Adaptive rates for trend filtering using dual certificates

Sara van de Geer

December, 2020

CIRM 2020

Joint work with
Francesco Ortelli

Contents

This course is based on ideas from

- Koltchinskii, Lounici and Tsybakov [2011]
- Dalalyan, Hebiri and Lederer [2016]
- Fuchs [2004]
- Candès and Fernandez-Granda [2014]
- ...

Lasso

- "Structured" design matrix $\Psi \in \mathbb{R}^{n \times p}$
- Response vector $Y \in \mathbb{R}^n$, $Y \sim \mathcal{N}(f^0, I)$

Goal: estimate the unknown f^0 using the Lasso

$$\hat{f} := \arg \min_{f_b = \Psi b,\ b \in \mathbb{R}^p} \left\{ \|Y - \Psi b\|_2^2 / n + 2\lambda \|b\|_1 \right\}.$$

Let $\Psi \subset (L_2(Q), \|\cdot\|_Q)$.

Notation

For a linear space $\mathcal{V} \subset L_2(Q)$ and $f \in L_2(Q)$ we let $f_{\mathcal{V}}$ be the projection of f on $\mathcal{V}$ and $f_{\mathcal{V}^\perp} := f - f_{\mathcal{V}}$. We let

$$\mathbf{u}(\mathcal{V}, \Psi) := \sup_{\psi \in \Psi} \|\psi_{\mathcal{V}^\perp}\|_Q.$$

Definition

For $N \in \mathbb{N}$ define the best N-term approximation

$$\gamma(N, \Psi) := \min\left\{\mathbf{u}(\mathcal{V}, \Psi) : \ \dim(\mathcal{V}) = N\right\}.$$

Let $\lambda_0(u) := \sqrt{2\log(2p) + 2u}$.

Theorem

Suppose $\gamma(N, \Psi) \leq AN^{-1/W}$.

With probability at least $1 - \exp[-u] - \exp[-v]$, *and for*

$$\delta_n^2 = \frac{(2A\lambda_0(u))^W}{n\lambda^W} + \frac{1+2v}{n}$$

we have

$$\|\hat{f} - f^0\|_2^2/n + \lambda\|\hat{\beta}\|_1 \leq \underbrace{\left\{\|f^* - f^0\|_2^2/n + 3\lambda\|\beta^*\|_1\right\}}_{\text{approximation error}}$$

$$+\delta_n^2.$$

$$\uparrow$$

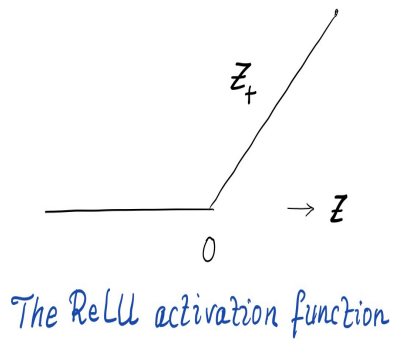
estimation error

Application: one-hidden-layer NN

Definition

The hinge, or ReLU, function is

$$z \mapsto z_+ : \begin{cases} z, & z \geq 0 \\ 0, & z < 0 \end{cases}$$



Notation

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

$$(v - c)_+ := \begin{pmatrix} (v_1 - c)_+ \\ \vdots \\ (v_n - c)_+ \end{pmatrix}.$$

Construction of the dictionary Ψ

- $d :=$ input dimension
- $X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^{n \times d}$ input matrix
- $W := (w_1, \ldots w_p) \in \mathbb{R}^{r \times p}$ matrix of weights
- $\|w_j\|_2 = 1, j \in \{1, \ldots, p\}$
- $\mathbf{c} = (c_1, \ldots, c_p) \in \mathbb{R}^p$ vector of biases
- $\psi_j := (Xw_j - c_j)_+ \in \mathbb{R}^n, j = 1, \ldots, p$
- $\Psi := \{\psi_1, \ldots, \psi_p) = (XW - \mathbf{c})_+ \in \mathbb{R}^{n \times p}$

Lemma *Suppose* $\max_{1 \le i \le n} \|x_i\|_2 \le 1$. *Then*

$$\gamma(N, \Psi) \le AN^{-1/d}.$$

Lasso for 1-hidden-layer NN

Define

$$\mathcal{NN} := \left\{ f_b = (XW - \mathbf{c})_+ b : \ b \in \mathbb{R}^p \right\}.$$

Let

$$\hat{f} = f_{\hat{\beta}} := \arg \min_{f_b \in \mathcal{NN}} \left\{ \|Y - f_b\|_2^2 / n + 2\lambda \|b\|_1 \right\}.$$

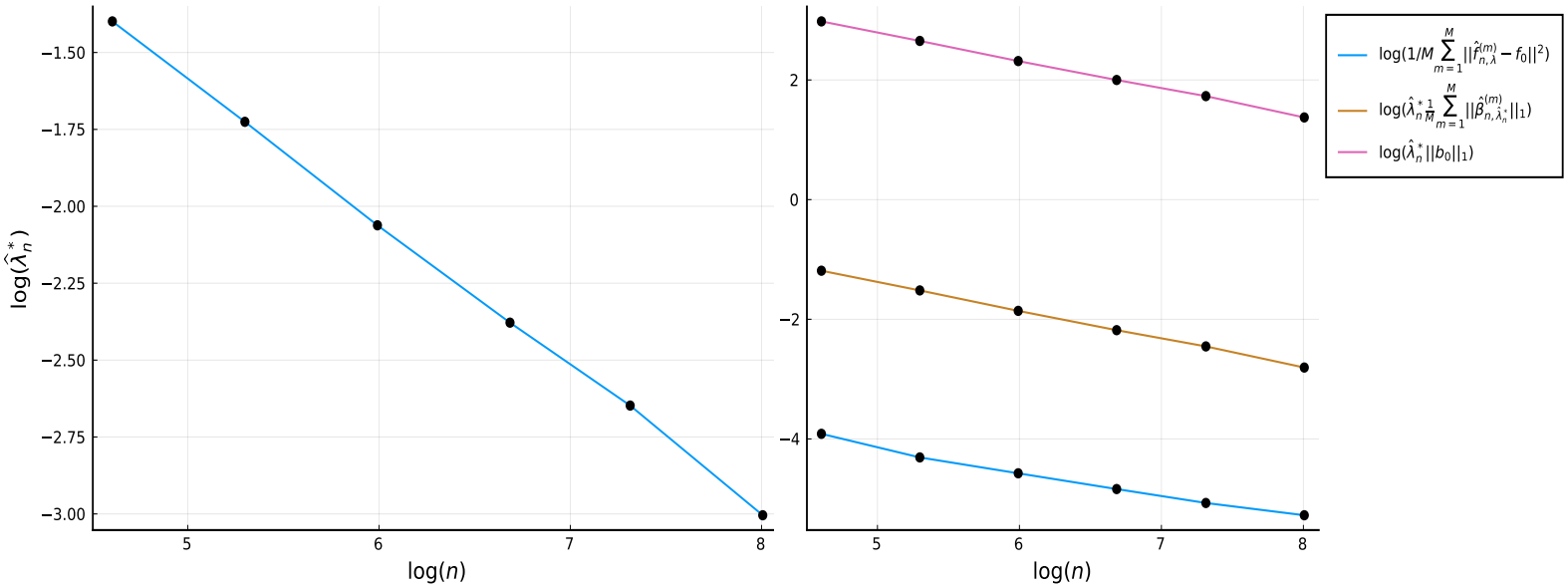
[Parhi and Nowak , 2020]

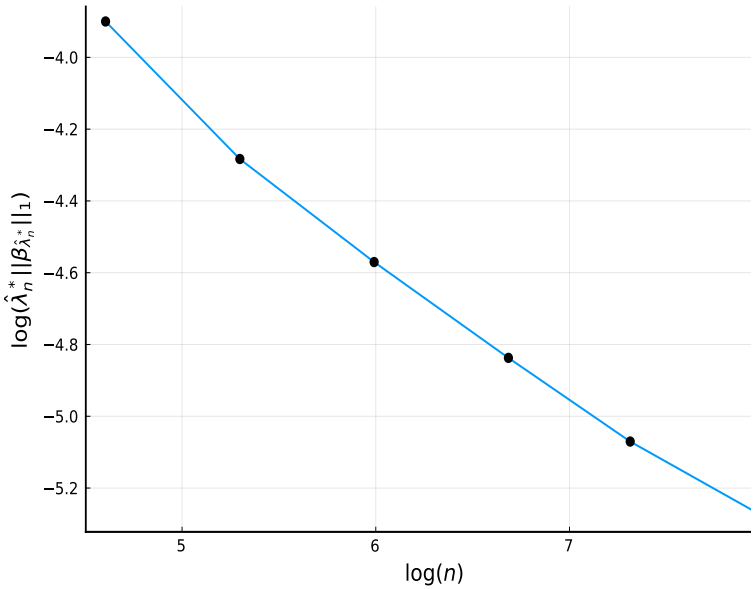
Simulation by Peter Hinz

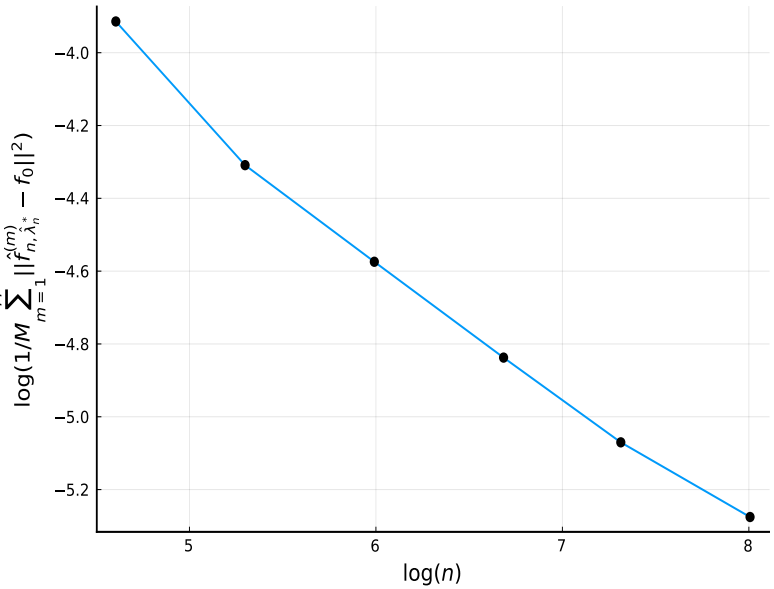
$d := 100$

$p := 10000$

$\sigma^2 := 1$


−1.50
−1.75
−2.00
−2.25
−2.50
−2.75
−3.00
log($\hat{\lambda}_n^*$)
5
6
7
8
log(n)
2
0
−2
−4
5
6
7
8
log(n)
log(1/M $\sum_{m=1}^{M}$ $||\hat{f}_{n,\hat{\lambda}}^{(m)} - f_0||^2$)
log($\hat{\lambda}_n^* \frac{1}{M}$ $\sum_{m=1}^{M}$ $||\hat{\beta}_{n,\hat{\lambda}_n^*}^{(m)}||_1$)
log($\hat{\lambda}_n^* ||b_0||_1$)


−4.0
−4.2
−4.4
−4.6
−4.8
−5.0
−5.2
5
6
7
log(n)
log(λ̂*_n ||β_{λ̂*_n}||_1)

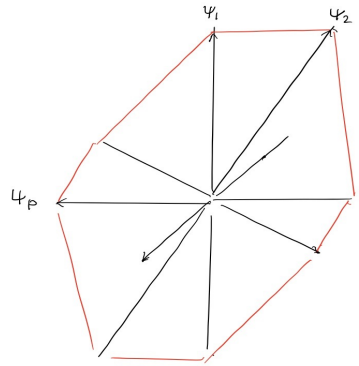

−4.0
−4.2
−4.4
−4.6
−4.8
−5.0
−5.2
5
6
7
8
log(n)
log(1/M Σ_{m=1}^{M} ||f̂^{(m)}_{n,λ̂*_n} − f_0||^2)

Question:

How good are the bounds using best N-term approximations?

Note:

- $\{\Psi b : \ b \in \mathbb{R}^p, \|b\|_1 = 1\} = \mathrm{conv}(\pm\Psi)$ is the convex hull of $\pm\Psi$

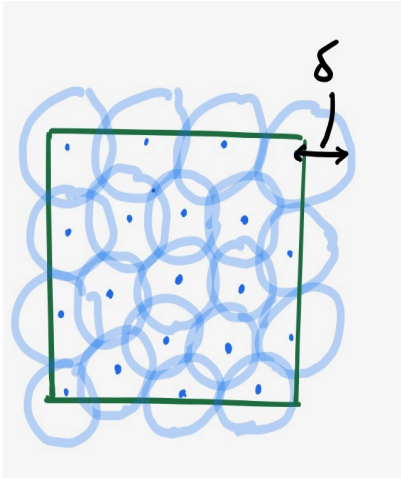


- Ψ “structured": it has polynomial covering numbers $N(\cdot)$
- [Ball and Pajor, 1990] $\rightsquigarrow$ entropy $H(\cdot)$ of $\mathrm{conv}(\pm\Psi)$
- Estimation error for LSE over $\mathrm{conv}(\pm\Psi)$ is $\delta_n^2 \asymp H(\delta_n)/n$

Covering and entropy

Let (S, d) be a metric space.

Definition *For $\delta > 0$, δ-covering number $N(\delta, S, d) :=$ minimum # balls with radius δ necessary to cover S*



Definition *Entropy* $H(\cdot, S) := \log N(\cdot, S)$.

Entropy of convex hulls

Let $\Psi \in (L_2(Q), \|\cdot\|_Q)$.

Theorem [Ball & Pajor, 1990] ["Covering Theorem"]

Suppose for some positive constants A and W

$$N(\delta, \Psi) \leq A\delta^{-W}, \ \delta > 0.$$

Then for a constant C

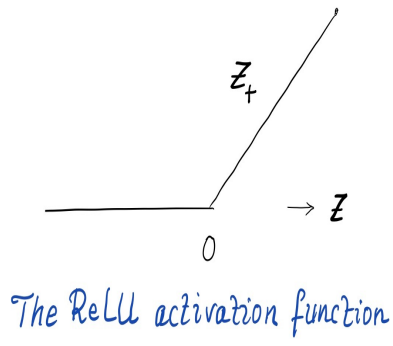
$$H(\delta, \mathrm{conv}(\Psi)) \leq C\delta^{-\frac{2W}{2+W}}, \ \delta > 0.$$

The entropy bound based on covering numbers can be too loose

Definition

The hinge, or ReLU, function is

$$z \mapsto z_+ : \begin{cases} z, & z \geq 0 \\ 0, & z < 0 \end{cases}$$



Let

$$X := \left(\frac{1}{n}, \frac{2}{n}, \cdots, 1\right)^\top,$$

$$\psi_1 \equiv 1,\ \psi_2 := X - 1/n,\ \psi_j := \left(X - \frac{j-1}{n}\right)_+,\ j \in \{3, \ldots, n\}.$$

Define the space of linear functions

$$\mathcal{N} := \{z = \psi_1 b_1 + \psi_2 b_2 : (b_1, b_2)^\top \in \mathbb{R}^2\}$$

Then

$$\left\{z + f : \ z \in \mathcal{N}, \ f \in \mathrm{conv}(\{\psi_j\}_{j=3}^n)\right\} = \left\{f \in \mathbb{R}^n : \ n\|\Delta^2 f\|_1 = 1\right\}$$

with

$$(\Delta^2 f)_j := f_j - 2f_{j-1} + f_{j-2}, \ j \in \{3, \ldots, n\}.$$

Let

$$\|v\|_{Q_n}^2 := \frac{1}{n}\sum_{i=1}^n v_i^2, \ v \in \mathbb{R}^n.$$

By the [Covering Theorem] with $W = 1$

$$H(\delta, \mathrm{conv}(\{\psi_j\}_{j=3}^{n}), \|\cdot\|_{Q_n}) \leq C\delta^{-\frac{2}{3}}, \delta > 0.$$

But by [Babenko, 1979]

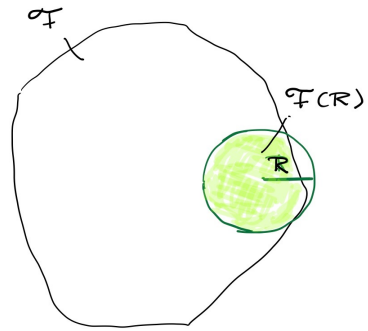
$$H(\delta, \mathrm{conv}(\{\psi_j\}_{j=3}^{n}), \|\cdot\|_{Q_n}) \leq C\delta^{-\frac{1}{2}}, \delta > 0!$$

Notation

For $R > 0$,

$\mathcal{F}(R) :=$

$\{f \in \overline{\text{conv}}(\Psi) : \ \|f\|_Q \leq R\}$



Theorem ["Projection Theorem"]
Suppose that for some constants positive A and W

$$\gamma(N, \Psi) \leq A N^{-\frac{1}{W}}.$$

For all $R > 0$ *and* $\delta > 0$,

$$\begin{aligned} & H(\delta, \mathcal{F}(R)) \\ \leq\ & \underbrace{2\left(\frac{4A}{\delta}\right)^{\frac{2W}{2+W}}}_{\text{important part}} \underbrace{\log^{\frac{2}{2+W}}\left(\frac{4R+\delta}{\delta}\right) H^{\frac{W}{2+W}}(\delta/4, \Psi)}_{\text{just some log terms}} \\ +\ & \underbrace{\log\left(\frac{4R+\delta}{\delta}\right) + H_2(\delta/4, \Psi)}_{\text{some more log terms}}. \end{aligned}$$

In other words,

$$H(\delta, \mathcal{F}(R)) \lesssim \delta^{-\frac{2W}{2+W}} \log(1/\delta)$$

and for the local entropy where $R \asymp \delta$ we get

$$H(\delta, \mathcal{F}(R)) \lesssim \delta^{-\frac{2W}{2+W}} \log^{\frac{W}{2+W}}(1/\delta).$$

Example: second order discrete derivatives.

$$\mathcal{F} := \left\{ f \in \mathbb{R}^n : \ n\|\Delta^2 f\|_1 = 1 \right\} = \left\{ \text{linear} + \text{conv}(\underbrace{\{\psi_j\}_{j=3}^n}_{\text{ReLU functions}}) \right\}$$

Then

$$\begin{aligned} & N(\delta, \{\psi_j\}_{j=3}^n, Q_n) \asymp \delta^{-1} : \ W_{\text{cov}} = 1 \\ \overset{\text{[Covering Thm]}}{\Rightarrow} \quad & H(\delta, \text{conv}\{\psi_j\}_{j=3}^n) \lesssim \delta^{-\frac{2W_{\text{cov}}}{2+W_{\text{cov}}}} = \delta^{-\frac{2}{3}}. \end{aligned}$$

$$\begin{aligned} & \gamma(N, \{\psi_j\}_{j=3}^n) \leq AN^{-\frac{1}{2/3}} : \ W_{\text{proj}} = \frac{2}{3} \\ \overset{\text{[Projection Thm]}}{\Rightarrow} \quad & H(\delta, \text{conv}\{\psi_j\}_{j=3}^n) \lesssim \delta^{-\frac{2W_{\text{proj}}}{2+W_{\text{proj}}}} \log n = \delta^{-\frac{1}{2}} \log n. \end{aligned}$$

Example: k^{th} order discrete derivatives

.

$$\mathcal{F} := \left\{ f \in \mathbb{R}^n : \; n^{k-1} \|\Delta^k f\|_1 = 1 \right\}$$

$$= \left\{ (k-1)^{th} \text{ order polynomial} + \text{conv}(\{\psi_j\}_{j=k+1}^n) \right\},$$

where $\{\psi_j\}_{j=k+1}^n$ is the falling factorial basis.

[Wang, Smola, Tibshirani, 2014]

Then

$$N(\delta, \{\psi_j\}_{j=k+1}^n) \asymp \delta^{-1} : W_{\text{cov}} = 1$$

$$\overset{\text{[Covering Thm]}}{\Rightarrow} \quad H(\delta, \{\psi_j\}_{j=k+1}^n) \lesssim \delta^{-\frac{2W_{\text{cov}}}{2+W_{\text{cov}}}} = \delta^{-\frac{2}{3}}.$$

$$\gamma(N, \{\psi_j\}_{j=k+1}^n) \leq AN^{-\frac{1}{2/(2k-1)}} : \; W_{\text{proj}} = \frac{2}{2k-1}$$

$$\overset{\text{[Projection Thm]}}{\Rightarrow} \quad H(\delta, \{\psi_j\}_{j=k+1}^n) \lesssim \delta^{-\frac{2W_{\text{proj}}}{2+W_{\text{proj}}}} \log n = \delta^{-\frac{1}{k}} \log n.$$

Higher-dimensional extension (MARS)

Let $x_i = (\xi_{i,1}, \ldots, \xi_{i,r}) \in \{1/n_0, \ldots, 1\}^d$ and let $n = n_0^r$. Let $\{\psi_j\}_{j=1}^p$ be real-valued functions on $\{1/n_0, \ldots, 1\}$ with $\max_{1 \le j \le p} \|\psi_j\|_\infty \le 1$. Let for $\mathbf{j} = (j_1, \ldots, j_r) \in \{1, \ldots, p\}^d$

$$\psi_{\mathbf{j}}(x_i) = \prod_{t=1}^{d} \psi_{j_t}(\xi_{i,t}).$$

Then functions of the form

$$f = \sum_{\mathbf{j} \in \{1,\ldots,p\}^d} \psi_{\mathbf{j}} b_{\mathbf{j}}$$

with ψ_j the falling factorial basis functions corresponds to those used in the context of multiplicative adaptive regression splines (MARS) [Friedman, 1991] .

Let

$$\mathcal{F} := \left\{ f = \sum_{\mathbf{j} \in \{1,\ldots,p\}^d} \psi_{\mathbf{j}} b_{\mathbf{j}} : \sum_{\mathbf{j} \in \{1,\ldots,p\}^d} |b_{\mathbf{j}}| \leq 1 \right\}.$$

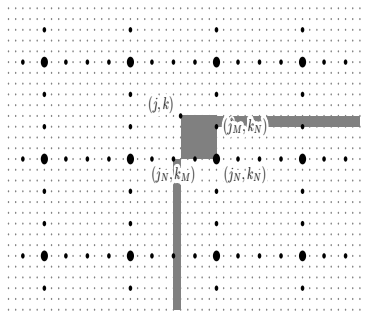
Then

$$\gamma(N, \{\psi_{\mathbf{j}}\}) \quad \leq \quad AN^{-\frac{2k-1}{2[1+\frac{1}{2}+\cdots+\frac{1}{d}]}}.$$

$$W = W_{\text{proj}} = \frac{2[1+\frac{1}{2}+\cdots+\frac{1}{d}]}{2k-1}$$

$$\left(W_{\text{cov}} = 2d \right)$$

which gives



$$\frac{2W}{2+W} = \frac{2h}{2k+h-1}, h = 1 + \frac{1}{2} + \cdots + \frac{1}{d}.$$

We then find by [Projection Theorem]

$$H_2(\delta, \mathcal{F}(R)) \lesssim \delta^{-\frac{2h}{2k+h-1}} \log p.$$

For $k = 1$ still not tight

$$\frac{1}{\delta} \sim H_2(\delta, \mathcal{F}(R)) \lesssim \delta^{-\frac{2h}{h+1}} \log p$$

Part II: Adaptation for the trend filtering problem

Trend filtering problem:

$$\begin{aligned}\hat{f} &:= \min_{f \in \mathbb{R}^n} \left\{ \|Y - f\|_n^2 + 2\lambda \|\Delta^k f\|_1 \right\} \\ &= \min_{f_b = \Psi b} \left\{ \|Y - f_b\|_n^2 + 2\lambda \|b_{-\{1,\ldots,k\}}\|_1 \right\}\end{aligned}$$

where Ψ is the falling factorial basis.

[Wang, Smola, Tibshirani, 2014]

Notation

$S \subset \{k+1, \ldots, n\}$, $s := |S|$

$$
\begin{aligned}
\mathcal{N}_{-S} &:= \{f \in \mathbb{R}^n : \ (\Delta^k f)_j = 0 \ \forall \, j \notin S\} \\
r_S &:= \dim(\mathcal{N}_{-S}) = k + s = \boxed{\text{sparsity}} \\
\|f^* - f^0\|_2^2/n &:= \text{approximation error}, \ f^* \in \mathcal{N}_{-S} \\
r_S/n &:= \text{estimation error}
\end{aligned}
$$

Goal: Show that $\|\hat{f} - f^0\|_n^2$ trades off approximation error and estimation error r_S/n.

Notation:

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

$$D_S f := \{(Df)_j : \ j \in S\}, \ D_{-S} f := \{(Df)_j : \ j \notin S\}$$

Thus

$$\mathcal{N}_{-S} = \{f \in \mathbb{R}^n : \ D_{-S} f = 0\}$$

Choice of the tuning parameter

Let $\mathcal{V} \supset \mathcal{N}_{-S}$ be a linear space.
Recall

$$\mathbf{u}(\mathcal{V}, \Psi) := \sup_{\psi \in \Psi} \|\psi_{\mathcal{V}^\perp}\|_{Q_n}.$$

Definition *For $N \geq r_S$ the best N-term approximation is*

$$\begin{aligned} \gamma_S(N, \Psi) &:= \min\left\{\mathbf{u}(\mathcal{V}, \Psi) : \ \mathcal{V} \supset \mathcal{N}_{-S}, \ \dim(\mathcal{V}) = N\right\} \\ &:= \mathbf{u}(\mathcal{V}_N, \Psi). \end{aligned}$$

Define for $u > 0$

$$\lambda_0(u) := \sqrt{\frac{2(\log(2n) + u)}{n}}.$$

Take

$$\lambda \geq \underbrace{\lambda_0(u)}_{\asymp \sqrt{\frac{\log n}{n}}} \underbrace{\gamma_S(N, \Psi)}_{\substack{\text{best } N-\text{term} \\ \text{approximation}}}$$

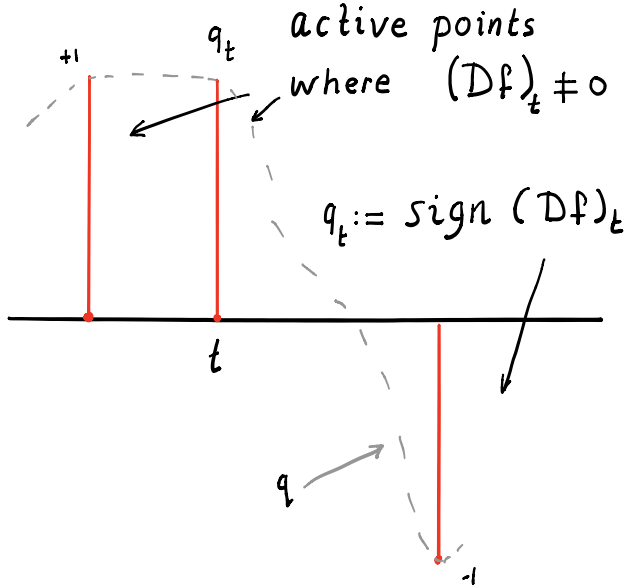
Effective sparsity without noise

Definition

Let $q \in \mathbb{R}^{n-k}$ where $q_S \in \{\pm 1\}^s$ ($s := |S|$) is a fixed sign vector. We call q an interpolating vector (that interpolates the signs at S).

Definition *The noiseless effective sparsity is*

$$\Gamma^2(q_S) := \min\left\{ n\|D^\top q\|_2^2 : \ q \text{ interpolating } \ |q_j| \le 1 \ \forall\, j \notin S \right\}.$$



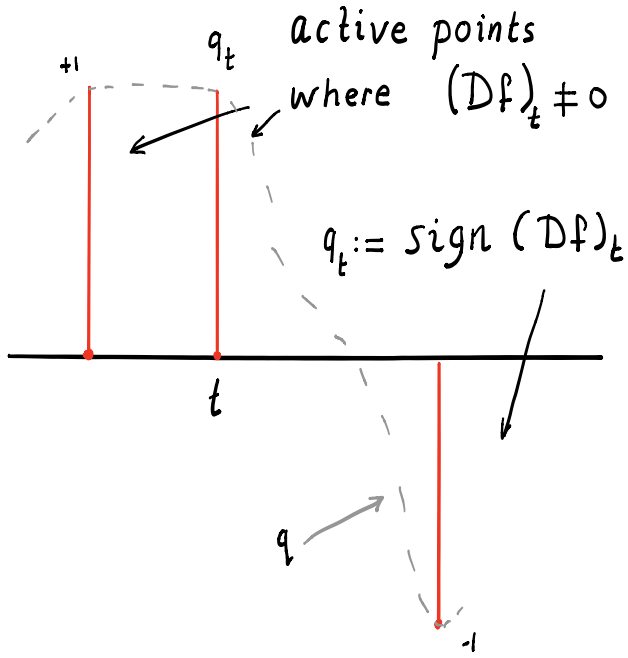
Remark This is a quantified version of the dual certificates used in noiseless compressed sensing.

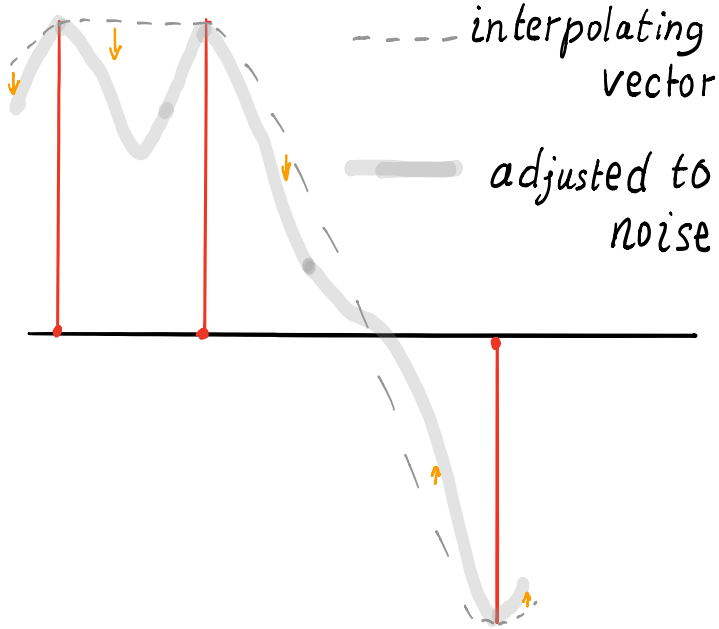
Definition Define the normalized noise weights

$$w_j := \lambda_0 \|\psi_{\mathcal{V}_N^\perp, j}\|_{Q_n}/\lambda, \ j \notin S.$$

The effective sparsity is

$$\Gamma^2(q_S, w_{-S}) := \min\left\{ n\|D^\top q\|_2^2 : \ q \text{ interpolating } \ |q_j| \le 1 - w_j \ \forall\, j \notin S \right\}.$$





Oracle inequality

Theorem *Let $f^* \in \mathbb{R}^n$ be arbitrary and $S := \{j : (Df^*)_j \neq 0\}$. Let $\lambda \geq \lambda_0(u)\mathcal{U}_S(N, \Psi)$. For all $u > 0$ and $v > 0$, we have with probability at least $1 - \exp[-u] - \exp[-v]$*

$$\|\hat{f} - f^0\|_2^2/n \leq \overbrace{\|f^* - f^0\|_2^2/n}^{\text{approximation error}} + \overbrace{\left(\sqrt{\frac{N}{n}} + \sqrt{\frac{2v}{n}} + \lambda\Gamma(q_S, w_{-S})\right)^2}^{\text{estimation error}}$$

where

$$\Gamma^2(q_S, w_{-S}),\ q_S := \text{sign}(D_S f^*),\ j \in S$$

is the effective sparsity.

[Dalalyan, Hebiri and Lederer; 2017]

Example
Total variation (TV) in $\mathbb{R}$

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

with

$$\|Df\|_1 := \sum_{j=2}^{n} |f_j - f_{j-1}| = \mathrm{TV}(f)$$

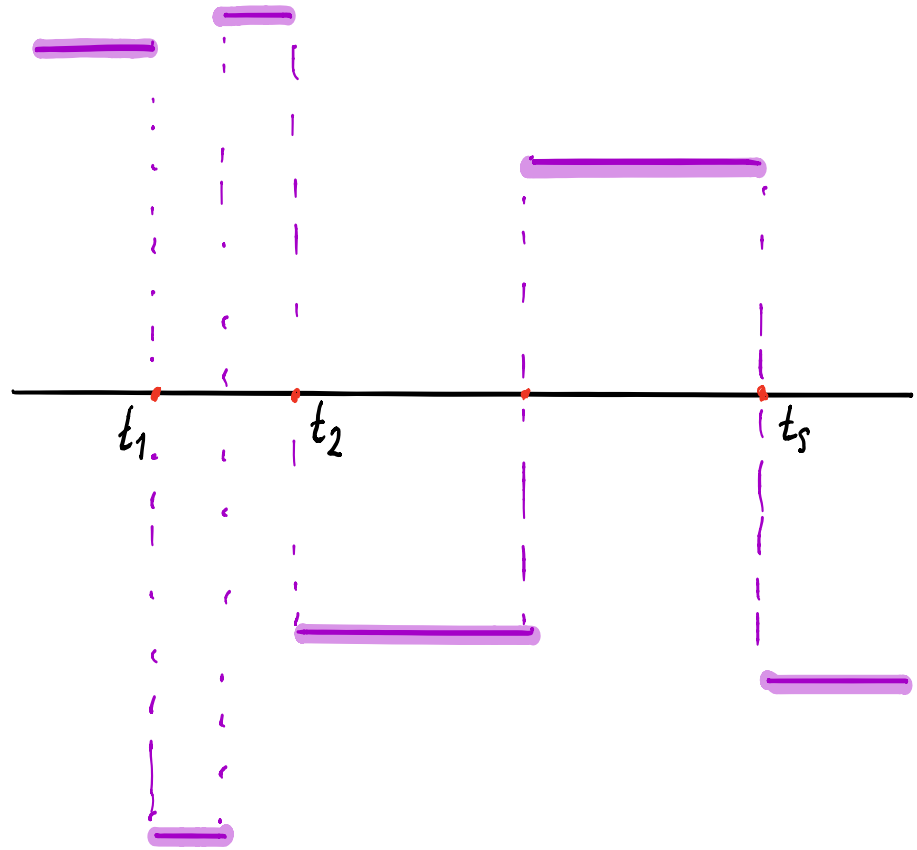
Thus

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

$$D^\top = - \begin{pmatrix} +1 & 0 & \cdots & 0 \\ -1 & +1 & \cdots & 0 \\ 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & +1 \\ 0 & 0 & & -1 \end{pmatrix}$$

Let

- $f \in \mathbb{R}^n$,
- s be its number of jumps,
- $1 < t_1 < \ldots < t_s < n$ be the location of the jumps,



- $n_1 = t_1 - 1$, $n_2 = t_2 - t_1$, $\ldots$, $n_{s+1} = n + 1 - t_s$ be the distances between jumps.

We take

$$N = r_S,\ \mathcal{V}_N = \mathcal{N}_{-S}.$$

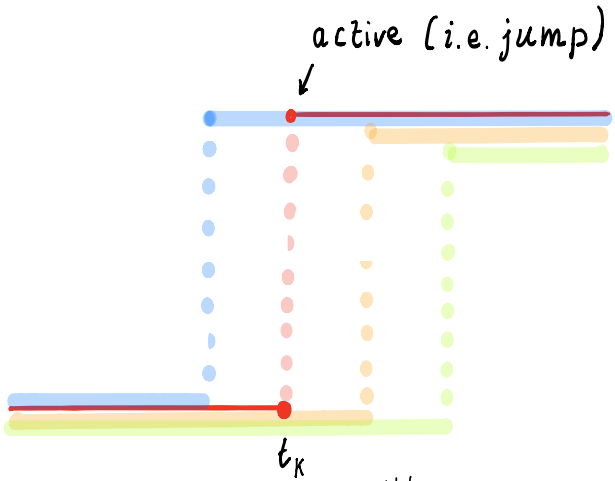
Then have noise weights

$$\|\psi_{\mathcal{V}_N^{\perp}, t_{j-1}+i}\|_2 = \sqrt{\frac{i(n_j - i)}{d_j}},\ i \in [1, n_j],$$

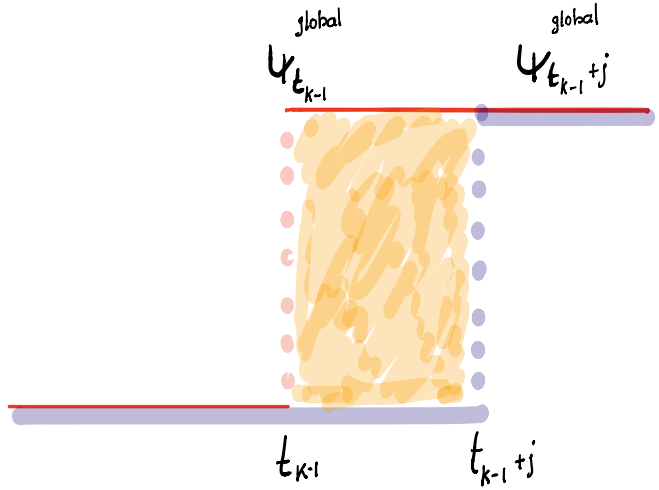
$$\gamma_S(N, \Psi) = \sqrt{\frac{n_{\max}}{4n}} \qquad (1/W = 1/2)$$

Choice of tuning parameter:

$$\lambda \geq \lambda_0(u)\sqrt{\frac{n_{\max}}{4n}}.$$

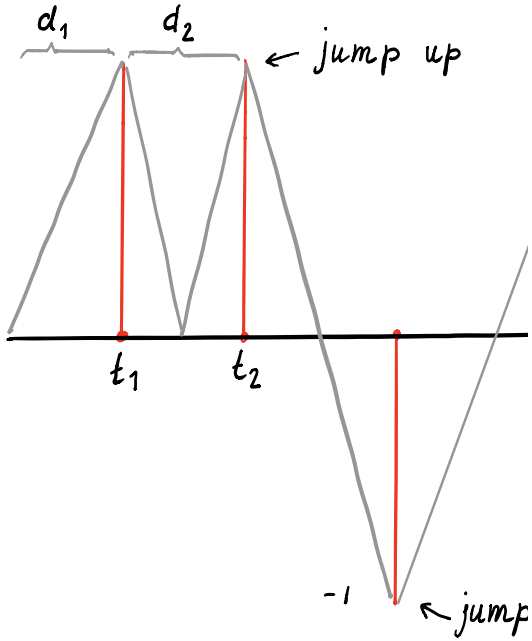


dictionary $\psi_1^{\text{global}}, \dots \psi_n^{\text{global}}$
for the total variation penalty

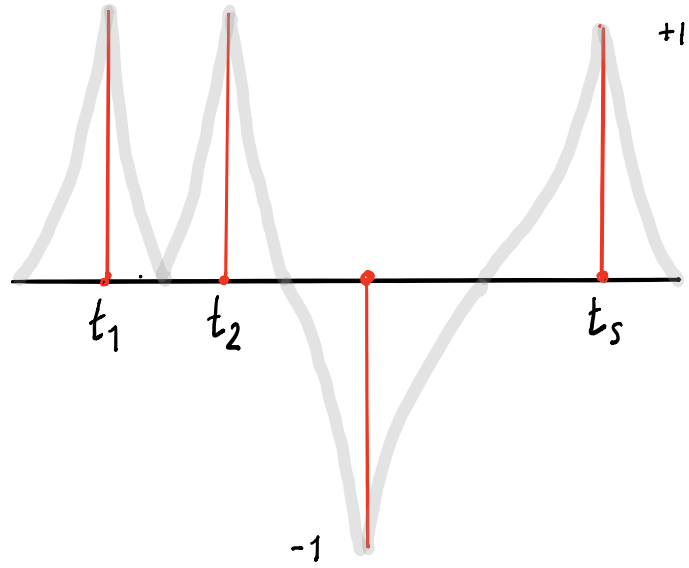


$= j \quad \Rightarrow \quad \| \psi_{t_{K-1}}^{\text{global}} - \psi_{t_{K-1}+j}^{\text{global}} \|_2 = \sqrt{j}$

no noise



noisy



Application of the theorem to the total variation penalty:

Take

$$\lambda \geq \underbrace{\frac{\sqrt{2(\log(2n)+u)}}{n}}_{\asymp\sqrt{\log n}/n} \sqrt{\frac{n_{\max}}{4}}.$$

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

$$\begin{aligned} \|\hat{f} - f^0\|_2^2/n &\leq \|f^* - f^0\|_2^2/n \\ &+ \left(\sqrt{\frac{s+1}{n}} + \sqrt{\frac{2v}{n}} + \lambda\Gamma(q_S, w_{-S})\right)^2 \\ &+ 4\lambda\|D_{-S}f^*\|_1 \end{aligned}$$

where

$$\Gamma^2(q_S, w_{-S}) \leq \frac{n\log n_1}{n_1} + \sum_{j=2}^{s} \frac{4n\log(n_j/2)}{n_j} + \frac{n\log n_{s+1}}{n_{s+1}}.$$

Special case: equally spaced jumps

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

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

We may take

$$\lambda \asymp \underbrace{\lambda_0(u)}_{\asymp\sqrt{\frac{\log n}{n}}} \underbrace{\sqrt{\frac{1}{s+1}}}_{\asymp\gamma_S(r_S,\Psi)}.$$

So we get

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

Example
total variation of first discrete derivative

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

with

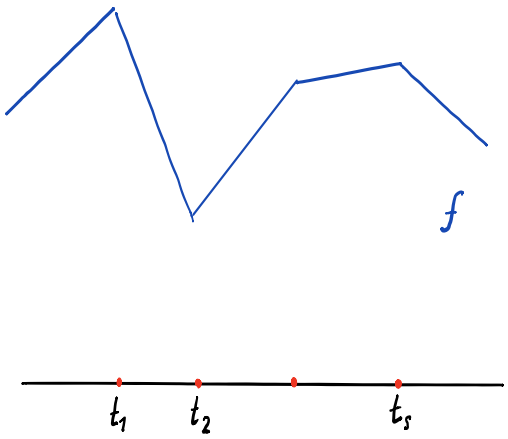
$$Df := \Delta^2 f$$

the second order discrete derivative

$$(\Delta^2 f)_j = f_j - 2f_{j-1} + f_{j-2}, \; j \geq 3.$$

[Tibshirani; 2014], [Guntuboyina, Lieu, Chatterjee & Sen; 2020]
Then

$$\| Df \|_1 := \| \Delta^2 f \|_1 = \mathrm{TV}(\Delta f).$$



Let

- $S := \{t_1, \ldots, t_s\}$, $t_1 < \cdots < t_s$,
- $n_j := t_j - t_{j-1}$, $j = 1, \ldots, s+1$,
- $n_{\max} := \max_{1 \le j \le s+1} n_j$

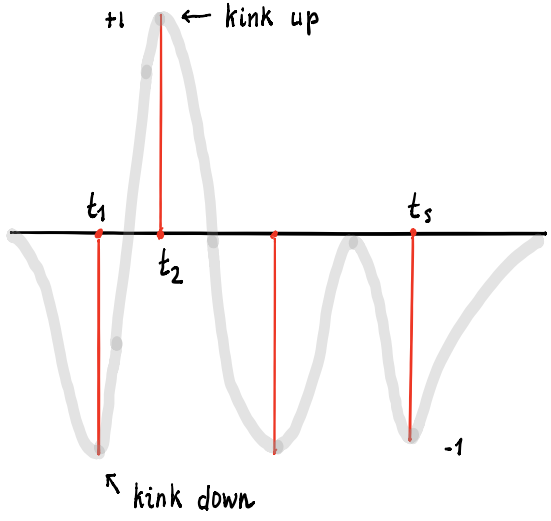
Let

$$\mathcal{V}_N := \mathcal{N}_{-S} \cup \operatorname{span}\{\psi_{t_1+1}, \ldots, \psi_{t_s+1}\},$$

$$N = r_S + s = 2(s+1)$$

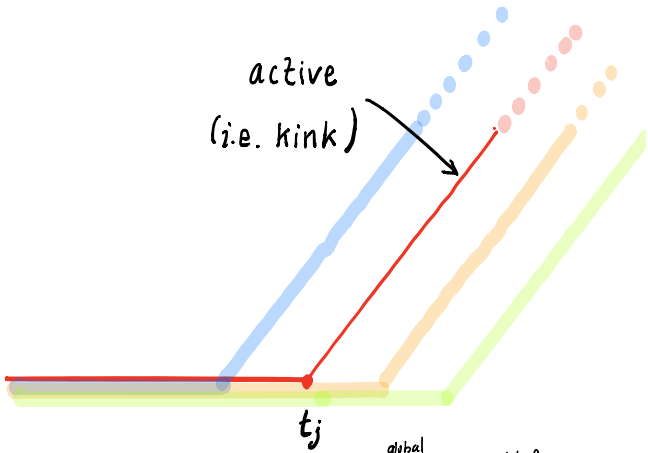
We get

$$\gamma_S(N, \Psi) \leq A n_{\max}^{3/2}/\sqrt{n} \qquad (1/W = 3/2).$$

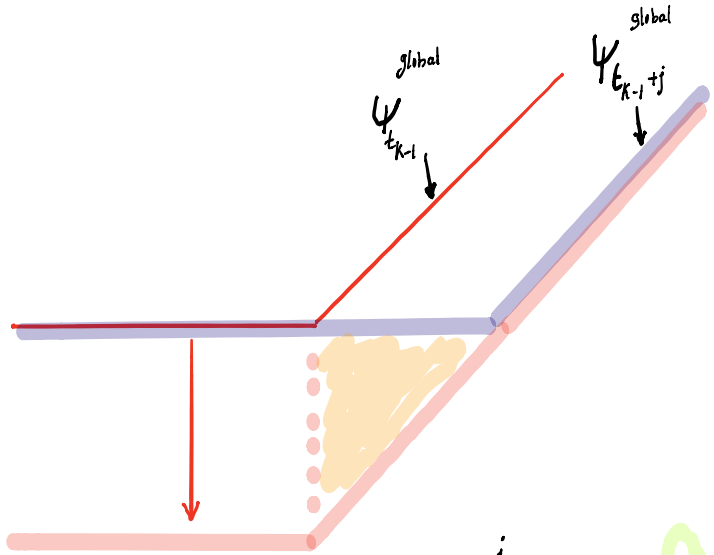


$$\Gamma^2(q_S) \leq n\|D^\top q_{\text{noiseless}}\|_2^2 \leq 16\sum_{j=1}^{s+1}\frac{n}{n_j^3}.$$

Noise and tuning: noise weights.



dictionary $\psi_1^{\text{global}}, \ldots, \psi_n^{\text{global}}$
for the 2^{nd} derivatives penalty



$$\text{distance}^2 \text{ to } \mathcal{N}_{-3} \leq \approx \int_0^j x^2 dx = \frac{j^3}{3}$$

Application of the theorem to the total variation penalty on Δf:

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

$$\lambda \geq c_2 n \lambda_0(u) \sqrt{n_{\max}^3 / n}$$

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

$$\|\hat{f} - f^0\|_2^2 / n \leq \|f^* - f^0\|_2^2 / n$$

$$+ \left(\sqrt{\frac{2(s+1)}{n}} + \sqrt{\frac{2v}{n}} + \lambda \Gamma(S, w_{-S}) \right)^2$$

with

$$\underbrace{\Gamma^2(q_S, w_{-S})}_{\text{“effective sparsity”}} \leq C_2^2 \sum_{j=1}^{s+1} \frac{n \log(n_j).}{n_j^3}.$$

Example
total variation penalty on higher order differences

Fix $k \in \mathbb{N}$.

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

with

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

Let

- $S := \{t_1, \ldots, t_s\}$, $t_1 < \cdots < t_s$, $t_0 := N$, $t_{s+1} := n+1$,
- $n_j := t_j - t_{j-1}$, $j = 1, \ldots, s+1$,
- $n_{\max} := \max_{1 \le j \le s+1} n_j$

$$\mathcal{V}_N := \mathcal{N}_{-S} \cup \{\psi_{t_j+1}, \ldots, \psi_{t_j+k-1}\}_{j=1}^{s}$$

$$N = r_S + (k-1)s = k(s+1)$$

$$\gamma_S(N, \Psi) \le A n_{\max}^{\frac{2k-1}{2}} / \sqrt{n} \qquad \left(1/W = \frac{2k-1}{2}\right)$$

Application of the theorem to the total variation penalty on $\Delta^{k-1} f$:

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

$$\lambda \geq c_k n^{k-1} \lambda_0(u) \sqrt{n_{\max}^{2k-1}/n}$$

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

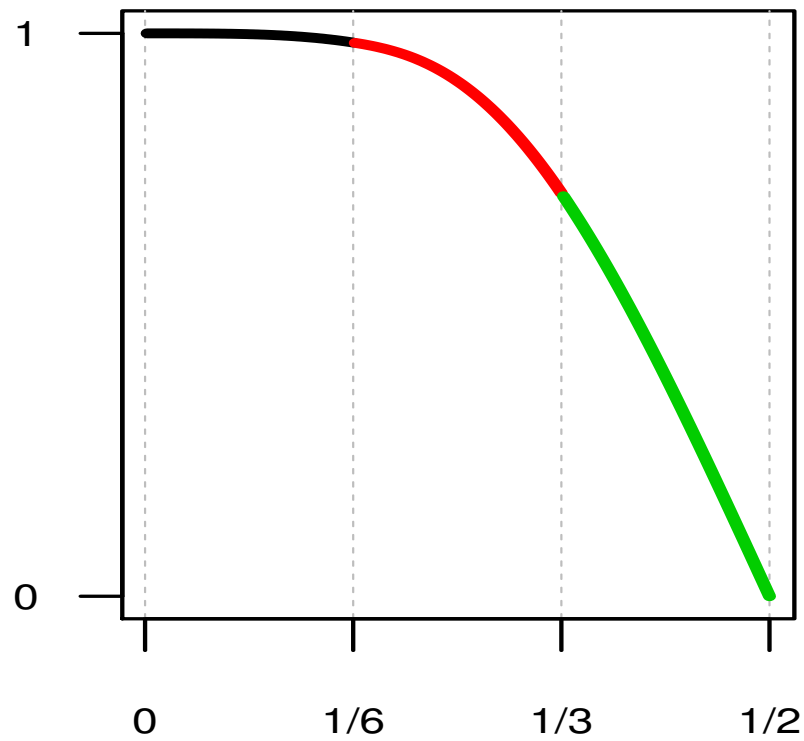
$$\|\hat{f} - f^0\|_2^2/n \leq \|f^* - f^0\|_2^2/n$$

$$+ \left(\sqrt{\frac{k(s+1)}{n}} + \sqrt{\frac{2v}{n}} + \lambda \Gamma(q_S, w_{-S}) \right)^2$$

with

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

Interpolating vector for $k = 4$



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

$$\|\hat{f} - f^0\|_2^2/n = \mathcal{O}_{\mathbb{P}}\left(\frac{(s_0+1)\log^2 n}{n}\right)$$

provided

$$\lambda \asymp n^{k-1}\sqrt{\frac{\log n}{n}}\left(\frac{1}{s_0+1}\right)^{\frac{2k-1}{2}}.$$

Example
Total variation in higher dimensions

[Fang, Guntuboyina & Sen; 2019]

Let

$$f \in \mathbb{R}^{n_1 \times n_2} \text{ (a matrix)}$$

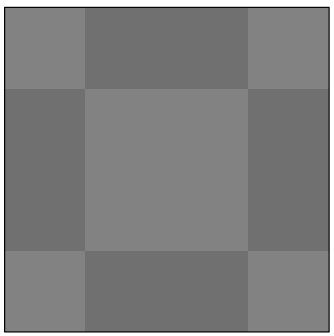
and

$$\|D_1 f D_2^\top\|_1 := \sum_{j=2}^{n_1} \sum_{k=2}^{n_2} |f_{j,k} - f_{j-1,k} - f_{j,k-1} + f_{j-1,k-1}|$$

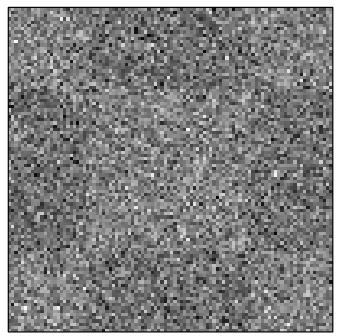
Hardy-Krause variation

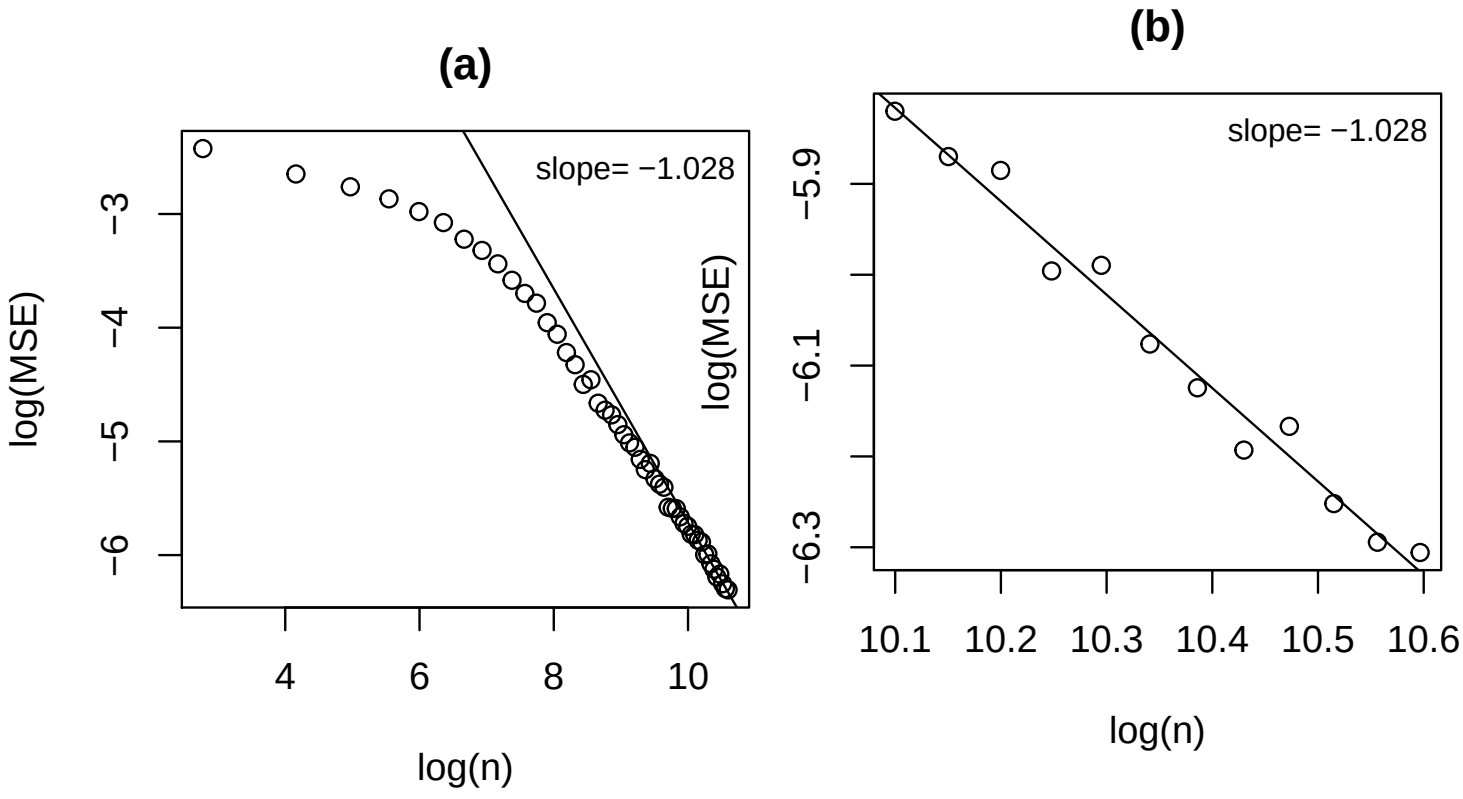
Useful for pictures $\cdots$

(a)



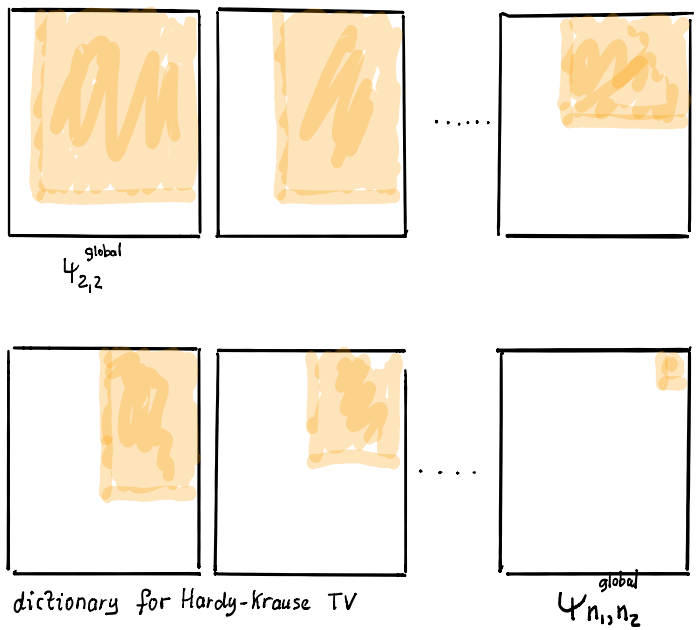
(b)



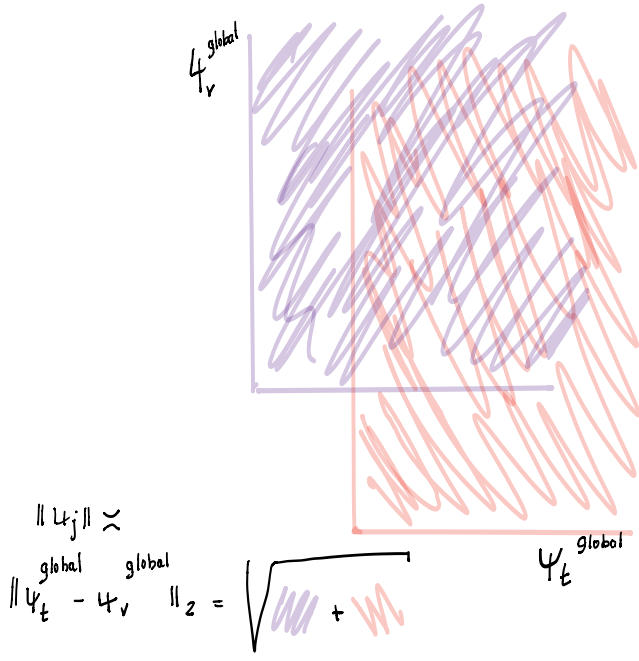

(a)
slope= −1.028
log(MSE)
log(MSE)
−3
−4
−5
−6
4
6
8
10
log(n)
(b)
slope= −1.028
−5.9
−6.1
−6.3
10.1
10.2
10.3
10.4
10.5
10.6
log(n)

Tuning and noise

dictionary

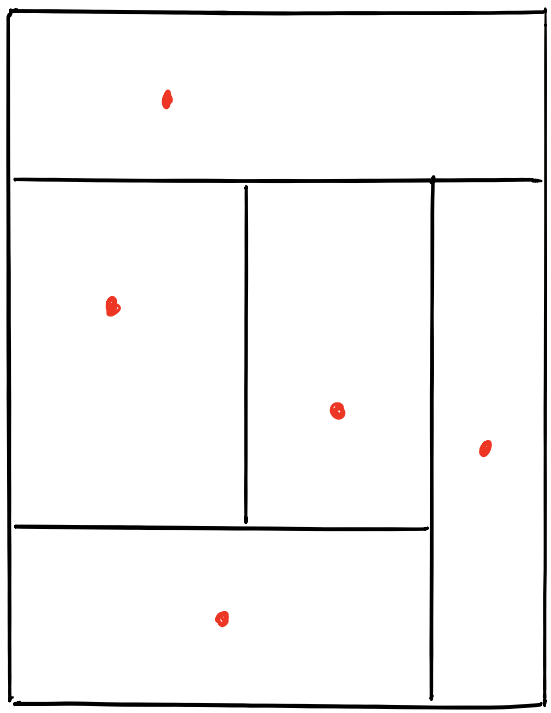


projections

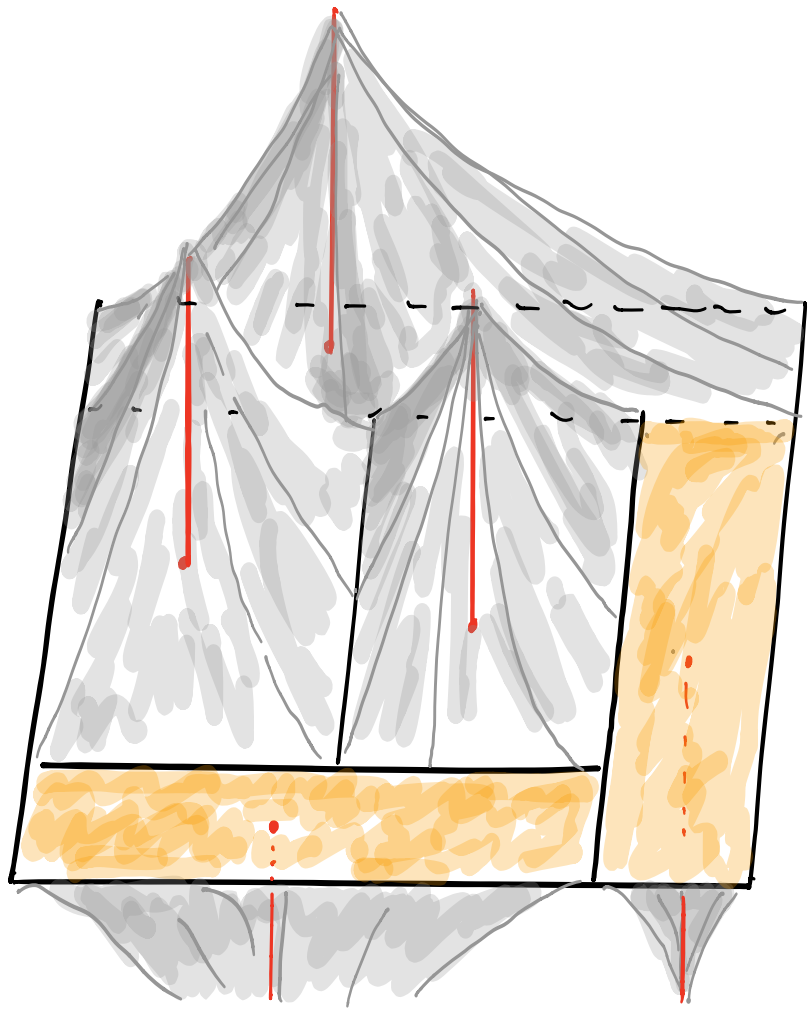


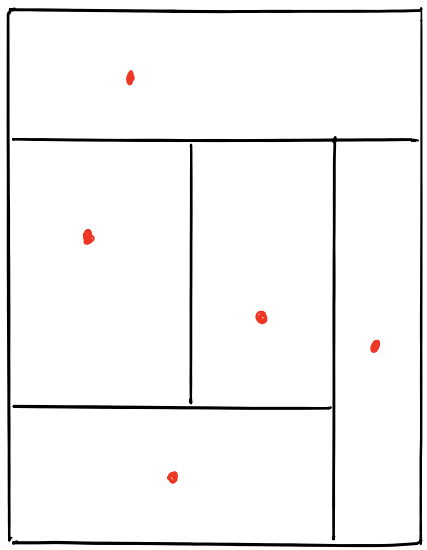
Effective sparsity

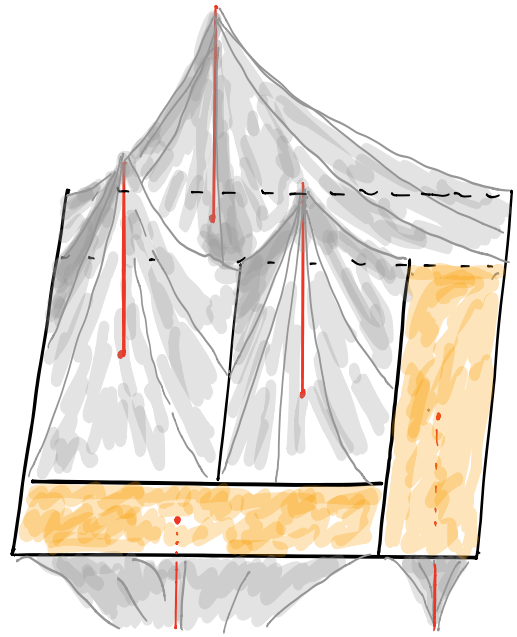
Let $\{t_1, \ldots, t_s\} \subset \mathbb{R}^2$ be the locations of the jumps. Construct a tiling of the rectangle $[1 : n_1] \times [1 : n_2]$ into s sub-rectangles, such that each sub-rectangle contains one jump.



Interpolating vector q





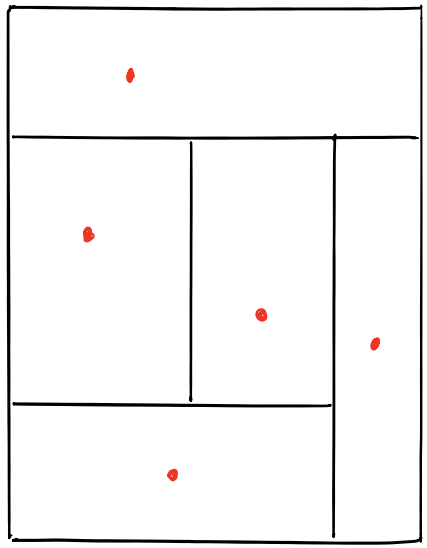


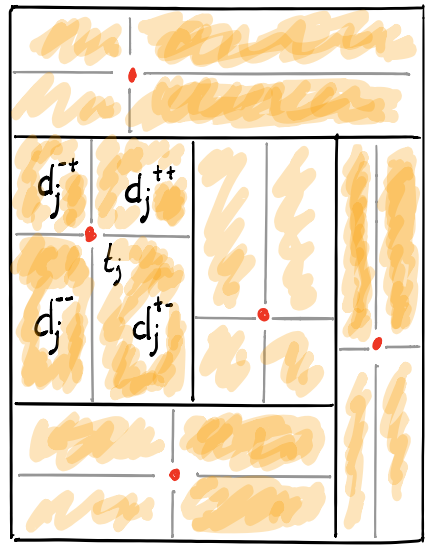
Let $(n_j^{--}, n_j^{+-}, n_j^{-+}, n_j^{++})$ contain the areas between t_j and the corners of its sub-rectangle, $j = 1, \ldots, s$.
Then

$$\Gamma^2(q_S, w_{-S}) \leq n\|D_1^\top q D_2\|_2^2$$

$$\leq \sum_{j=1}^{s} \left(\frac{n}{n_j^{--}} + \frac{n}{n_j^{+-}} + \frac{n}{n_j^{-+}} + \frac{n}{n_j^{++}} \right) \left(\log(1 + n_1) + \log(1 + n_2) \right)$$

The areas $(n_j^{--}, n_j^{+-}, n_j^{-+}, n_j^{++})$, $j = 1, \ldots, s$.





Application of the theorem to the 2dim total variation penalty

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

$$\lambda \geq 2\sqrt{\frac{2(\log(2n) + 2u)}{n}}\sqrt{n_{1,\max}/n_1 + n_{2,\max}/n_2}$$

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

$$\|\hat{f} - f^0\|_2^2/n \leq \|f^* - f^0\|_2^2/n$$

$$+\left(\sqrt{\frac{(s+1)}{n}} + \sqrt{\frac{2v}{n}} + \lambda\Gamma(q_S, w_{-S})\right)^2$$

with

$$\underbrace{\Gamma^2(q_S, w_{-S})}_{\text{“effective sparsity”}}$$

$$\leq \sum_{j=1}^{s}\left(\frac{n}{n_j^{--}} + \frac{n}{n_j^{+-}} + \frac{n}{n_j^{-+}} + \frac{n}{n_j^{++}}\right)\left(\log(1+n_1) + \log(1+n_2)\right)$$

Corollary *Suppose the jumps are roughly on a regular grid. Then the rate of convergence is*

$$\|\hat{f} - f^0\|_2^2/n = \mathcal{O}_{\mathbb{P}}\left(\frac{(s_0+1)\log^2 n}{n}\right)\sqrt{s_0+1}$$

provided

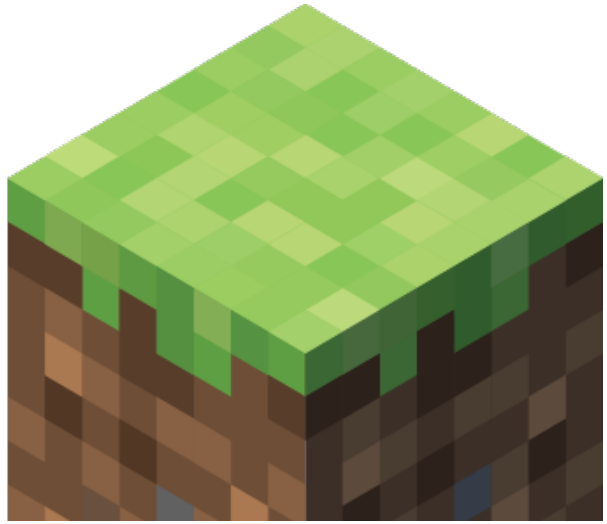
$$\lambda \asymp \sqrt{\frac{\log n}{n}}\left(\frac{1}{s_0+1}\right)^{\frac{1}{4}}$$

Remark The extra factor is

$$\frac{\text{distance in } \mathbb{R}^d}{\text{area in } \mathbb{R}^d} = \frac{\left(\frac{1}{s_0+1}\right)^{\frac{1}{d}}}{\frac{1}{s_0+1}}$$

$$= \begin{cases} 1 & \text{for } d=1 \\ \sqrt{s_0+1} & \text{for } d=2 \end{cases}$$

Example: Total variation in dimension d



Application of the theorem ...

Corollary *Suppose the jumps are roughly on a regular grid. Then the rate of convergence is*

$$\|\hat{f} - f^0\|_2^2/n = \mathcal{O}_{\mathbb{P}}\left(\frac{(s_0+1)\log^2 n}{n}\right)(s_0+1)^{1-\frac{1}{d}}$$

provided

$$\lambda \asymp \sqrt{\frac{\log n}{n}}\left(\frac{1}{s_0+1}\right)^{\frac{1}{2d}}$$

remark For mesh grid

$$\|\hat{f} - f^0\|_2^2/n = \mathcal{O}_{\mathbb{P}}\left(\frac{(s_0+1)\log^2 n}{n}\right)(s_0+1)^{1-\frac{1}{1+\frac{1}{2}+\cdots+\frac{1}{d}}}$$

Example: MARS in dimension d

Application of the theorem ...

Example
Total variation on graphs

$\mathcal{G} := (\mathbf{V}, \mathcal{E})$ be a connected graph,
$\mathbf{V} :=$ set of vertices
$n := |\mathbf{V}|$ # vertices
$\mathcal{E}$ edges of of $\mathcal{G}$.

Signal

$$f^0 = \{f_v^0\}_{v \in \mathbf{V}} \in \mathbb{R}^n.$$

Observations

$$Y_v = f_v^0 + \epsilon_v, \ v \in \mathbf{V}, \ \{\epsilon_v\}_{v \in \mathbf{V}} \text{ i.i.d. } \mathcal{N}(0,1).$$

Total variation

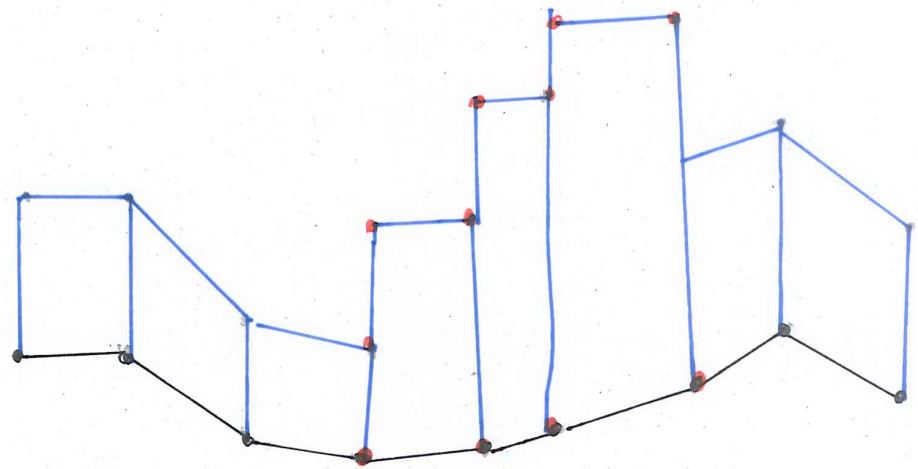
$$\mathrm{TV}(f) := \sum_{v \sim v'} |f_v - f_{v'}|, \ f = \{f_v\}_{v \in \mathbf{V}}$$

Analysis estimator

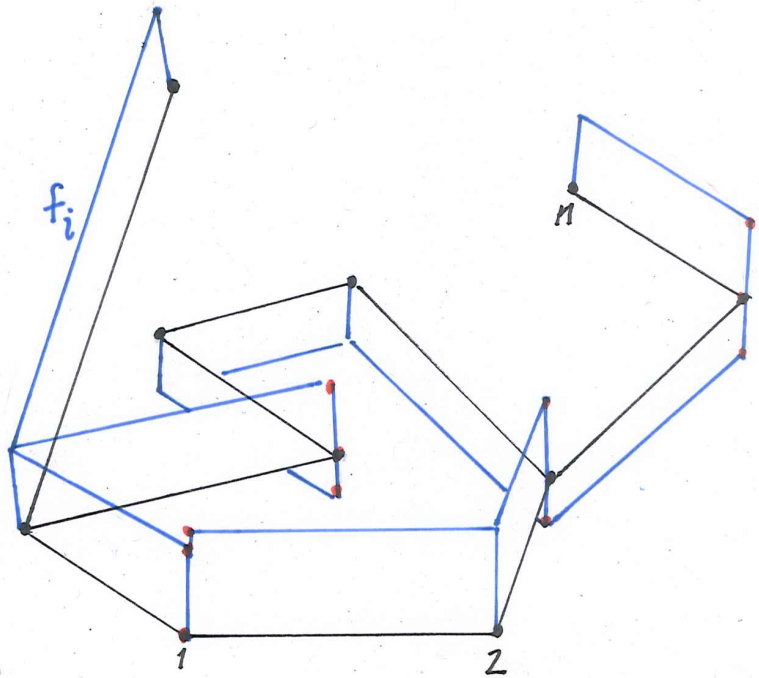
$$\hat{f} := \arg\min_f \left\{ \sum_{v \in \mathbf{V}} (Y_v - f_v)^2 / n + 2\lambda \mathrm{TV}(f) \right\}.$$

We have considered the path graph

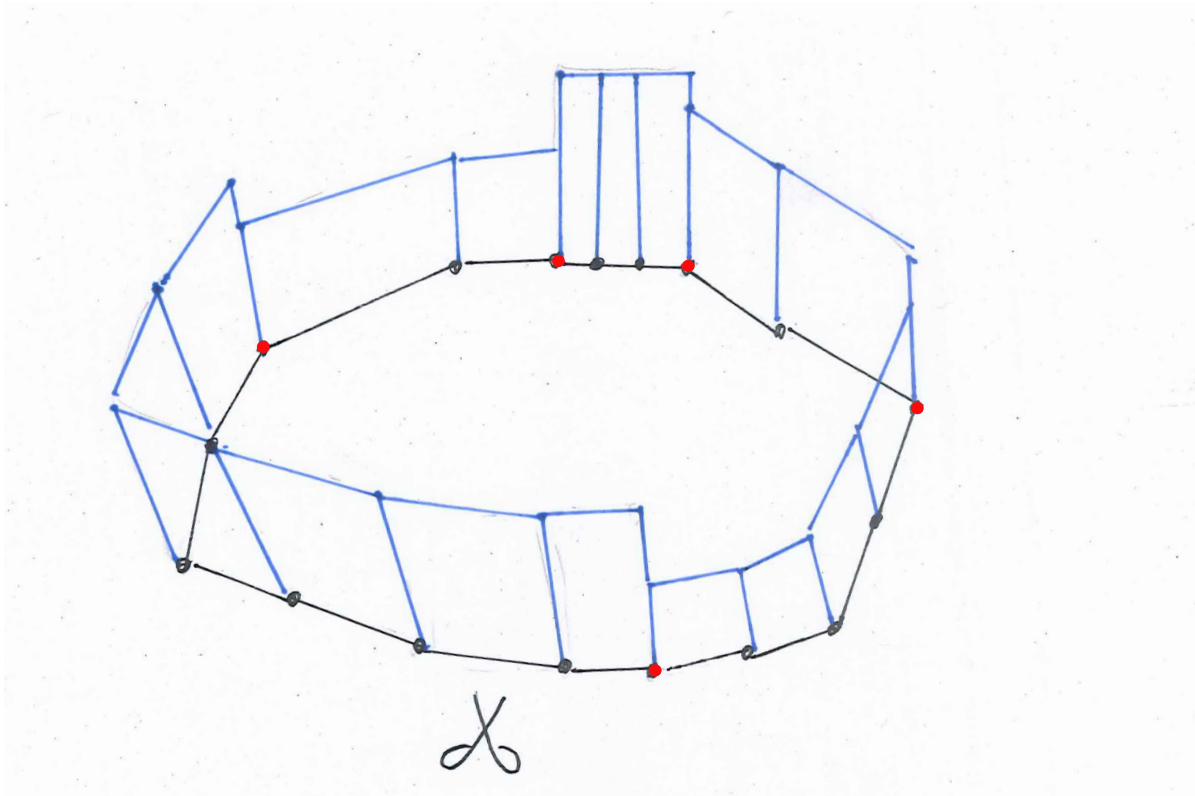
At each node v is attached a value f_v $(v = 1, \ldots, n)$



We now consider more general graphs



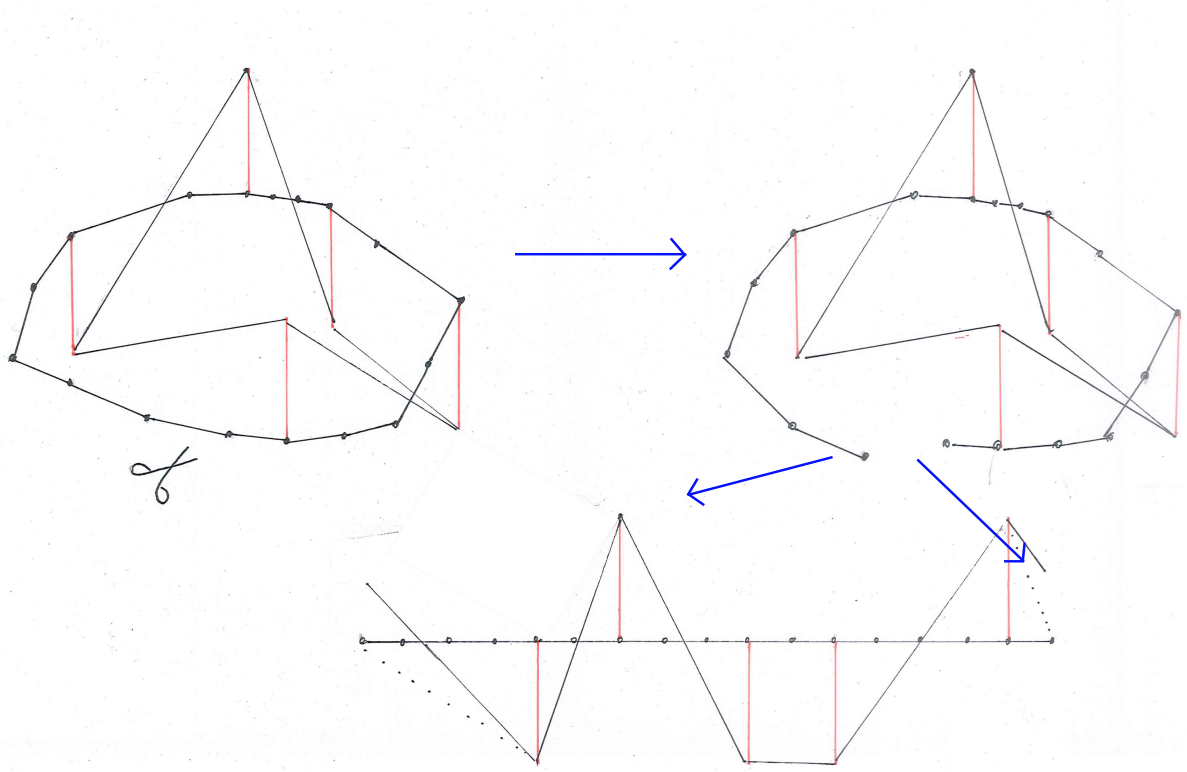
For example cycles



By cutting it we get back the path graph!

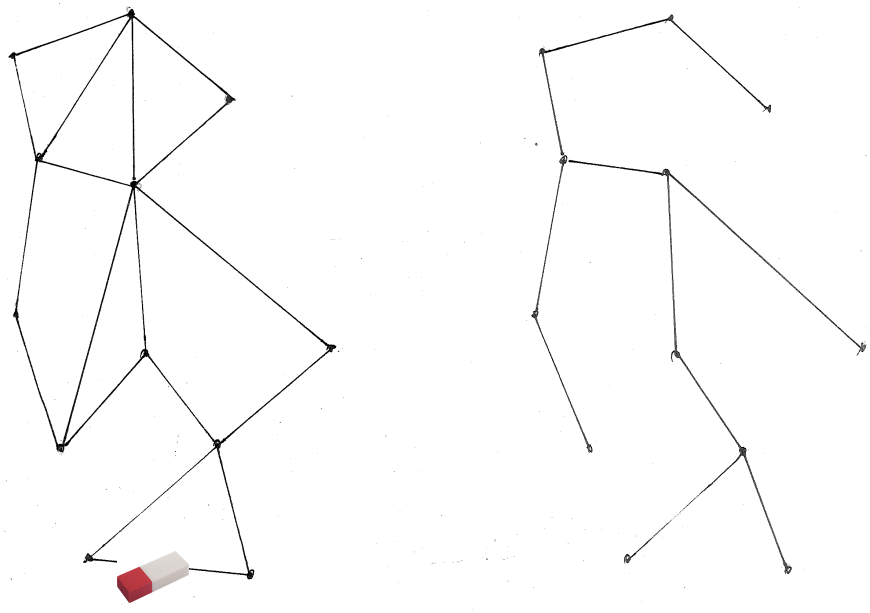
Effective sparsity without noise for graphs

The interpolating vector q for the cycle



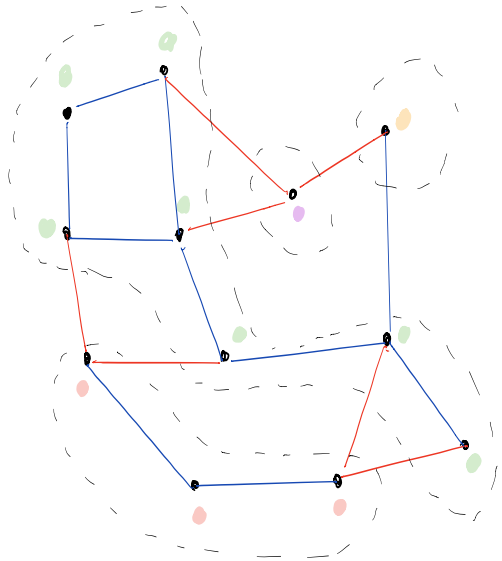
The interpolating vector is as for the path graph

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.



Tuning and noise for graphs

We now have
$f \in \mathcal{N}_{-S}$
$\Leftrightarrow$ f is constant on certain connected sub-graphs
$(f_{\mathcal{N}_{-S}})_v$ = local average within the sub-graph ,
$v \in$ sub-graph
$\rightsquigarrow$
to find the dictionary
$\{\psi_e\}_{e \notin S}$
we do a local version of centring f.



Global dictionary

Notation: $v \rightsquigarrow v' \subset \mathcal{E}$ path from v to v'

$$
\begin{aligned}
n|\psi_{v,e}^{\text{global}}| &:= \left| \left\{ v' \in \mathcal{V} : \ e \in v \rightsquigarrow v' \right\} \right| \\
&= \#\text{ paths starting at } v \\
&\quad \text{that contain the edge } e
\end{aligned}
$$

$$
v \in \mathbf{V}, \ e \in \mathcal{E}.
$$

Let

$$\bar{f} := \frac{\sum_{v' \in \mathcal{V}} f_{v'}}{n}$$

Then

$$f_v - \bar{f} = \frac{\sum_{v' \in \mathbf{V}} (f_v - f_{v'})}{n}$$

Define now for all $e := u \sim u' \in \mathcal{E}$,

$$b_e := f_u - f_{u'}.$$

Then

$$f_v - f_{v'} = \sum_{e \in v \rightsquigarrow v'} b_e.$$

and so

$$f_v - \bar{f} \quad = \quad \sum_{e \in \mathcal{E}} \psi_{v,e}^{\text{global}} \, b_e.$$

$\rightsquigarrow$ global noise weights

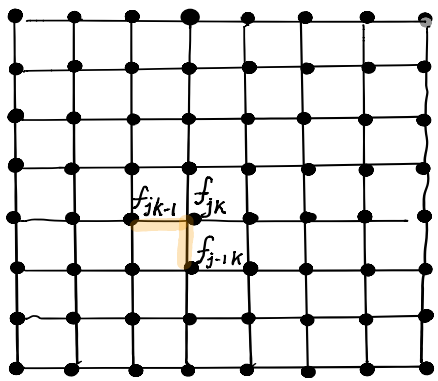
$$\|\psi_e^{\text{global}}\|_2^2 = \sum_{v \in \mathbf{V}} (\psi_{v,e}^{\text{global}})^2, \ e \in \mathcal{E}.$$

Now to choose the paths in such a way that this is small is “small".

Example
Total variation in 2d bis

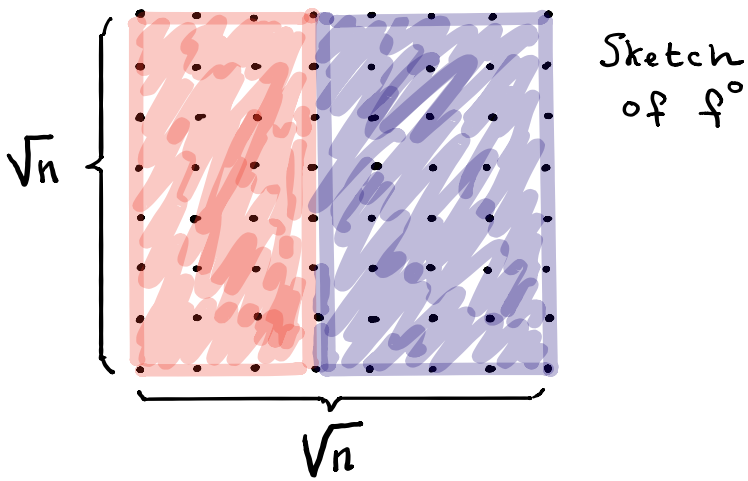
[Sadhalana, Wang & Tibshirani; 2016],
[Sadhalana, Wang, Sharpnack & Tibshirani; 2017]

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



"dimension" (f^0)=2 $\quad \stackrel{?}{\Rightarrow} \|\hat{f} - f^0\|_2^2/n \asymp \frac{2}{n}$?

jumps = $\sqrt{n}$ $\quad \stackrel{?}{\Rightarrow} \|\hat{f} - f^0\|_2^2/n \asymp \frac{\sqrt{n}}{n}$?

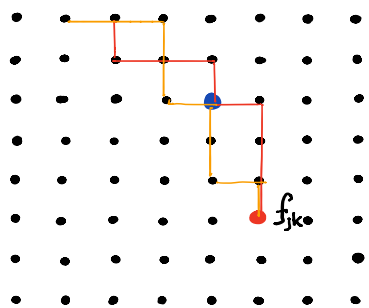


As paths from v to v' we take grid lines

$$x \mapsto y = \lfloor ax + c \rfloor$$

For example, for $v = (j, k)$ and $v' = (j', k')$ with $j' \leq j$, $k' \leq k$ we let take the path following the grid line $x \mapsto y$ with

$$k - y = \left\lfloor \left(\frac{k - k'}{j - j' + c} (j - x + c) \right) \right\rfloor .$$



Count the number
of paths
that contain the
horizontal/vertical
edge at the blue point

Lemma *Let* $v = (j, k)$, $e = (j'', k'') \sim (j'' - 1, k'')$. *Then*

$$n\psi_{v,e}^{\text{global}} \leq (j + c)\left(\frac{j''}{j - j''} \wedge \frac{k''}{k - k''}\right)$$

$$\begin{aligned} n^2 \|\psi_{v,e}^{\text{global}}\|^2 &\leq 2j''k''(n_1 - j'' + 1)(n_1 - j'' + 2c) \\ &+ 4j''^3 k'' \left(\log\left(1 + \frac{n_1 - j''}{c}\right) + \frac{1}{c}\right) \end{aligned}$$

Corollary *Taking* $c = 1$ *gives*

$$\max_{e \notin S} \|\psi_e\|_2^2 \asymp \log n.$$

$\rightsquigarrow$

tuning parameter

$$\log n/n \lesssim \lambda$$

effective sparsity

$$\Gamma^2(q_S, 1) \asymp n\sqrt{n}$$

and the (minimax [Hütter & Rigollet, 2016]) rate

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

Corollary *Taking* $c = \sqrt{n}$ *gives*

$$\max_{e \notin S} \|\psi_e\|_2^2 \asymp n^{1/4} \log n$$

and similarly for the vertical edges.

$\rightsquigarrow$

tuning parameter

$$n^{-\frac{7}{8}} \log n \lesssim \lambda$$

effective sparsity

$$\Gamma^2(q_S, w_{-S}) \asymp n \log n$$

and the (adaptive) rate

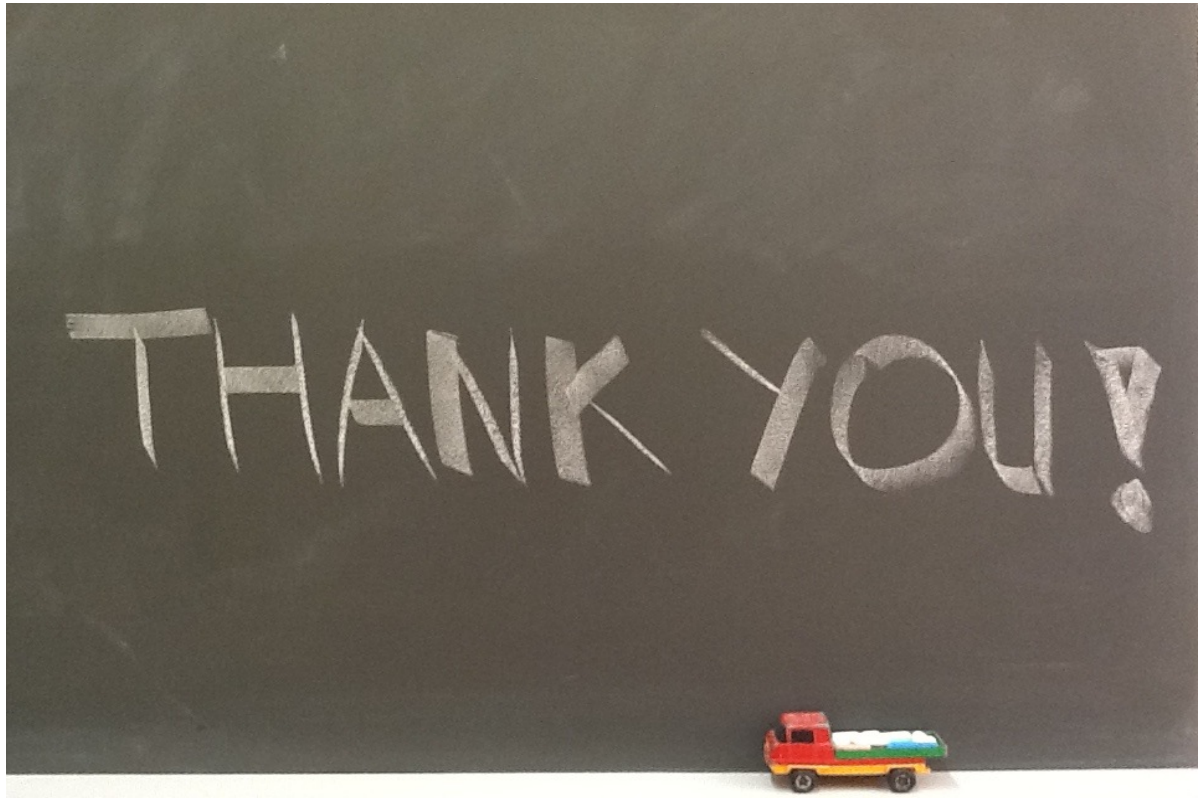
$$\|\hat{f} - f^0\|_2^2 / n = \mathcal{O}_{\mathbb{P}}\left(\frac{\log^3 n}{n^{3/4}}\right).$$

as in [Chatterjee and Goswami, 2019]

Conclusion

- “Lasso" can also work for highly correlated design
- Overparametrization vs entropy ...
- interpolating vectors useful for structured problems
  $\rightsquigarrow$ improvement of restricted eigenvalue conditions
- TV-type penalties are adaptive

Remark $\exists$ extensions to other loss functions
(e.g. logistic loss)


THANK YOU!