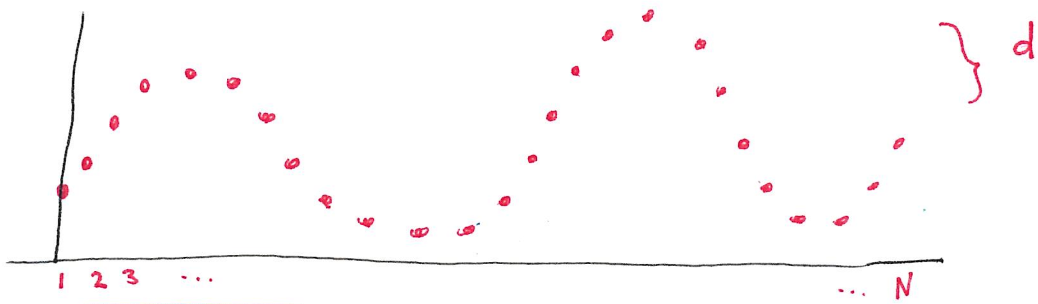
PCMI Lecture #4

Regularization

we start by continuing to explore the signal reconstruction from data problem.

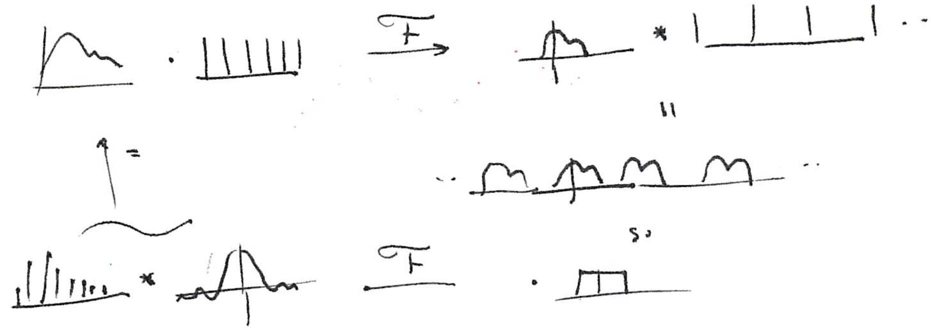


$\boxed{\text{Problem: find signal off of the samples.}}$

Recalling the fact that there is a codimension N family of interpolations of this data, we see that our job is highly ill-posed, inspite of the fact that our eyes and preconceived ideas tell us the curve is wavelike.

Noiseless Case: interpolation

(1) Band limited → shannon reconstruction. gives a perfect reconstruction



another way to look at it is simply the sampled sines / cosines span $\mathbb{R}^N$...

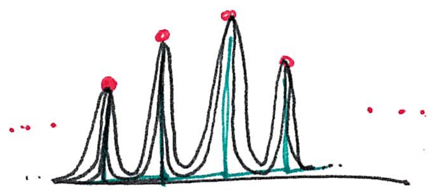
(2) min $\|\nabla f\|_2$ → piecewise linear interpolation

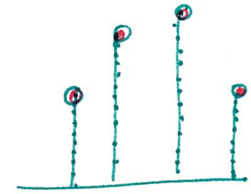
Simple calculation shows the minimizer is piecewise linear

(3) $\min \begin{matrix} \|K\|_{sup} \\ \|K\|_\infty \\ \|K\|_2 \end{matrix}$ $\longrightarrow$ more complicated but easy to describe

minimize the rate of turn of the tangent vector. Easiest thing to do is ~~think~~ express it as a arc length parameterized path that must pass through samples. ~~candidate~~ ~~constraint~~ ~~candidate~~ ~~solution~~ $\rightarrow$ splines.

(4) $\min \|f\|_2$ $\longrightarrow$ no nontrivial minimizer (if we want $f \in C[0,L]$ then $\exists$ no minimum)

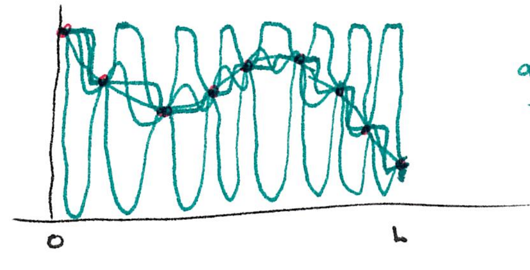




$\simeq 0$ in L^2

same for $\|f\|_1$

(5) $\min \|f\|_\infty$ $\longrightarrow$ lots of minimizers $\rightarrow$ no help at all ☺



all curves minimize the objective ~ regularized energy!

(6) $\min \left\| \frac{d^2 f}{dx^2} \right\|_p$ very similar to $\|K\|_2$

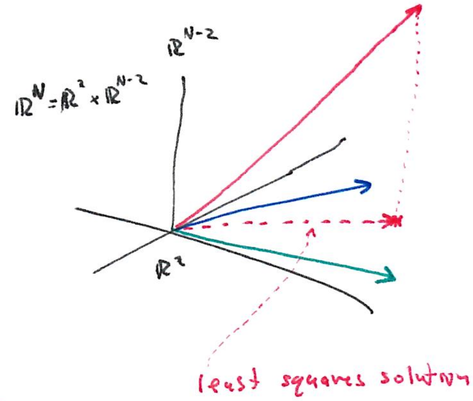
Key point: choice of regularization $\longrightarrow$ has a large influence on form of interpolation, interpolating minimizer.

Noisy Case: interpolation with errors.

(1) $\min_{\alpha, \beta} \|d - \alpha \vec{x} - \beta \cdot \vec{1}\|_2 \longrightarrow$ least squares

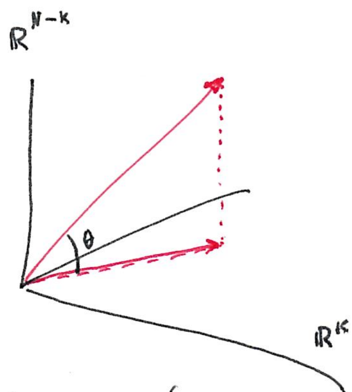
least squares is solved by projection:

$$d = \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_N \end{bmatrix} \quad x = \begin{bmatrix} 0 \\ 1 \\ 2 \\ \vdots \\ N-1 \end{bmatrix} \quad \vec{1} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$



(this is really max likelihood)

② least squares again



$\cos\theta$ = correlation

$$d = \alpha_1 f_1 + \alpha_2 f_2 + \cdots + \alpha_K f_K$$

f_i are somehow nice, regular, simpler, defined on all of $[0, L]$

③

$$D \equiv \begin{array}{c c} & M \\ N & \begin{bmatrix} | & | & & | \\ d^1 & d^2 & \cdots & d^M \\ | & | & & | \end{bmatrix} \end{array}$$

$M >> N$

$\Downarrow$

compute SVD

$$D = \begin{array}{c c} & N \\ N & \begin{bmatrix} | & | & & | \\ l_1 & l_2 & \cdots & l_N \\ | & | & & | \end{bmatrix} \end{array} \begin{bmatrix} \sigma_1 & & & \vdots & \\ & \sigma_2 & 0 & \vdots & \\ & & \ddots & \vdots & 0 \\ 0 & & \sigma_N & \vdots & \end{bmatrix} \begin{array}{c c} & M \\ M & \begin{bmatrix} \text{---} & r_1 & \text{---} \\ \text{---} & r_2 & \text{---} \\ & \vdots & \\ \text{---} & r_M & \text{---} \end{bmatrix} \end{array}$$

$$= L \Sigma R^T$$

Now compute interpolations for $\ell_1, \ell_2, \ldots \ell_N \Rightarrow f_1, f_2, \ldots, f_N$

This is done once. $F \equiv [f_1 f_2 \cdots f_N]$

Now ~~[illegible]~~ compute

$$f_d \equiv F L^T d$$ This interpolates the data

$$f_d^\varepsilon \equiv F L_\varepsilon^T d$$ where L_ε are the columns corresponding to those singular values $> \varepsilon$. This interpolates the data with errors, useful for noisy data.

This is very similar to the first, noiseless case of Shannon reconstruction.

(4)

$$\min_{d^*, f} \left\{ \| d - d^* \|_p + \alpha \begin{matrix} \|\nabla f\|_2 \\ \text{or} \\ \| K \|_2 \end{matrix} \right\}$$ where f interpolates d^*

comments on these examples

- underlying the notion & practice of regularization is the notion that things are simple, sparse, low dimensional.
- This problem is similar to the Dirichlet problem and it's generalization to higher dimensions includes such problems (of course we can define the problem as $\Delta f = 0$ and $f(a) = f_1$ $f(b) = f_2$ but it is so simple we usually move onto $\mathbb{R}^2$ & $\mathbb{R}^3$ rather quickly)
- There is a great deal of content here that arises again in higher dimensional contexts ($\mathbb{R}^2, \mathbb{R}^3$) with image problems. Examples: (1) Inpainting and (2) super-resolution
- regularization = prior knowledge

  Again, without further assumptions, such as those implicit in the choice of regularization, data is just a sequence of numbers.

Regularization: more generally

noiseless case:

$$\underset{\text{data}}{d} = \underset{\text{measurement operator}}{M}(\underset{\text{state we would like to know}}{u})$$

(or image or signal or etc)

we are given d and want u "=" $M^{-1}(d)$

The problem is that M need not be one-to-one or even if it is one-to-one it need not have a continuous inverse.

Typical approach:

$$u^* = \underset{v}{\text{argmin}} \; \|d - M(v)\| + \underset{\text{regularization energy}}{R(v)}$$

Idea with R is that the smaller R is, the more regular u is

$$\cancel{\left(u^* = \underset{v}{\text{argmin}} \; \|d - M(v)\| + R(v)\right)}$$

Alternatively:

$$u^* = M^{-1}(d) \cap \{\text{low dimensional set}\}$$

or

$$\{v \in \text{low dimensional set} \mid \rho(v, M^{-1}(d)) \text{ is minimized}\}$$

The low dimensional set has low complexity, high regularity ... in fact one might take as a definition of regularity Membership in a <u>predetermined</u> low dimensional set/manifold/subspace.

noisy case:

$$d = M(u) + \eta \qquad \text{(we will stick to additive noise)}$$

Now we might consider

$$u^* = \underset{v}{\text{argmin}} \; R(v) \text{ subject to } \|d - M(v)\| = \underset{\text{noise level}}{\sigma}$$

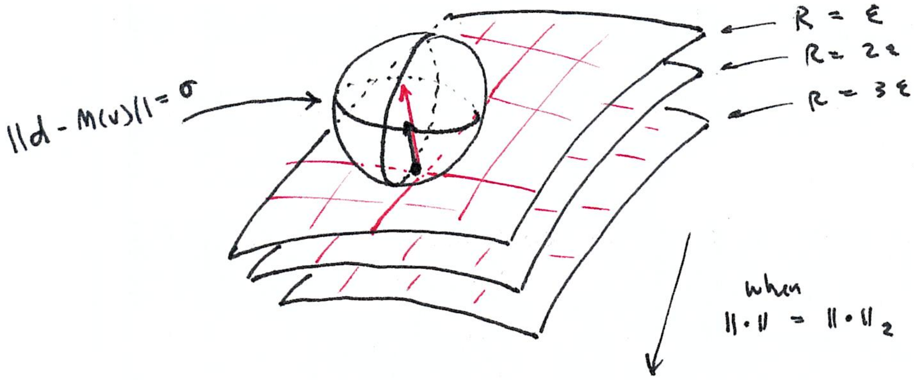
or

subject to $\|d - M(v)\| \leq \sigma$

This leads to:

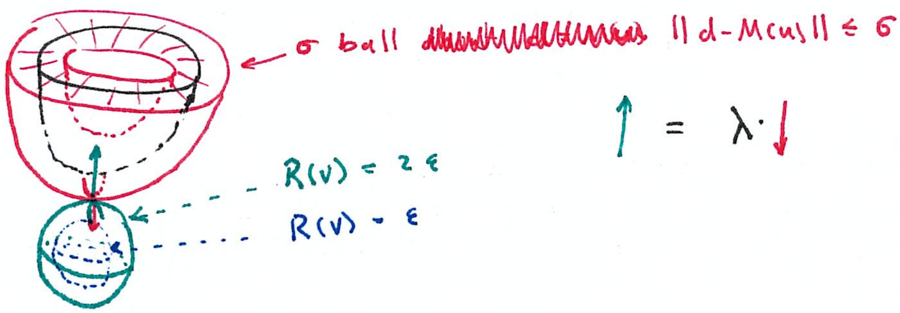
$$u^* = \underset{v}{\operatorname{argmin}} \; R(v) + \lambda \| d - M(v) \|$$

choosing λ correctly gives constraint.



when $\| \cdot \| = \| \cdot \|_2$

$$DR(v) + \lambda (d - M(v)) = 0$$



$\boxed{\text{Some Popular (and some less well known) } R(u)}$

<u>For images</u>

$\int |\nabla u| \, dx$ Total variation seminorm

$\int |\nabla u|^2 \, dx$ H^1 seminorm

$\int |\nabla u|^{p(|\nabla u|)} \, dx$ Blomgren et al., Bollt et. al.

$\nabla \cdot \left(\frac{g'(|\nabla u|)}{|\nabla u|} \nabla u \right)$ $\int g(|\nabla u|) \, dx$ various methods use more general g's including anisotropic BV, perona-malik like methods

Linear Problems

$$Ax = b \qquad {}_m\big[\overset{n}{A}\big]$$

$\boxed{m < n}$ null space (N) has dimension $n-m$ (Full Rank assumption)

If x_b is any solution of $Ax = b$ then all $x \in x_b + N$ is also a solution. We call $x_b + N$ the <u>solution space</u>.

$\boxed{m = n}$ null space is just $\{0\}$ (Full Rank assumption again)

$$x = A^{-1} b$$

$\boxed{m > n}$ null space is again $\{0\}$ (Full Rank assumption)

Typically $Ax \neq b$ $\forall$ possible x.
We can seek $\min_x \|Ax - b\|_2$ for example... this is least squares.

<u>Fundamental Tool, Mentioned above:</u> SVD

$$A = \underbrace{\begin{bmatrix} | & | & & | \\ u_1 & u_2 & \cdots & u_m \\ | & | & & | \end{bmatrix}}_{\text{orthogonal}} \overbrace{\left[\begin{array}{ccc|c} \sigma_1 & & 0 & \\ & \ddots & & 0 \\ 0 & & \sigma_m & \end{array}\right]}^{\text{digonal, } \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_m \geq 0} \underbrace{\begin{bmatrix} — & v_1 & — \\ & \vdots & \\ — & v_n & — \end{bmatrix}}_{\text{orthogonal}}$$

(illustrated in the case $m < n$)

If σ_K is small enough, we might choose to consider all σ_i $i \geq K$ to be zero. computing the pseudo-inverse then projects onto the first $K-1$ right singular vectors

$$A \rightarrow A^{\varepsilon} = \begin{bmatrix} L \end{bmatrix} \begin{bmatrix} \sigma & & & \vdots & \\ & \sigma & & \vdots & \\ & & \sigma & \vdots & 0 \\ & & & 0 & \\ & & & & 0 \end{bmatrix} \begin{bmatrix} R \end{bmatrix}$$

σ below this line $< \varepsilon$

$$(A^{\varepsilon})^{-1} d = R^T (\Sigma^{\varepsilon})^{-1} L^T d$$

$$(\Sigma^{\varepsilon})^{-1} = \begin{bmatrix} \frac{1}{\sigma} \end{bmatrix} \begin{bmatrix} 1/\sigma & & & 0 \\ & 1/\sigma & & \\ & & 1/\sigma & \\ & 0 & & 0 \; 0 \ldots \\ & & 0 & \end{bmatrix}$$

This regularizes the inverse.

note that

$$r_K \longrightarrow \sigma_K l_K$$

now suppose $d \cdot l_K = \alpha$ then on inversion we will have a component in the r_K direction of magnitude $\frac{\alpha}{\sigma_K}$ ... i.e. $X = \sum_{i=1}^{n} c_i r_i$

where $c_K = \frac{\alpha}{\sigma_K}$ so if α is $O(1)$ and $\sigma_K << 1$ then $c_K >> 1$. Now is this is a noise direction, we have just magnified noise.

Thus truncating the singular values is a regularization that is important for noisy data!

Regularization and Regularity

- strictly speaking, regularization is simply doing something to turn an ill-posed problem into a well posed one.
- regularity is how nice some measure, set, or function is

They are closely related: one usually picks $R(u)$ or your lower dimensional set so that all minimizers or elements are regular.

Other issues

Replacing $u = M^{-1}(d) \longrightarrow u = \arg\min_{v} F(v)$

Can give you unique u with continuous dependence on d BUT the real question is what is the sensitivity of u wrt d? How shallow is the minimum at u? How fast does the minimal region change as we change d?

When $u = \arg\min_{v} F(v)$ is not unique how wild is the nonuniqueness.

(Example: when $F(\cdot)$ is convex, the minimizers set is convex)

Final example of priors (= regularization)

Tomography with different priors