as $F(\boldsymbol{y}, \boldsymbol{G}\boldsymbol{\alpha}) = \iota(\boldsymbol{y} - \boldsymbol{G}\boldsymbol{\alpha})$, where $\iota : \mathbb{R}^L \to \{0, +\infty\}$ is the indicator function

$$\iota(\boldsymbol{z}) = \begin{cases} 0 & \text{if } \boldsymbol{z} = \boldsymbol{0}, \\ +\infty & \text{otherwise.} \end{cases}$$

Problem (3) becomes then a generalised interpolation problem:

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^N} \iota(\boldsymbol{y} - \boldsymbol{G}\boldsymbol{\alpha}) \quad + \quad \lambda \mathscr{R}(\boldsymbol{\alpha}) = \min_{\boldsymbol{\alpha} \in \mathbb{R}^N, \boldsymbol{y} = \boldsymbol{G}\boldsymbol{\alpha}} \mathscr{R}(\boldsymbol{\alpha}).$$

Penalised Problems with Strictly Convex Cost Functional are Interpolation Problems

Under the assumptions of the Theorem "Unicity of Predicted Measurements" on Slide 22 we have that:

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^N} F(\boldsymbol{y}, \boldsymbol{G}\boldsymbol{\alpha}) \quad + \quad \lambda \mathscr{R}(\boldsymbol{\alpha}) = \min_{\boldsymbol{\alpha} \in \mathbb{R}^N, \boldsymbol{y}^\star = \boldsymbol{G}\boldsymbol{\alpha}} \mathscr{R}(\boldsymbol{\alpha}),$$

for some (unknown) $\boldsymbol{y}^\star \in \mathbb{R}^L$. Hence, every penalised optimisation problem with strictly convex cost functional is equivalent to a generalised interpolation problem.

Choosing the Cost Functional (Noisy Case)

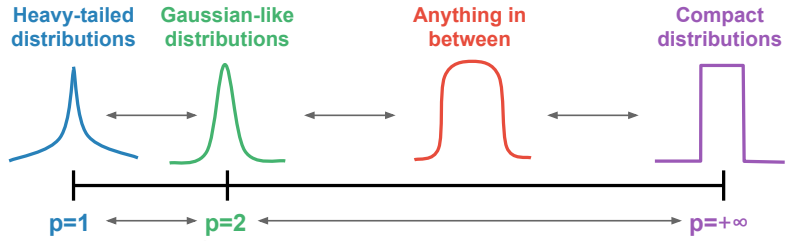
In a noisy setup, consistency is not desired anymore, as it almost always leads to overfitting the noisy data. One approach consists then in using the negative log-likelihood of the data $\boldsymbol{y}$ as a measure of discrepancy:

$$F(\boldsymbol{y}, \boldsymbol{G\alpha}) = -\ell(\boldsymbol{\alpha}|\boldsymbol{y}) = -\log p_{Y_1,\dots,Y_L}\left(y_1,\dots,y_L|\boldsymbol{\alpha}\right).$$

When the noise distribution is not fully known or the likelihood too complex, one can also use general ℓ_p cost functionals

$$F(\boldsymbol{y}, \boldsymbol{G\alpha}) = \|\boldsymbol{y} - \boldsymbol{G\alpha}\|_p^p = \sum_{i=1}^{L} \left| y_i - \sum_{n=1}^{N} G_{in}\alpha_n \right|^p,$$

where $p \in [1, +\infty]$ is typically chosen according to the tail behaviour of the noise distribution [2].



Example: Cost Functional for Gaussian Noise

Assume the following multivariate Gaussian noise model:

$$\boldsymbol{Y} = \boldsymbol{G\alpha} + \boldsymbol{N}, \quad \text{where} \quad \boldsymbol{N} \overset{\text{d}}{\sim} \mathcal{N}_L(\boldsymbol{0}, \boldsymbol{\Sigma}), \quad p_{\boldsymbol{N}}(\boldsymbol{y}) = \frac{1}{|\boldsymbol{\Sigma}|^{1/2}(2\pi)^{L/2}} \exp\left(-\frac{1}{2}\boldsymbol{y}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y}\right), \quad \boldsymbol{y} \in \mathbb{R}^L.$$

Then we have:

$$\begin{aligned} F(\boldsymbol{y}, \boldsymbol{G\alpha}) &= -\ell(\boldsymbol{\alpha}|\boldsymbol{y}) = -\log p_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{\alpha}) \\ &= -\log\left(\frac{1}{|\boldsymbol{\Sigma}|^{1/2}(2\pi)^{L/2}} \exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{G\alpha})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{G\alpha})\right)\right) \\ &= \frac{1}{2}\left\|\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{y}-\boldsymbol{G\alpha})\right\|_2^2 + \underbrace{\frac{1}{2}\log|\boldsymbol{\Sigma}| + \frac{L}{2}\log(2\pi)}_{\text{Independent of } \boldsymbol{\alpha}} \\ &\propto \left\|\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{y}-\boldsymbol{G\alpha})\right\|_2^2. \end{aligned}$$

This is the weighted least-squares functional. For white noise, we have $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{I}_L$ and the cost functional becomes proportional to $\|\boldsymbol{y}-\boldsymbol{G\alpha}\|_2^2$, which is the regular least-squares functional.

Example: Cost Functional for Laplacian Noise

Assume the following Laplacian white noise model:

$$Y_i = (\boldsymbol{G\alpha})_i + N_i, \quad \text{where} \quad N_i \overset{\text{i.i.d.}}{\sim} \text{Laplace}\,(0,\sigma), \quad p_N(x) = \frac{1}{2\sigma}\exp\left(-\frac{|x|}{\sigma}\right), \quad x \in \mathbb{R}.$$

Then we have:

$$\begin{aligned} F(\boldsymbol{y}, \boldsymbol{G\alpha}) &= -\ell(\boldsymbol{\alpha}|\boldsymbol{y}) = -\log p_{Y_1,\ldots,Y_L}\left(y_1,\ldots,y_L|\boldsymbol{\alpha}\right) \\ &= -\log\left(\frac{1}{(2\sigma)^L}\prod_{i=1}^{L}\exp\left(-\frac{|y_i - (\boldsymbol{G\alpha})_i|}{\sigma}\right)\right) \\ &= \frac{1}{\sigma}\sum_{i=1}^{L}|y_i - (\boldsymbol{G\alpha})_i| + \underbrace{L\log(2\sigma)}_{\text{Independent of } \boldsymbol{\alpha}} \\ &\propto \left\|\boldsymbol{y} - \boldsymbol{G\alpha}\right\|_1. \end{aligned}$$

This is the least absolute deviations functional. It is less affected by outliers than the least-squares functional. The weighted least absolute deviations functional can also be defined but cannot be interpreted as the negative log-likelihood of a multivariate Laplacian distribution.

Example: Cost Functional for Poisson Noise

Assume positive measurements $\boldsymbol{Y}: \Omega \to \mathbb{R}_+^L$ and the following Poisson noise model:

$$Y_i \overset{\text{ind}}{\sim} \text{Poisson}\left((\boldsymbol{G}\boldsymbol{\alpha})_i\right), \quad p_{Y_i}(k) = \frac{(\boldsymbol{G}\boldsymbol{\alpha})_i^k e^{-(\boldsymbol{G}\boldsymbol{\alpha})_i}}{k!}, \quad \forall k \in \mathbb{N}.$$

Then we have:

$$\begin{aligned} F(\boldsymbol{y}, \boldsymbol{G}\boldsymbol{\alpha}) &= -\ell(\boldsymbol{\alpha}|\boldsymbol{y}) = -\log p_{Y_1,\dots,Y_L}\left(y_1, \dots, y_L|\boldsymbol{\alpha}\right) \\ &= -\log\left(\prod_{i=1}^{L} \frac{(\boldsymbol{G}\boldsymbol{\alpha})_i^{y_i} e^{-(\boldsymbol{G}\boldsymbol{\alpha})_i}}{y_i!}\right) \\ &= \sum_{i=1}^{L} (\boldsymbol{G}\boldsymbol{\alpha})_i - y_i \log\left((\boldsymbol{G}\boldsymbol{\alpha})_i\right) + \underbrace{\log(y_i!)}_{\text{Independent of } \boldsymbol{\alpha}} \\ &\propto \sum_{i=1}^{L} (\boldsymbol{G}\boldsymbol{\alpha})_i - y_i \log\left((\boldsymbol{G}\boldsymbol{\alpha})_i\right) \end{aligned}$$

Example: Cost Functional for Poisson Noise (Continued)

$$\propto \sum_{i=1}^{L} (\boldsymbol{G\alpha})_i - y_i \log\big((\boldsymbol{G\alpha})_i\big) + \underbrace{y_i \log(y_i) - y_i}_{\text{Can add anything independent of } \boldsymbol{\alpha}}$$

$$= \sum_{i=1}^{L} y_i \log\left(\frac{y_i}{(\boldsymbol{G\alpha})_i}\right) + (\boldsymbol{G\alpha})_i - y_i$$

$$= D_{KL}(\boldsymbol{y} \| \boldsymbol{G\alpha}),$$

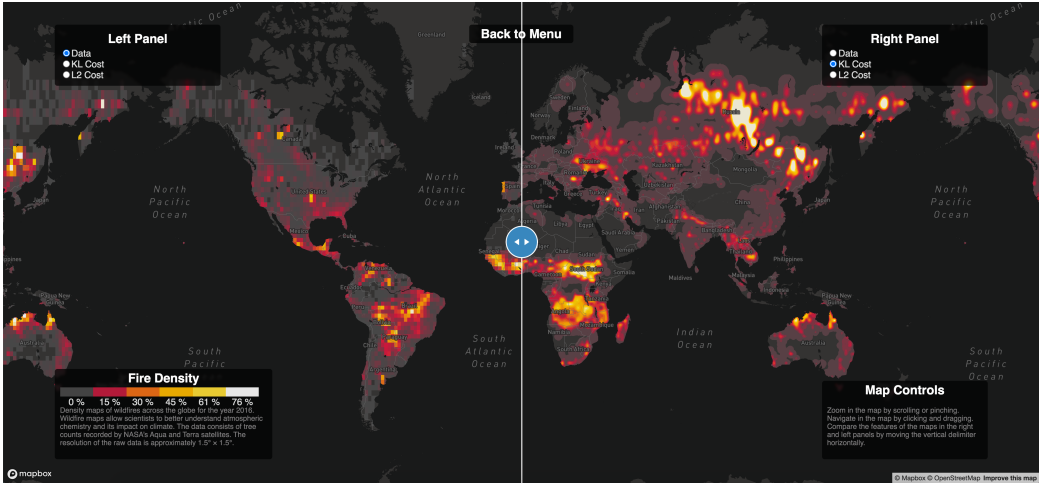
where

$$D_{KL}(\boldsymbol{y}\|\boldsymbol{z}) = \sum_{i=1}^{L} y_i \log\left(\frac{y_i}{z_i}\right) - y_i + z_i, \quad \forall \boldsymbol{y}, \boldsymbol{z} \in \mathbb{R}_+^L, \tag{5}$$

is the generalised Kullback-Leibler (KL) divergence [3] for discrete positive vectors which do not necessarily sum to one. In information theory, and in the case where $\mathbf{1}^T \boldsymbol{z} = \mathbf{1}^T \boldsymbol{y} = 1$,[5] the KL-divergence (5) can be interpreted as the relative entropy of $\boldsymbol{y}$ with respect to $\boldsymbol{z}$, i.e. the amount of information lost when using $\boldsymbol{z}$ to approximate $\boldsymbol{y}$. Note that the KL-divergence is not a distance (no symmetry/subadditivity).

[5] so that $\boldsymbol{z}$ and $\boldsymbol{y}$ can be interpreted as discrete probability distributions

Real-Life Example: Wild Fires



https://matthieumeo.github.io/fire_density.html

Choosing the Regularisation Functional

The regularisation functional is used to favour physically-admissible solutions with simple behaviours. It can be interpreted as implementing Occam's razor principle:

Occam's Razor Principle (*Lex parsimoniae*)

Occam's razor principle is a philosophical principle also known as the "law of briefness" or in Latin "lex parsimoniae". It was supposedly formulated by William of Ockham in the 14th century, who wrote in Latin *"Entia non sunt multiplicanda praeter necessitatem"*. In English, this translates to *"More things should not be used than are necessary"*.

In essence, this principle states that when two equally good explanations for a given phenomenon are available, one should always favour the simplest, i.e. the one that introduces the least explanatory variables.
What exactly is meant by "simple" solutions will depend on the specific application at hand.

(generalised) Tikhonov Regularisation

A common regularisation strategy consists in penalising the squared ℓ_2-norm of the solutions, i.e.

$$\mathscr{R}(\boldsymbol{\alpha}) = \|\boldsymbol{\alpha}\|_2^2, \qquad \boldsymbol{\alpha} \in \mathbb{R}^N. \tag{6}$$

This strategy is called Tikhonov regularisation and tends to favour smooth solutions. Different notions of smoothness can be achieved by introducing a positive semi-definite finite-difference differential operator $\boldsymbol{D} \in \mathbb{R}^{N \times N}$ in (6), yielding a generalised Tikhonov regularisation:

$$\mathscr{R}(\boldsymbol{\alpha}) = \|\boldsymbol{D}\boldsymbol{\alpha}\|_2^2, \qquad \boldsymbol{\alpha} \in \mathbb{R}^N. \tag{7}$$

The Tikhonov functional (6) is strictly convex, hence yielding unique solutions when used in conjunction with a convex cost functional. The generalised Tikhonov functional (7) is strictly convex if $\boldsymbol{D}$ is injective and simply convex otherwise. In the latter case, solutions to (3) exist if $\mathscr{N}(\boldsymbol{G}) \cap \mathscr{N}(\boldsymbol{D}) = \{\boldsymbol{0}\}$ and F is coercive but are in general non unique.[6]

[6] A sufficient condition for uniqueness is that F is proper strictly convex.

Form of Solutions with generalised Tikhonov Regularisation

Representer Theorem: (generalised Tikhonov Regularisation)

Assume that:

1. $\boldsymbol{G} \in \mathbb{R}^{L \times N}$ is surjective (i.e. full row rank), $\boldsymbol{D}$ is positive semi-definite and $\mathcal{N}(\boldsymbol{G}) \cap \mathcal{N}(\boldsymbol{D}) = \{\boldsymbol{0}\}$.
2. $F(\boldsymbol{y}, \cdot) : \mathbb{R}^L \to \mathbb{R}_+$ is proper strictly convex, coercive and lower semi-continuous for every $\boldsymbol{y} \in \mathbb{R}^L$.

Then the optimisation problem:

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^N} F(\boldsymbol{y}, \boldsymbol{G}\boldsymbol{\alpha}) \quad + \quad \lambda \|\boldsymbol{D}\boldsymbol{\alpha}\|_2^2$$

admits a unique solution which can be written as

$$\boldsymbol{\alpha}^\star = \left(\boldsymbol{D}^T\boldsymbol{D}\right)^\dagger \boldsymbol{G}^T\boldsymbol{\beta}^\star + \boldsymbol{\gamma}^\star,$$

for some $\boldsymbol{\beta}^\star \in \mathbb{R}^L$ and $\boldsymbol{\gamma}^\star \in \mathcal{N}(\boldsymbol{0})$.

When $\boldsymbol{D} = \boldsymbol{I}_N$ (standard Tikhonov regularisation) of $\boldsymbol{D}$ is invertible then the theorem holds for F proper convex and lwsc and we get $\boldsymbol{\alpha}^\star = \left(\boldsymbol{D}^T\boldsymbol{D}\right)^{-1} \boldsymbol{G}^T\boldsymbol{\beta}^\star$ this case is discussed in [4, Corollary 7]).

Ridge Estimate

strict cvx + proper + lwsc (coercive and strict)

$$\min_{\alpha \in \mathbb{R}^N} \underbrace{\frac{1}{2}\|y - G\alpha\|_2^2}_{J} + \underbrace{\frac{\lambda}{2}\|\alpha\|_2^2}_{D=I}. \quad \leftarrow \frac{\partial J}{\partial \alpha} = 0$$

Theorem : $\alpha^* = G^T \beta^*$? $\beta^* \in \mathbb{R}^L$.

$$\frac{\partial J}{\partial \alpha}(\alpha) = G^T G \alpha^* - G^T y + \lambda \alpha^* = 0$$

$$\Leftrightarrow (G^T G + \lambda I)\alpha^* = G^T y$$

$$\Leftrightarrow \alpha^* = \underbrace{(G^T G + \lambda I)^{-1} G^T}_{\in \mathcal{R}(G^T)} y$$

Stability of Ridge Estimate

$$y \mapsto \alpha^*(y) = (G^T G + \lambda I)^{-1} G^T y$$

bounded?

$$\|\alpha^*(y)\|_2 \leq \|(G^T G + \lambda I)^{-1}\|_2 \, \|G^T\|_2 \, \|y\|_2$$

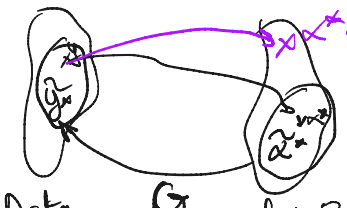
$$\|G^T\|_2 = \sqrt{\lambda_{max}(G^T G)}$$

$$= \frac{\sqrt{\lambda_{max}(G^T G)}}{\lambda_{min}(G^T G) + \lambda} \|y\|_2 \qquad \lambda > 0$$

$\lambda_{min}(G^T G) \geq 0$

$$\leq C \|y\|_2 \qquad C \ll +\infty$$

without λ



ℓ_1/TV Regularisation

A common regularisation strategy consists in penalising the ℓ_1-norm of the solutions, i.e.

$$\mathscr{R}(\boldsymbol{\alpha}) = \|\boldsymbol{\alpha}\|_1, \qquad \boldsymbol{\alpha} \in \mathbb{R}^N. \tag{8}$$

This strategy tends to favour sparse solutions with only a few non zero coefficients. Different notions of sparsity can be achieved by introducing a positive semi-definite finite-difference differential operator $\boldsymbol{D} \in \mathbb{R}^{N \times N}$ in (8), yielding a total variation (TV) regularisation:

$$\mathscr{R}(\boldsymbol{\alpha}) = \|\boldsymbol{D}\boldsymbol{\alpha}\|_1, \qquad \boldsymbol{\alpha} \in \mathbb{R}^N. \tag{9}$$

The ℓ_1 and TV functionals are convex. Solutions to (9) exist if $\mathscr{N}(\boldsymbol{G}) \cap \mathscr{N}(\boldsymbol{D}) = \{\boldsymbol{0}\}$ and F is coercive but are in general non unique.[7]

Examples:

- **LASSO/Penalised Basis Pursuit:** $\min_{\boldsymbol{\alpha} \in \mathbb{R}^N} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{G}\boldsymbol{\alpha}\|_2^2 \quad + \quad \lambda \|\boldsymbol{\alpha}\|_1.$
- **Generalised LASSO:** $\min_{\boldsymbol{\alpha} \in \mathbb{R}^N} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{G}\boldsymbol{\alpha}\|_2^2 \quad + \quad \lambda \|\boldsymbol{D}\boldsymbol{\alpha}\|_1.$

[7] Sufficient conditions for uniqueness are: $\boldsymbol{G}$ is injective and F is strictly convex.

Form of Solutions with TV Regularisation

Representer Theorem I: (TV Regularisation)

Assume that:

1. $\boldsymbol{G} \in \mathbb{R}^{L \times N}$ is invertible, $\boldsymbol{D}$ is positive semi-definite.
2. $F(\boldsymbol{y}, \cdot) : \mathbb{R}^L \to \mathbb{R}_+$ is proper strictly convex, coercive and lower semi-continuous for every $\boldsymbol{y} \in \mathbb{R}^L$.

Then the optimisation problem:

$$\mathcal{V} = \underset{\boldsymbol{\alpha} \in \mathbb{R}^N}{\arg\min}\, F(\boldsymbol{y}, \boldsymbol{G}\boldsymbol{\alpha}) \quad + \quad \lambda \|\boldsymbol{D}\boldsymbol{\alpha}\|_1$$

admits a unique solution of the form:

$$\boldsymbol{\alpha}^\star = \boldsymbol{D}^\dagger \boldsymbol{\beta}_K^\star + \boldsymbol{\gamma}^\star,$$

for some K-sparse vector $\boldsymbol{\beta}_K^\star \in \mathbb{R}^N$, $K \leq L$ and $\boldsymbol{\gamma}^\star \in \mathcal{N}(\boldsymbol{D})$.

When $\boldsymbol{D} = \boldsymbol{I}_N$ (ℓ_1 regularisation) or $\boldsymbol{D}$ is invertible then the theorem holds for F proper strictly convex and lwsc (no coercivity needed) and we have $\boldsymbol{\alpha}^\star = \boldsymbol{D}^{-1} \boldsymbol{\beta}_K^\star$.

Form of Solutions with TV Regularisation

Representer Theorem II: (TV Regularisation)

Assume that:

1. $\boldsymbol{G} \in \mathbb{R}^{L \times N}$ is surjective (i.e. full row rank), $\boldsymbol{D}$ is positive semi-definite and $\mathcal{N}(\boldsymbol{G}) \cap \mathcal{N}(\boldsymbol{D}) = \{\boldsymbol{0}\}$.
2. $F(\boldsymbol{y}, \cdot) : \mathbb{R}^L \to \mathbb{R}_+$ is proper convex, coercive and lower semi-continuous for every $\boldsymbol{y} \in \mathbb{R}^L$.

Then the solution set:

$$\mathcal{V} = \underset{\boldsymbol{\alpha} \in \mathbb{R}^N}{\arg\min}\, F(\boldsymbol{y}, \boldsymbol{G}\boldsymbol{\alpha}) \quad + \quad \lambda \|\boldsymbol{D}\boldsymbol{\alpha}\|_1$$

is non empty, compact and the convex-hull of extreme point solutions of the form:

$$\boldsymbol{\alpha}^\star = \boldsymbol{D}^\dagger \boldsymbol{\beta}_K^\star + \boldsymbol{\gamma}^\star,$$

for some K-sparse vector $\boldsymbol{\beta}_K^\star \in \mathbb{R}^N$, $K \leq L$ and $\boldsymbol{\gamma}^\star \in \mathcal{N}(\boldsymbol{D})$.

When $\boldsymbol{D} = \boldsymbol{I}_N$ (ℓ_1 regularisation) or $\boldsymbol{D}$ is invertible then the theorem holds for F proper convex and lwsc (no coercivity needed) and we have $\boldsymbol{\alpha}^\star = \boldsymbol{D}^{-1} \boldsymbol{\beta}_K^\star$ (this case is discussed in [4, Corollary 8]).

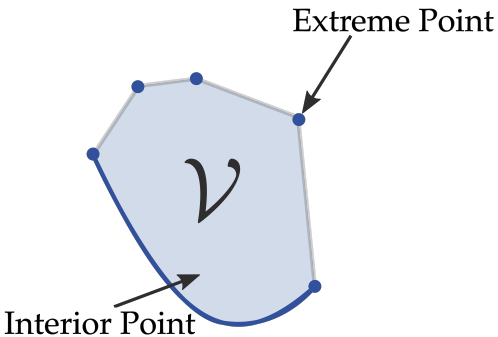
Solution Set is the Convex-Hull of Sparse Extreme Points

Definiton: (Extreme Point)

Let $\mathcal{V}$ be a convex set. An extreme point $v \in \mathcal{V}$ is a point such that

$$\nexists (w, v) \in \mathcal{V}^2, \theta \in]0,1[: \quad v = \theta w + (1-\theta) v.$$

In plain words, v is a point in $\mathcal{V}$ which does not lie in any open line segment joining two points of $\mathcal{V}$.



Example: Finite Difference Operator in $\mathbb{R}^{3\times 3}$

$$D = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix} \qquad D\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 - x_1 \\ x_3 - x_2 \end{pmatrix} \rightarrow \text{finite-difference operator}$$

$$V = \underset{x \in \mathbb{R}^3}{\operatorname{argmin}} \; F(y, Gx) + \lambda \|Dx\|_1$$

$$y \in \mathbb{R}^2, \quad G \in \mathbb{R}^{2\times 3}$$

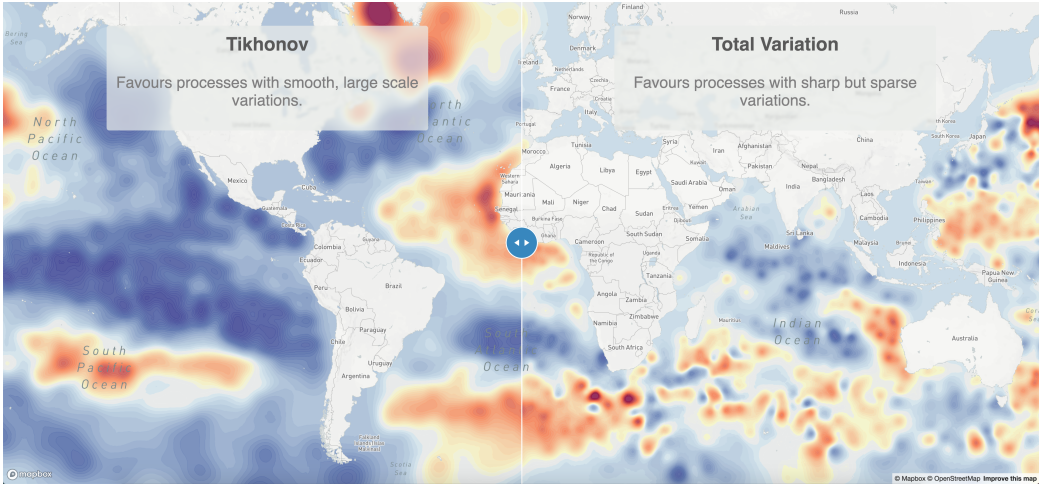
V cvx-hull of extreme points of the form:

$$x^* = D^{-1} B_k^* = \sum_{k=1}^{K} B_k D^{-1}_{:,n_k} \qquad K \leq 2$$

$K \leq 2$

$D^{-1} =$ 

Real-Life Example: Sea Surface Temperatures



https://matthieumeo.github.io/tikhonov_vs_tv_en.html

Maximum Entropy Regularisation

The maximum entropy regularisation strategy considers the following regularisation functional:

$$\mathscr{R}(\boldsymbol{\alpha}) = \sum_{n=1}^{N} \alpha_i \log(\alpha_i), \qquad \boldsymbol{\alpha} \in \mathbb{R}_+^N. \tag{10}$$

When $\mathbf{1}^T \boldsymbol{\alpha} = 1$ this si the negative Shannon entropy [5, 6] of $\boldsymbol{\alpha}$, a mathematical generalisation of entropy as introduced by Boltzmann in thermodynamics. It favours positive, featureless solutions which: smooth vectors indeed carry much less information than vectors with sharp, localised features, and hence have higher entropy.

This regularisation can be generalised by considering the negative relative entropy w.r.t. a reference discrete distribution $\boldsymbol{\eta} \in \mathbb{R}_+^N$:

$$\mathscr{R}(\boldsymbol{\alpha}) = \sum_{n=1}^{N} \alpha_i \log\left(\frac{\alpha_i}{\eta_i}\right), \qquad \boldsymbol{\alpha} \in \mathbb{R}_+^N. \tag{11}$$

The functional (11) favours solutions with similar features as the reference distribution $\boldsymbol{\eta}$. Both functionals (10) and (11) are strictly convex and coercive, hence yielding unique solutions when they exist.

Form of Solutions with Maximum Entropy Interpolation

Representer Theorem: (Maximum Entropy Interpolation)

Consider the generalised interpolation problem:

$$\min_{\substack{\boldsymbol{\alpha}\in\mathbb{R}^N \\ \boldsymbol{y}=\boldsymbol{G}\boldsymbol{\alpha}}} \sum_{n=1}^{N} \alpha_i \log(\alpha_i), \tag{12}$$

for some $\boldsymbol{y} \in \mathbb{R}^L$. If (12) admits a solution, the latter is unique and can be written as [7]:

$$\boldsymbol{\alpha}^\star = \gamma \exp\left(\boldsymbol{G}^T \boldsymbol{\beta}^\star\right),$$

for some $\gamma > 0$ and $\boldsymbol{\beta}^\star \in \mathbb{R}^L$.

The nonlinear exponential map kills low intensity features and boosts prominent ones.

Nonnegativity Regularisation

The nonnegativity regularisation strategy considers the following regularisation functional:

$$\mathscr{R}(\boldsymbol{\alpha}) = \begin{cases} 0 & \text{if } \boldsymbol{\alpha} \in \mathbb{R}_+^N, \\ +\infty & \text{otherwise.} \end{cases} \tag{13}$$

It constrains the solutions to be positive. This functional is convex and non coercive.

The functional (13) is sometimes replaced by the log-barrier functional (convex and non coercive):

$$\mathscr{R}(\boldsymbol{\alpha}) = -\sum_{n=1}^{N} \log(\alpha_i), \qquad \boldsymbol{\alpha} \in \mathbb{R}_+^N,$$

which also promotes positive solutions.

Nonnegative Least-Squares (NNLS):

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}_+^N} \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{G}\boldsymbol{\alpha}\|_2^2.$$

Form of Solutions with Nonnegativity Constraints

Representer Theorem: (Nonnegativity Constraints)

Assume that:

1. $\boldsymbol{G} \in \mathbb{R}^{L \times N}$ is injective.
2. $F(\boldsymbol{y}, \cdot) : \mathbb{R}^L \to \mathbb{R}_+$ is proper strictly convex, coercive and lower semi-continuous for every $\boldsymbol{y} \in \mathbb{R}^L$.

Then the optimisation problem:

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}_+^N} F(\boldsymbol{y}, \boldsymbol{G}\boldsymbol{\alpha})$$

admits a unique L-sparse solution.

The proof to this Theorem follows from [8, Proposition 4.1].

Bayesian Interpretation

In certain cases, the penalised optimisation problem (3) can be interpreted as a maximum a posteriori (MAP) problem. Adopting a Bayesian view, assume for example a Gaussian a priori distribution for α and a Gaussian likelihood function (i.e. a Gaussian white noise model):

$$p(\boldsymbol{\alpha}) \propto \exp\left(-\frac{1}{2\xi^2}\|\boldsymbol{\alpha}\|_2^2\right), \qquad p(\boldsymbol{y}|\boldsymbol{\alpha}) \propto \exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{y}-\boldsymbol{G}\boldsymbol{\alpha}\|_2^2\right). \tag{14}$$

From Baye's theorem, the posterior distribution of $\boldsymbol{\alpha}$ knowing the data $\boldsymbol{y}$ is then given by

$$p(\boldsymbol{\alpha}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\alpha})p(\boldsymbol{\alpha})}{\int_{\mathbb{R}^N} p(\boldsymbol{y}|\boldsymbol{\alpha})p(\boldsymbol{\alpha})d\boldsymbol{\alpha}}. \tag{15}$$

A maximum a posteriori (MAP) estimate is then defined as

$$\boldsymbol{\alpha}^\star_{MAP} \in \underset{\boldsymbol{\alpha}\in\mathbb{R}^N}{\arg\max}\, p(\boldsymbol{\alpha}|\boldsymbol{y}) = \underset{\boldsymbol{\alpha}\in\mathbb{R}^N}{\arg\max}\, p(\boldsymbol{y}|\boldsymbol{\alpha})p(\boldsymbol{\alpha}) = \underset{\boldsymbol{\alpha}\in\mathbb{R}^N}{\arg\max}\, L(\boldsymbol{\alpha}|\boldsymbol{y})p(\boldsymbol{\alpha}) = \underset{\boldsymbol{\alpha}\in\mathbb{R}^N}{\arg\min} -\ell(\boldsymbol{\alpha}|\boldsymbol{y}) - \log(p(\boldsymbol{\alpha})).$$

For the prior distribution and likelihood assumed in (14) this yields:

$$\boldsymbol{\alpha}^\star_{MAP} \in \underset{}{\arg\min}\, \frac{1}{2\sigma^2}\|\boldsymbol{y}-\boldsymbol{G}\boldsymbol{\alpha}\|_2^2 + \frac{1}{2\xi^2}\|\boldsymbol{\alpha}\|_2^2 = \underset{}{\arg\min}\, \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{G}\boldsymbol{\alpha}\|_2^2 + \frac{\lambda}{2}\|\boldsymbol{\alpha}\|_2^2, \quad \text{with} \quad \lambda = \frac{\sigma^2}{\xi^2}. \tag{16}$$

Bayesian Interpretation (continued)

We recognise in (16) the Ridge estimate (see Slide 34), obtained when choosing a least-squares cost functional and a Tikhonov penalty in (3). Notice moreover that the regularisation parameter λ is equal to the ratio of the likelihood and prior variances.

Note that the prior distribution can be improper (i.e. *not summable*) as long as $\int_{\mathbb{R}^N} p(\boldsymbol{y}|\boldsymbol{\alpha})p(\boldsymbol{\alpha})d\boldsymbol{\alpha} < +\infty$ so that (15) is still well-defined.

This allows us to extend the previous analysis to many classical optimisation problems:

Example: (Penalised Optimisation Pbs as MAP)

- **Weighted Least-Squares with Generalised Tikhonov** ($\min_{\boldsymbol{\alpha}\in\mathbb{R}^N} \frac{1}{2}\|\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{y}-\boldsymbol{G}\boldsymbol{\alpha})\|_2^2 + \frac{1}{2}\|\boldsymbol{D}\boldsymbol{\alpha}\|_2^2$)**:**
  $p(\boldsymbol{\alpha}) \propto \exp\left(-\frac{1}{2}\boldsymbol{\alpha}^T\boldsymbol{D}^T\boldsymbol{D}\boldsymbol{\alpha}\right), \qquad p(\boldsymbol{y}|\boldsymbol{\alpha}) \propto \exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{G}\boldsymbol{\alpha})^T\Sigma^{-1}(\boldsymbol{y}-\boldsymbol{G}\boldsymbol{\alpha})\right), \quad \lambda = 1.$
- **Least Absolute Deviations with Tikhonov** ($\min_{\boldsymbol{\alpha}\in\mathbb{R}^N} \|\boldsymbol{y}-\boldsymbol{G}\boldsymbol{\alpha}\|_1 + \frac{\lambda}{2}\|\boldsymbol{\alpha}\|_2^2$)**:**
  $p(\boldsymbol{\alpha}) \propto \exp\left(-\frac{1}{2\xi^2}\|\boldsymbol{\alpha}\|_2^2\right), \qquad p(\boldsymbol{y}|\boldsymbol{\alpha}) \propto \exp\left(-\frac{1}{\sigma}\|\boldsymbol{y}-\boldsymbol{G}\boldsymbol{\alpha}\|_1\right), \quad \lambda = \frac{\sigma}{\xi^2}.$

Bayesian Interpretation (continued)

Example: (Penalised Optimisation Pbs as MAP)

- **LASSO** ($\min_{\boldsymbol{\alpha}\in\mathbb{R}^N} \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{G\alpha}\|_2^2+\lambda\|\boldsymbol{\alpha}\|_1$)**:**
  $p(\boldsymbol{\alpha}) \propto \exp\left(-\frac{1}{\xi}\|\boldsymbol{\alpha}\|_1\right), \qquad p(\boldsymbol{y}|\boldsymbol{\alpha}) \propto \exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{y}-\boldsymbol{G\alpha}\|_2^2\right), \quad \lambda=\frac{\sigma^2}{\xi}.$
- **Generalised LASSO** ($\min_{\boldsymbol{\alpha}\in\mathbb{R}^N} \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{G\alpha}\|_2^2+\lambda\|\boldsymbol{D\alpha}\|_1$)**:**
  $p(\boldsymbol{\alpha}) \propto \exp\left(-\|\boldsymbol{D\alpha}\|_1\right), \qquad p(\boldsymbol{y}|\boldsymbol{\alpha}) \propto \exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{y}-\boldsymbol{G\alpha}\|_2^2\right), \quad \lambda=\sigma^2.$
- **Least-squares with Maximum Entropy** ($\min_{\boldsymbol{\alpha}\in\mathbb{R}^N} \frac{1}{2}\|(\boldsymbol{y}-\boldsymbol{G\alpha})\|_2^2+\lambda\sum_{n=1}^N \alpha_n\log(\alpha_n)$)**:**
  $p(\boldsymbol{\alpha}) \propto \exp\left(-\sum_{n=1}^N \alpha_n\log(\alpha_n)\right), \qquad p(\boldsymbol{y}|\boldsymbol{\alpha}) \propto \exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{y}-\boldsymbol{G\alpha})\|_2^2\right), \quad \lambda=\sigma^2.$
- **Nonnegative Least-Squares** ($\min_{\boldsymbol{\alpha}\in\mathbb{R}_+^N} \|\boldsymbol{y}-\boldsymbol{G\alpha}\|_2$)**:**
  $p(\boldsymbol{\alpha}) \propto \exp\left(-\iota_{\mathbb{R}_+^N}(\boldsymbol{\alpha})\right), \qquad p(\boldsymbol{y}|\boldsymbol{\alpha}) \propto \exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{y}-\boldsymbol{G\alpha}\|_2^2\right).$
- **KL-Divergence with Tikhonov** ($\min_{\boldsymbol{\alpha}\in\mathbb{R}^N} D_{KL}(\boldsymbol{y}||\boldsymbol{G\alpha})+\frac{\lambda}{2}\|\boldsymbol{\alpha}\|_2^2$)**:**
  $p(\boldsymbol{\alpha}) \propto \exp(-\frac{1}{2\xi^2}\|\boldsymbol{\alpha}\|_2^2), \qquad p(\boldsymbol{y}|\boldsymbol{\alpha}) \propto \exp\left(-D_{KL}(\boldsymbol{y}||\boldsymbol{G\alpha})\right), \quad \lambda=\frac{1}{\xi^2}.$

References I

[1] Harshit Gupta, Julien Fageot, and Michael Unser.
Continuous-domain solutions of linear inverse problems with tikhonov versus generalized tv regularization.
IEEE Transactions on Signal Processing, 66(17):4670–4684, 2018.

[2] John R Rice and John S White.
Norms for smoothing and estimation.
SIAM review, 6(3):243–256, 1964.

[3] Mario Bertero, Patrizia Boccacci, Giorgio Talenti, Riccardo Zanella, and Luca Zanni.
A discrepancy principle for poisson data.
Inverse problems, 26(10):105004, 2010.

[4] Michael Unser, Julien Fageot, and Harshit Gupta.
Representer theorems for sparsity-promoting l1 regularization.
IEEE Transactions on Information Theory, 62(9):5167–5180, 2016.

[5] Ramesh Narayan and Rajaram Nityananda.
Maximum entropy image restoration in astronomy.
Annual review of astronomy and astrophysics, 24(1):127–170, 1986.

References II

[6] David L Donoho, Iain M Johnstone, Jeffrey C Hoch, and Alan S Stern.
Maximum entropy and the nearly black object.
Journal of the Royal Statistical Society: Series B (Methodological), 54(1):41–67, 1992.

[7] Lawrence D Brown.
Fundamentals of statistical exponential families: with applications in statistical decision theory.
Ims, 1986.

[8] Claire Boyer, Antonin Chambolle, Yohann De Castro, Vincent Duval, Frédéric De Gournay, and Pierre Weiss.
On representer theorems and convex regularization.
SIAM Journal on Optimization, 29(2):1260–1281, 2019.