

Synthesis and Analysis for structured regression

Sara van de Geer

June 20, 2019

Joint work with
Francesco Ortelli
(errors are mine)







A photograph of a street in Amsterdam. In the foreground, a white bridge railing runs across the frame, with several bicycles parked along it. Behind the bridge is a row of multi-story brick buildings with many windows. To the left, a church tower with a blue dome and a golden cross is visible. The sky is overcast.

Happy Birthday
Aad!

Analysis and synthesis

Let $Y \in \mathbb{R}^n$ be the response vector.

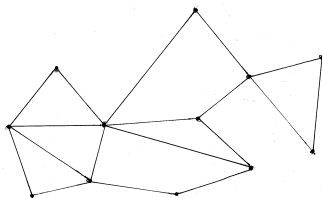
synthesis problem: with $X \in \mathbb{R}^{n \times p}$ given design matrix

$$\text{minimize over } b \in \mathbb{R}^p : \|Y - Xb\|_2^2/n + 2\lambda \|b\|_1.$$

analysis problem: with $D \in \mathbb{R}^{m \times n}$ given analysis matrix

$$\text{minimize over } f \in \mathbb{R}^n : \|Y - f\|_2^2/n + 2\lambda \|Df\|_1.$$

Example: $D :=$ incidence matrix of a graph.





Synthesis problem: $X = \text{dictionary}$

p variables of which perhaps only s are relevant
... and we do not know which ones.

Typically $p \gg n$.

Analysis problem:

m potential restrictions of which perhaps only s are violated
... and we do not know which ones.

Possibly $m \geq n$.

An example: total variation of higher order differences

A preview

Define the differences $(\Delta f)_j = f_j - f_{j-1}$ and

$$\Delta^0 f := f, \Delta^l f := \Delta \Delta^{l-1} f, l \in \mathbb{N}$$

Fix $k \in \mathbb{N}$ and let

$$Df := \Delta^k f.$$

Let $S := \{t_1, \dots, t_s\}$, $t_1 < \dots < t_s$, $t_0 := k$, $t_{s+1} := n$.

Let $d_\infty := \max_{1 \leq j \leq s+1} (t_j - t_{j-1})$

analysis:

$$\hat{f} = \arg \min_{f \in \mathbb{R}^n} \left\{ \|Y - f\|_2^2 / n + 2\lambda \|Df\|_1 \right\}$$

with

$$Df := \Delta^k f.$$

Suppose

$$Y = f^0 + \epsilon, \quad \epsilon \sim \mathcal{N}(0, I).$$

Theorem For all $u > 0$, $v > 0$ and for

$$\lambda \geq c_k \lambda_S(u) d_\infty^{\frac{2k-1}{2}} / \sqrt{n}$$

where

$$\lambda_S(u) = \sqrt{\frac{2 \log(2(n - (s + k))) + 2u}{n}}$$

we have with probability at least $1 - \exp[-u] - \exp[-v]$

$$\begin{aligned} \|\hat{f} - f^0\|_2^2/n &\leq \|f - f^0\|_2^2/n \\ &+ \left(\sqrt{\frac{k(s+1)}{n}} + \sqrt{\frac{2v}{n}} + \lambda \Gamma(S, w_{-s}) \right)^2 \end{aligned}$$

with

$$\underbrace{\Gamma^2(S, w_{-s})}_{\text{"effective sparsity"}} \leq C_k^2 \sum_{j=1}^{s+1} \frac{n \log(t_j - t_{j-1})}{(t_j - t_{j-1})^{2k-1}}.$$

Remark The “effective sparsity” is a version of the inverse of the restricted eigenvalue. This lecture is about how to upper-bound this quantity (when the design is highly correlated but structured).

Theorem For all $u > 0$, $v > 0$ and for

$$\lambda \geq c_k \lambda_S(u) d_\infty^{\frac{2k-1}{2}} / \sqrt{n}$$

where

$$\lambda_S(u) = \sqrt{\frac{2 \log(2(n - (s + k))) + 2u}{n}}$$

we have with probability at least $1 - \exp[-u] - \exp[-v]$

$$\begin{aligned} \|\hat{f} - f^0\|_2^2/n &\leq \|f - f^0\|_2^2/n \\ &+ \left(\sqrt{\frac{k(s+1)}{n}} + \sqrt{\frac{2v}{n}} + \lambda \Gamma(S, w_{-s}) \right)^2 \end{aligned}$$

with

$$\underbrace{\Gamma^2(S, w_{-s})}_{\text{"effective sparsity"}} \leq C_k^2 \sum_{j=1}^{s+1} \frac{n \log(t_j - t_{j-1})}{(t_j - t_{j-1})^{2k-1}}.$$

Remark The “effective sparsity” is a version of the inverse of the restricted eigenvalue. This lecture is about how to upper-bound this quantity (when the design is highly correlated but structured).

Corollary *If the jumps of $\Delta^{k-1} f^0$ are “roughly” equidistant the rate of convergence is*

$$\|\hat{f} - f^0\|_2^2/n = \mathcal{O}_{\mathbb{P}}\left(\frac{(s+1)\log^2 n}{n}\right)$$

provided

$$\lambda \asymp \sqrt{\frac{\log n}{n}} \left(\frac{n}{s+1}\right)^{\frac{2k-1}{2}} / \sqrt{n}.$$

Contents

- Noiseless analysis
- Explaining the restricted eigenvalue issue
 $\leadsto$ “new” compatibility constant (and effective sparsity)
- Noisy analysis

Notation

For $v \in \mathbb{R}^n$

$$\begin{aligned}\|v\|_{Q_n}^2 &:= \sum_{i=1}^n v_i^2 / n \\ &= \|v\|_2^2 / n.\end{aligned}$$

$S \subset [1, m]$ potentially active variables

for $b \in \mathbb{R}^m$ let $b_S := \{b_j\}_{j \in S}$, $b_{-S} := \{b_j\}_{j \notin S}$

Direct analysis of the analysis problem without noise

Let

$$f^* := \arg \min_{f \in \mathbb{R}^n} \left\{ \|f - f^0\|_{Q_n}^2 + 2\lambda \|Df\|_1 \right\}.$$

Compatibility constant

$$\kappa^2(S) := s \min \left\{ \|f\|_{Q_n}^2 : \|D_S f\|_1 - \|D_{-S} f\|_1 = 1 \right\}.$$

[Dalalyan, Hebiri, Lederer, 2017]

Effective sparsity

$$\Gamma^2(S) := \frac{s}{\kappa^2(S)}$$

Lemma *We have*

$$\|f^* - f^0\|_{Q_n}^2 \leq \min_{f, S} \left\{ \|f - f^0\|_{Q_n}^2 + \lambda^2 \Gamma^2(S) \right\}.$$

A bound for the effective sparsity

Dual certificate

Let $z_S \in \{\pm 1\}^S$ and $q \in \mathbb{R}^m$ be an interpolating vector:

$$q_S = z_S, \quad \|q\|_\infty \leq 1.$$

Then

$$\Gamma^2(S) \leq n \|D^T q\|_2^2.$$

[Candès and Fernandez-Granda, 2013]

Remark

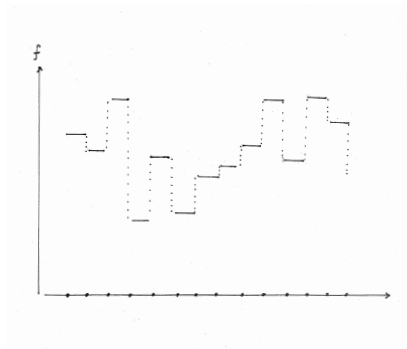
Unlike in the synthesis formulation there is no range condition on q

Example: the total variation penalty

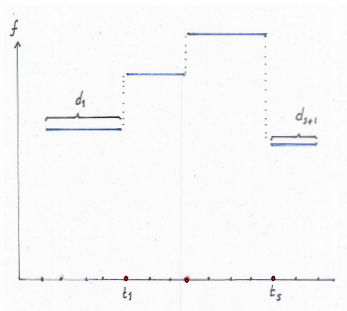
Consider the analysis problem

$$f^* := \arg \min \left\{ \|f - f^0\|_{Q_n}^2 + 2\lambda \sum_{i=2}^n |f_i - f_{i-1}| \right\}.$$

$TV(f)$ large



$TV(f)$ small



Remark

The analysis problem

$$f^* := \arg \min \left\{ \|f - f^0\|_{Q_n}^2 + 2\lambda \sum_{i=2}^n |f_i - f_{i-1}| \right\}.$$

corresponds to the synthesis problem

$$\beta^* := \arg \min_{b \in \mathbb{R}^n} \left\{ \|Xb - f^0\|_{Q_n}^2 + 2\lambda \sum_{j=2}^n |b_j| \right\}.$$

where

$$X = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ 1 & 1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \cdots & 1 \end{pmatrix}$$

$$\|Df\|_1 := \sum_{j=2}^n |f_j - f_{j-1}| \qquad D^T q = - \begin{pmatrix} q_1 \\ q_2 - q_1 \\ \vdots \\ -q_{n-1} \end{pmatrix}$$

with

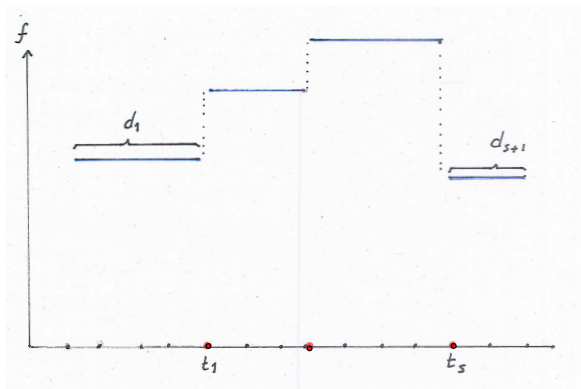
$$D = \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & & \cdots & -1 & 1 \end{pmatrix}$$

Let $f \in \mathbb{R}^n$.

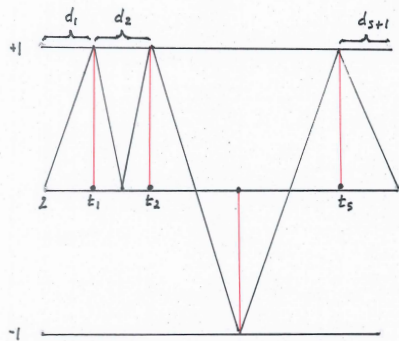
Let s be the number of jumps of f .

Let $1 < t_1 < \dots < t_s < n$ be the location of the jumps.

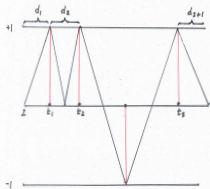
Let $d_1 = t_1$, $d_2 = t_2 - t_1$, $\dots$, $d_{s+1} = n - t_s$ be the distances
between jumps.



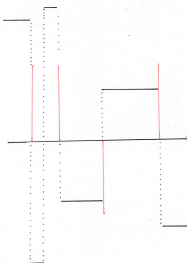
Interpolating vector q for the total variation penalty on f



Interpolating vector q



The vector $D^T q$



$$D^T q = - \begin{pmatrix} q_1 \\ q_2 - q_1 \\ \vdots \\ -q_{n-1} \end{pmatrix} = \begin{pmatrix} -1/d_1 \\ \vdots \\ -1/d_1 \\ +2/d_2 \\ \vdots \\ -2/d_2 \\ \vdots \\ +1/d_{s+1} \\ \vdots \\ +1/d_{s+1} \end{pmatrix}$$

Thus

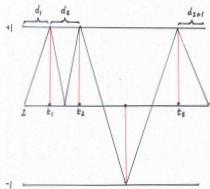
$$\Gamma^2(S) \leq \|D^T q\|_2^2 = \frac{n}{d_1} + \frac{4n}{d_2} + \cdots + \frac{4n}{d_s} + \frac{n}{d_{s+1}}$$

Remark

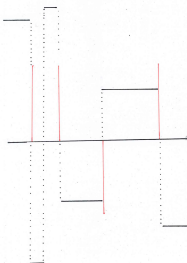
In the background we are simply doing partial integration:

$$(q, Df) = (D^T q, f).$$

Interpolating vector q



The vector $D^T q$



Corollary

We have

$$\|f^* - f^0\|_{Q_n}^2 \leq \|f - f^0\|_{Q_n}^2 + \lambda^2 \left(\frac{n}{d_1} + \sum_{j=2}^s \frac{4n}{d_j} + \frac{n}{d_{s+1}} \right).$$

Note

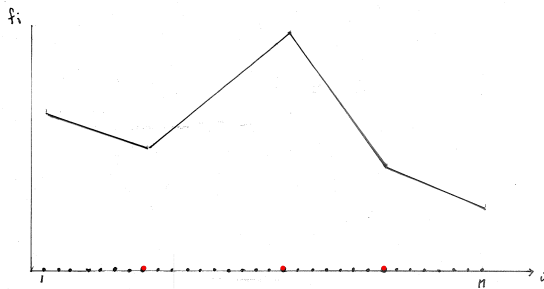
$$\left(\frac{n}{d_1} + \sum_{j=2}^s \frac{4n}{d_j} + \frac{n}{d_{s+1}} \right) \leq \frac{4(s+1)}{\min_{j=1}^{s+1} d_j/n}$$

Example: total variation penalty of second order differences

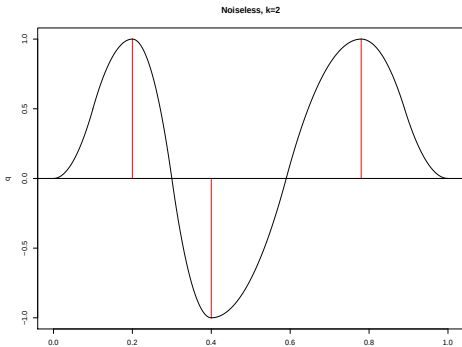
Let

$$\begin{aligned}\|Df\|_1 &:= \sum_{j=3}^n |\Delta f_j - \Delta f_{j-1}| \\ &:= \sum_{j=3}^n |f_j - 2f_{j-1} + f_{j-2}| \\ &:= \text{TV}(\Delta f).\end{aligned}$$

A function f with $\text{TV}(\Delta f)$ small



Interpolating vector q for the total variation penalty on Δf



One finds

$$\Gamma^2(S) \leq n \|D^T q\|_2^2 \leq 16 \sum_{j=1}^{s+1} \left(\frac{n}{d_j} \right)^3.$$

Example: total variation penalty of higher order differences

Let

$$\begin{aligned}\|Df\|_1 &:= \|\Delta^k f\|_1 \\ &:= \text{TV}(\Delta^{k-1} f).\end{aligned}$$

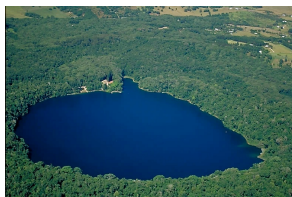
One finds

$$\Gamma^2(S) \leq n \|D^T q\|_2^2 \leq C_k^2 \sum_{j=1}^{s+1} \left(\frac{n}{d_j} \right)^{2k-1}.$$

Total variation in higher dimensions

$$\|Df\|_1 := \sum_{j=2}^{n_1} \sum_{k=2}^{n_2} |f_{j,k} - f_{j-1,k} - f_{j,k-1} + f_{j-1,k-1}|$$

Useful for pictures . . .

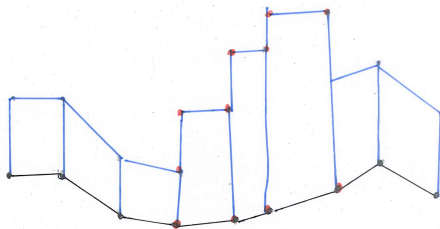


Example: total variation on graphs

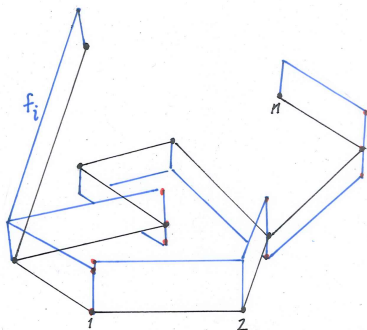
We have considered the path graph



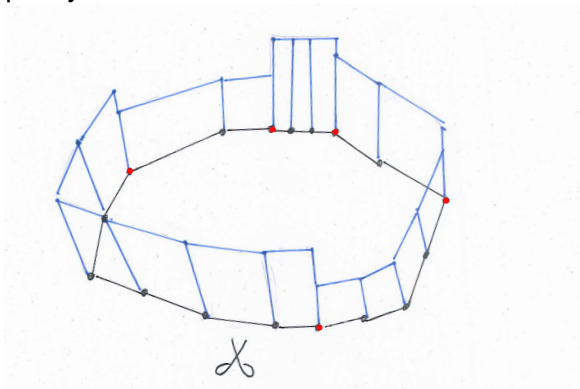
At each node i is attach a weight f_i ($i = 1, \dots, n$)



We can also consider more general graphs

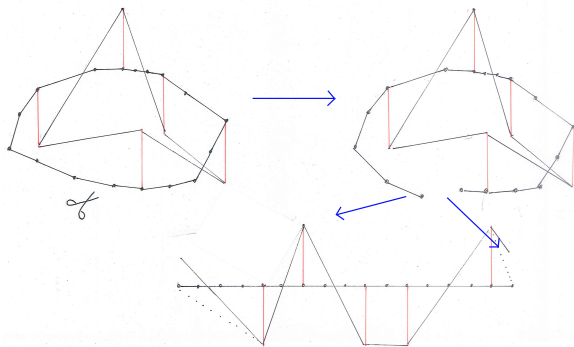


For example cycles



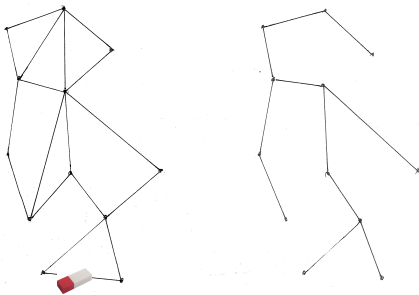
By cutting it we get back the path graph!

The interpolating vector q for the cycle



The interpolating vector is as for the path graph
It is actually easier now,
because we are free in choosing where to cut!

For general graphs, we first construct a generating tree and then split up the tree into path graphs. Both steps depend on where the jumps are.



Conclusion

The theory goes through for total variation penalties on graphs.

Adding noise

We consider the model

$$Y = f^0 + \epsilon, \quad \epsilon \sim \mathcal{N}(0, I)$$

analysis:

$$\hat{f} = \arg \min_{f \in \mathbb{R}^n} \left\{ \|Y - f\|_{Q_n}^2 + 2\lambda \|Df\|_1 \right\}$$

Notation

For $N \in \mathbb{N}$ and vectors $a \in \mathbb{R}^N$ and $b \in \mathbb{R}^N$ with the same length we write

$$ab := \begin{pmatrix} a_1 b_1 \\ a_2 b_2 \\ \vdots \\ a_N b_N \end{pmatrix}.$$

Analysis

Compatibility constants for noisy analysis

Let $w_{-S} \in [0, 1]^{p-s}$ be a vector of weights and

$$\kappa^2(S, w_{-S}) = s \min \left\{ \|f\|_{Q_n}^2 : \|D_S f\|_1 - \|(1 - w_{-S})D_{-S} f\|_1 = 1 \right\}.$$

Let

$\mathcal{N}_{-S} :=$ the null space of D_{-S}

$r_S := \text{rank}(\mathcal{N}_{-S})$

$\tilde{\omega}_j^{-1} :=$ length of the antiprojection of the j^{th} column of D_{-S}
on the other columns of D_{-S}

and

$$\omega_j := \begin{cases} \tilde{\omega}_j & \text{if } \tilde{\omega}_j \text{ is among the } n - r_S \text{ smallest } \tilde{\omega}_j \\ 0 & \text{else} \end{cases}$$

Remark If D_{-S} has full rank we have

$$\tilde{\omega}_{-S}^2 = \text{diag}(D_{-S}D_{-S}^T)^{-1}/n$$

Define

$$\lambda(t) := \sqrt{\frac{2 \log(2(n - r_S)) + 2t}{n}} \quad \text{“noise level”}$$

Take

$$\lambda \geq \lambda(t) \|\omega_{-S}\|_{\infty}$$

$$w_{-S} := \frac{\lambda(t)}{\lambda} \times \omega_{-S}$$

Oracle inequality for the analysis problem

Theorem

For all $t > 0$ and $x > 0$,
we have with probability at least $1 - \exp[-t] - \exp[-x]$

$$\begin{aligned} \|\hat{f} - f^0\|_{Q_n}^2 &\leq \overbrace{\|f - f^0\|_{Q_n}^2}^{\text{approximation error}} \\ &\quad + \overbrace{\left(\sqrt{\frac{r_S}{n}} + \sqrt{\frac{2x}{n}} + \lambda \Gamma(S, w_{-S}) \right)^2}^{\text{estimation error}} \\ &\quad + \overbrace{4\lambda \|D_{-S} f\|_1}^{\text{possibly left over small coefficients}} \end{aligned}$$

where

$$\Gamma^2(S, w_{-S}) = \frac{s}{\kappa^2(S, w_{-S})}$$

is the effective sparsity for the noisy analysis problem.

Define an interpolating vector q adjusted to noise as follows:

$$q_j = \{\pm 1\}, j \in S$$

$$|q_j| \leq 1 - w_j, j \notin S.$$

Then

$$\Gamma^2(S, w_{-S}) \leq n \|D^T q\|_2.$$

Thus

$$\|\hat{f} - f^0\|_{Q_n}^2 \leq \|f - f^0\|_{Q_n}^2 + \left(\sqrt{\frac{r_S}{n}} + \sqrt{\frac{2x}{n}} + \lambda \sqrt{n} \|D^T q\|_2 \right)^2.$$

Example: the total variation penalty for the noisy case

We have

$$\omega_{t_{j-1}+k}^2 = \frac{k(d_j - k)}{nd_j}, 1 \leq k \leq d_j - 1, j = 1, \dots, s+1$$

$$\|\omega^2\|_\infty = \frac{d_{\max}}{4n}, \quad d_{\max} := \max_{1 \leq j \leq s+1} d_j.$$

We therefore must take

$$\begin{aligned} \lambda &\geq \underbrace{\lambda(t)}_{\asymp \sqrt{\log n/n}} \sqrt{\frac{d_{\max}}{4n}} \\ &\asymp \frac{1}{n} \sqrt{d_{\max} \log n} \end{aligned}$$

An interpolating vector adjusted for noise has

$$|q_k| := 1 - \sqrt{\frac{d_1 - k}{d_1}} \quad , \quad 1 \leq k \leq d_1$$

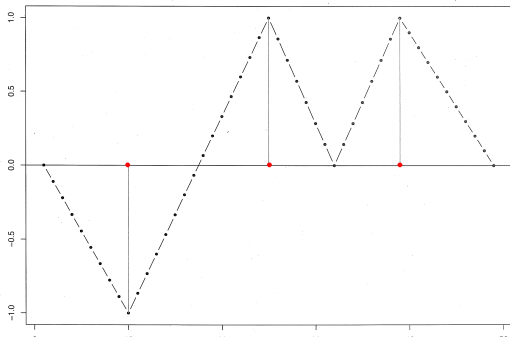
$$|q_{t_{j-1}+k}| = 1 - 2\sqrt{\frac{k(d_j - k)}{d_j^2}} \quad , \quad 1 \leq k \leq d_j - 1, \quad j = 2, \dots, s$$

$$|q_{t_s+k}| = 1 - \sqrt{\frac{k}{d_{s+1}}} \quad , \quad 1 \leq k \leq d_{s+1} - 1$$

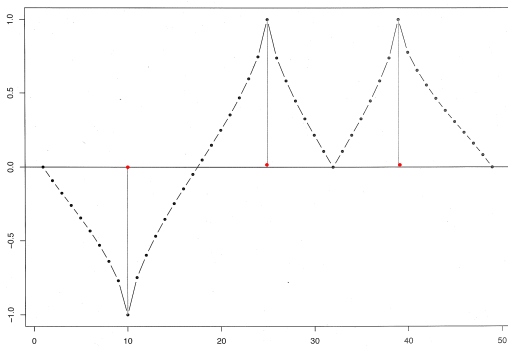
We get

$$\|q\|_2^2 \leq \frac{1}{d_1^2} + \frac{\log d_j}{d_j} + \sum_{j=2}^s \frac{4 \log(d_j/2)}{d_j} + \frac{\log d_{s+1}}{d_{s+1}} + \frac{1}{d_{s+1}^2}$$

Interpolating vector q for the total variation penalty on f



Adjusted to the noisy case



Application of the theorem gives for the total variation penalty:

For all $f \in \mathbb{R}^p$, $S \in [1, m]$, $t > 0$ and $x > 0$,
we have with

$$\lambda \geq \underbrace{\lambda(t)}_{\asymp \sqrt{\log n/n}} \sqrt{\frac{d_{\max}}{4n}}$$

and with probability at least $1 - \exp[-t] - \exp[-x]$

$$\begin{aligned} \|\hat{f} - f^0\|_{Q_n}^2 &\leq \|f - f^0\|_{Q_n}^2 \\ &+ \left(\sqrt{\frac{s+1}{n}} + \sqrt{\frac{2x}{n}} + \lambda \Gamma(S, w_{-s}) \right)^2 \\ &+ 4\lambda \|D_{-S} f\|_1 \end{aligned}$$

where

$$\begin{aligned} &\Gamma^2(S, w_{-s}) \\ &\leq \frac{n}{d_1^2} + \frac{n \log d_1}{d_1} + \sum_{j=2}^s \frac{4n \log(d_j/2)}{d_j} + \frac{n \log d_{s+1}}{d_{s+1}} + \frac{n}{d_{s+1}^2}. \end{aligned}$$

Special case: equally spaced jumps.

Suppose $d_j = \frac{n}{s+1}$. We get

$$\Gamma^2(S, w_{-S}) \leq 6(s+1)^2 \log n$$

The requirement $\lambda \geq \lambda(t) \sqrt{d_{\max}/(4n)}$ is met for

$$\lambda \asymp \underbrace{\sqrt{\frac{\log n}{n}}}_{\asymp \lambda(t)} \underbrace{\sqrt{\frac{1}{s+1}}}_{\asymp \sqrt{d_{\max}/n}}.$$

So we get

$$\begin{aligned} \lambda^2 \Gamma^2(S, w_{-S}) &= \mathcal{O}\left(\frac{s+1}{n}\right) \log^2 n \\ &= \mathcal{O}\left(\frac{\text{number of parameters}}{\text{number of observations}}\right) \log^2 n \end{aligned}$$

Conclusion for the general analysis problem with noise

In principle solved

e.g. for total variation on higher order derivatives

[Guntuboyina,
Lieu, Chatterjee
and Sen, 2014]

THANK YOU!

