Divided Differences, Falling Factorials, and Discrete Splines: Another Look at Trend Filtering and Related Problems

Ryan Tibshirani
Depts. of Statistics & Machine Learning
Carnegie Mellon University

https://www.stat.cmu.edu/~ryantibs/talks/dspline-2021.pdf

Tale of two estimators

Nonparametric regression: given data

$$y_i = f_0(x_i) + \epsilon_i, \quad i = 1, \ldots, n$$

Goal is to estimate f_0 (assumed smooth). Very rich literature. This talk focuses on two estimators:

Locally adaptive spline (Mammen & van de Geer, 1997):

$$\underset{f}{\text{minimize}} \sum_{i=1}^{n} \big(y_i - f(x_i)\big)^2 + \lambda \cdot \text{TV}(D^k f)$$

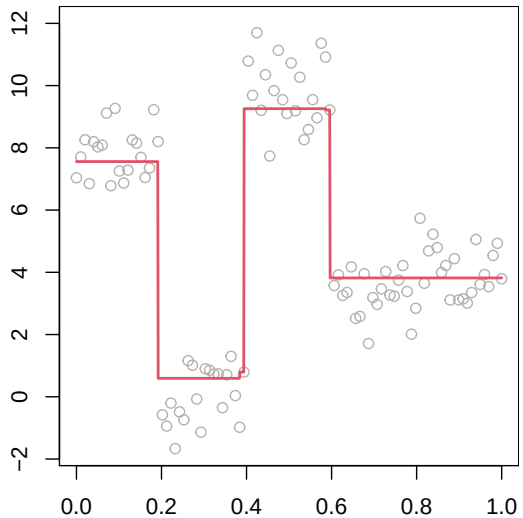
Trend filtering (Steidl et al., 2006; Kim et al., 2009):

$$\underset{\theta \in \mathbb{R}^n}{\text{minimize}} \ \|y - \theta\|_2^2 + \lambda \|D_n^{k+1} \theta\|_1$$

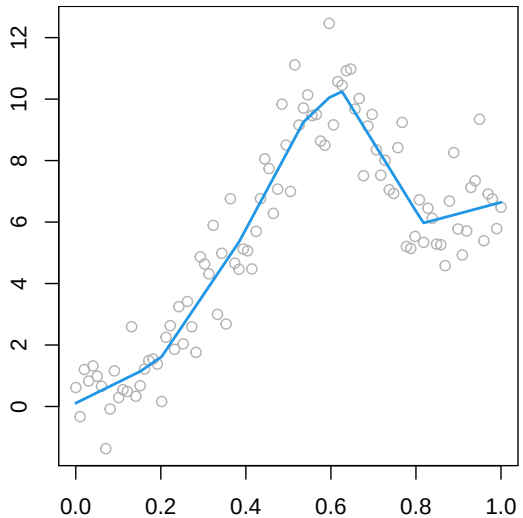
TF penalty operator

Penalty matrix D_n^{k+1} in TF is standard $(k+1)$st order (backward) difference operator

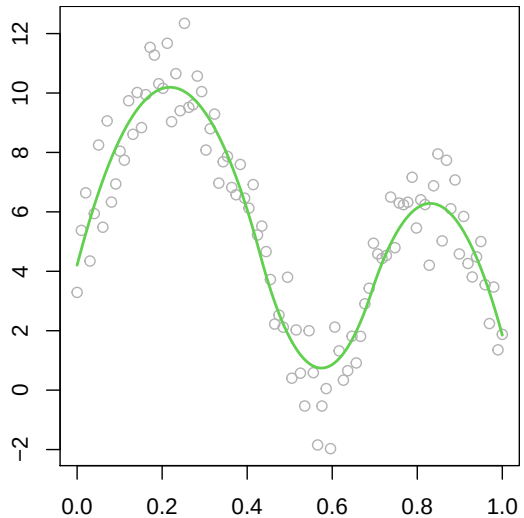
$k = 0$



$k = 1$



$k = 2$



Penalty terms:

$$|\theta_i - \theta_{i-1}|$$

$$|\theta_i - 2\theta_{i-1} + \theta_{i-2}|$$

$$|\theta_i - 3\theta_{i-1} + 3\theta_{i-2} - \theta_{i-3}|$$

What's the connection?

Let $\theta_i = f(x_i)$, $i = 1, \ldots, n$. By recursive property of differences:

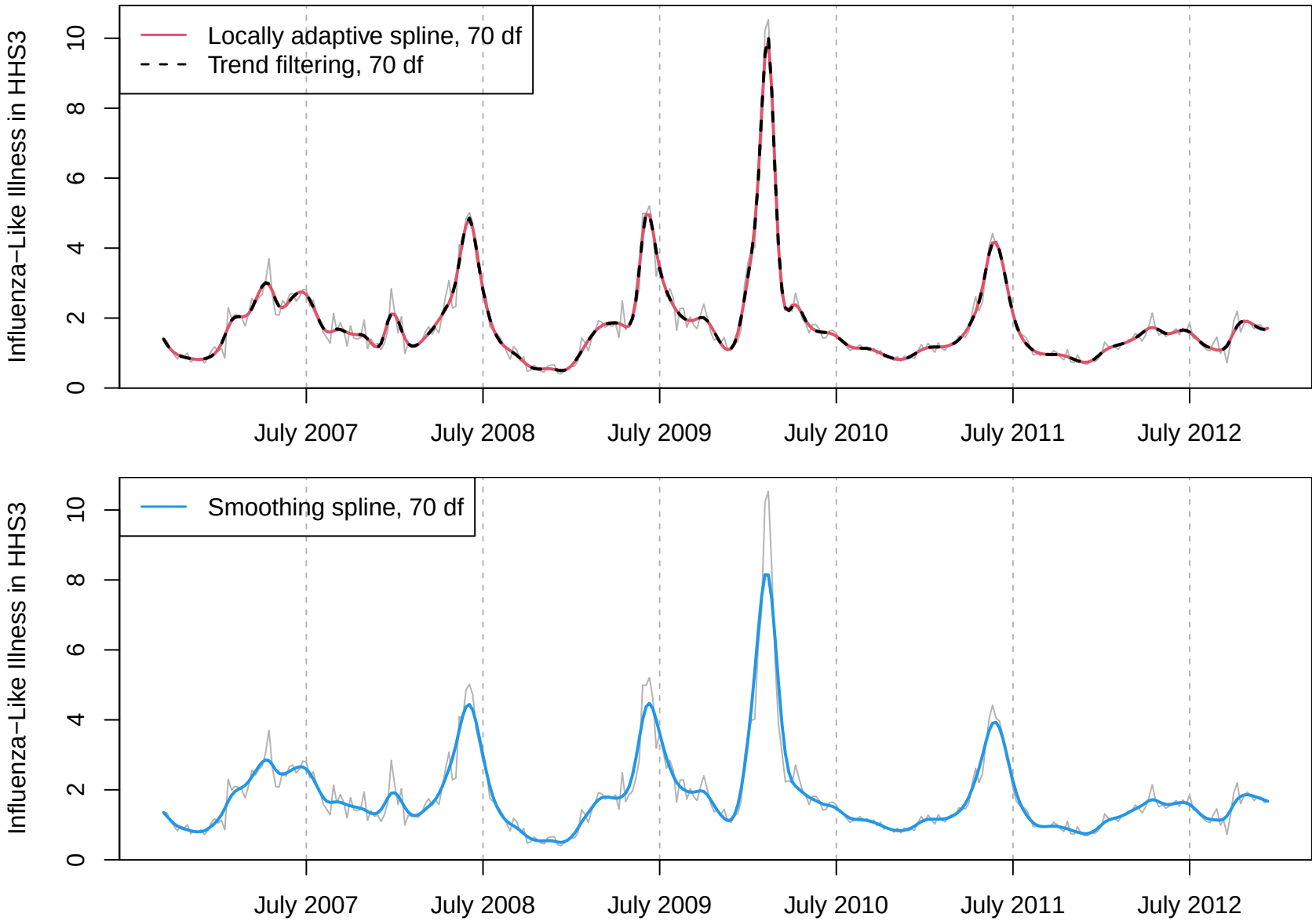
$$\begin{aligned} \|D_n^{k+1}\theta\|_1 &= \sum_{i=1}^{n-k-1} \left|(D_n^k\theta)_{i+1} - (D_n^k\theta)_i\right| \\ &\approx \sum_{i=k+2}^{n} \left|(D^k f)(x_{i+1}) - (D^k f)(x_i)\right| \\ &\approx \mathrm{TV}(D^k f) \end{aligned}$$

- Second step: kth differences of θ are $\approx$ kth derivatives of f
- Third step: TV is a supremum over all such partitions (further, this is exact when f is a kth degree piecewise polynomial with knots in $x_1, \ldots, x_n$)

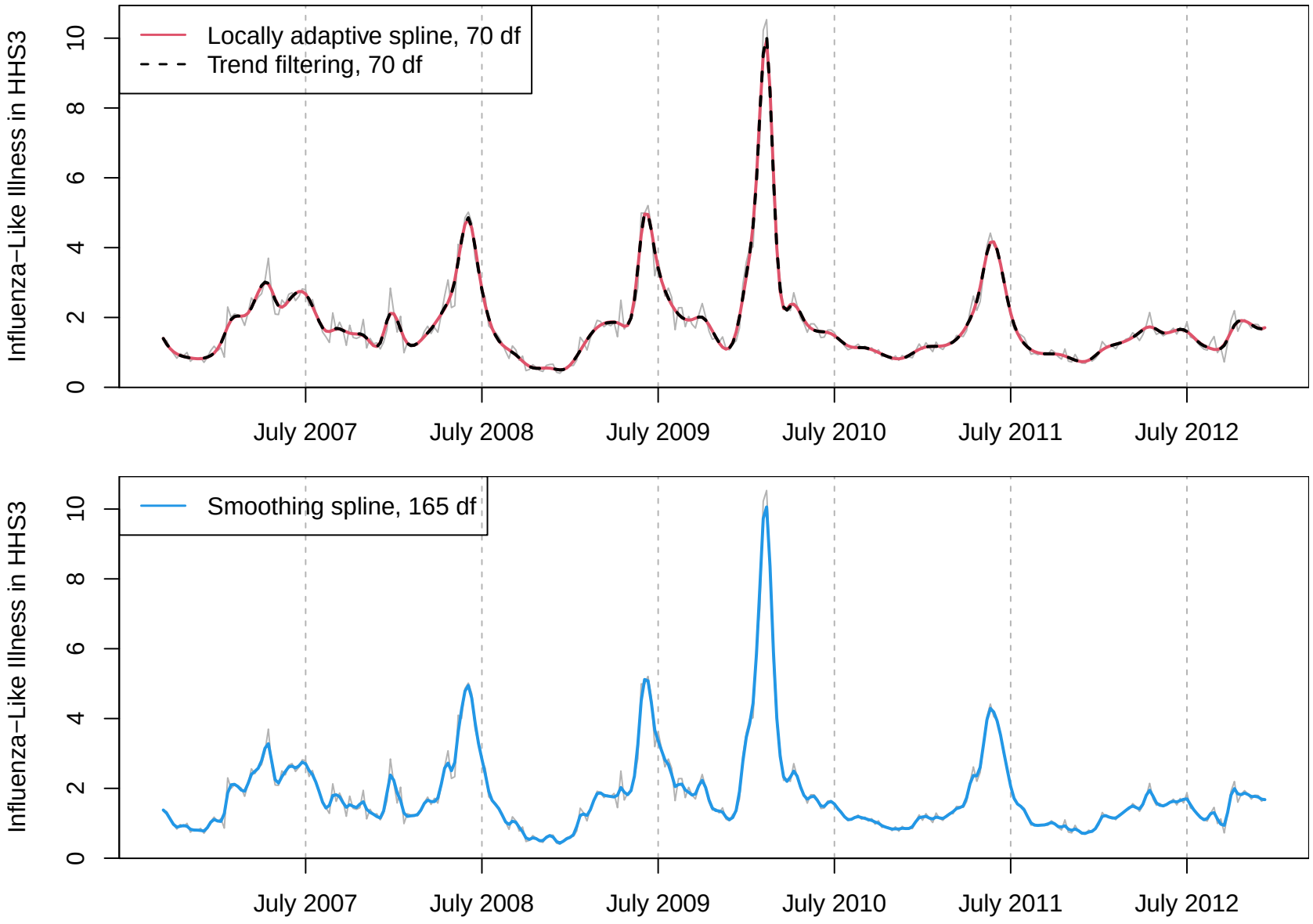
Some key properties

- Solution of locally adaptive spline problem is always kth degree spline; its knots are selected adaptively, but do not generally lie among $x_1, \ldots, x_n$
- Trend filtering has similar flavor of perfoming knot selection in kth degree spline, but done in discrete time
- Both are minimax optimal with respect to "big" function class: $\{f : \mathrm{TV}(D^k f) \leq \rho\}$
- Over this class, no linear smoother can be minimax optimal
- TF is more computationally efficient, thanks to banded D_n^{k+1} (it has bandwidth $k + 2$)

Illustrative example



Illustrative example



Digging deeper

For evenly-spaced $x_1, \ldots, x_n$, there exists functions f such that:

$$\begin{aligned} \|D_n^{k+1}\theta\|_1 &= \sum_{i=1}^{n-k-1} \big|(D_n^k\theta)_{i+1} - (D_n^k\theta)_i\big| \\ &= \sum_{i=k+2}^{n} \big|(D^k f)(x_{i+1}) - (D^k f)(x_i)\big| \\ &= \mathrm{TV}(D^k f) \end{aligned}$$

This holds for all $f \in \mathcal{H}_n^k = \mathrm{span}\{h_1^k, \ldots, h_n^k\}$, a "special" space of kth degree piecewise polynomials (not splines, general)

For arbitrary $x_1, \ldots, x_n$, extends to $\mathrm{TV}(D^k f) = \|D_n^{k+1}\theta\|_{w^k}$ for all $f \in \mathcal{H}_n^k$, a weighted ℓ_1 norm with weights $w_i^k = \frac{x_{i+k+1} - x_i}{k+1}$

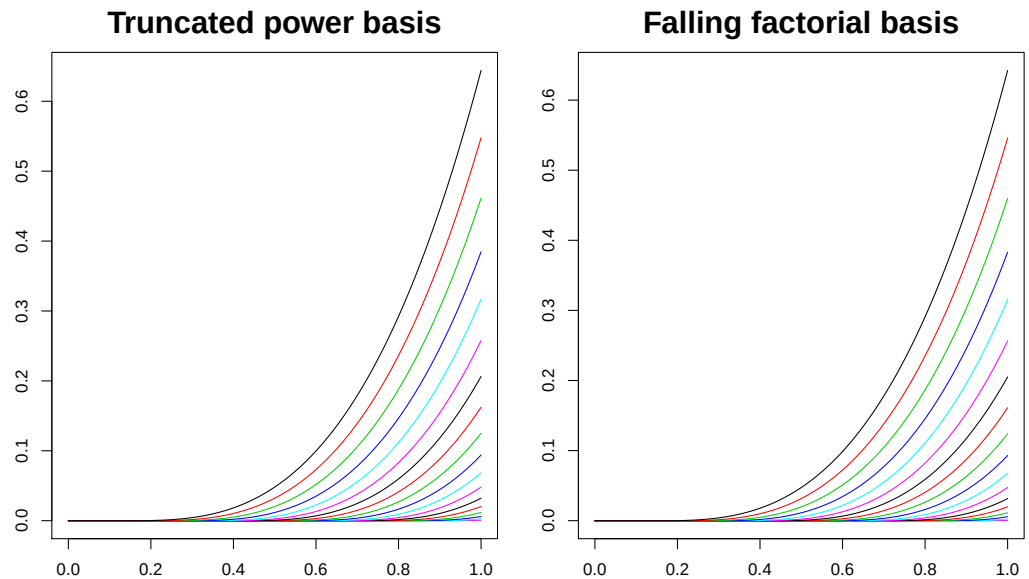
What are these functions?

These are defined by $h_i^k(x) = \frac{1}{i-1!}\prod_{\ell=1}^{i-1}(x - x_\ell)$, $i \leq k+1$, and

$$h_i^k(x) = \frac{1}{k!}\prod_{\ell=i-k+1}^{i}(x - x_\ell) \cdot 1\{x > x_i\}, \quad i \geq k+2$$

We call these the kth degree falling factorial basis functions. Note resemblance to truncated power function $g_i^k(x) = (x - x_i)_+^k/k!$

Interestingly, while h_i^k is piecewise polynomial of degree k, it is not a spline for $k \geq 2$



Discrete-continuous bridge

For $f \in \mathcal{H}_n^k$, it holds that $\mathrm{TV}(D^k f) = \|D_n^{k+1}\theta\|_{w^k}$ with $\theta_i = f(x_i)$, $i = 1, \ldots, n$, hence TF exactly solves (for arbitrary $x_1, \ldots, x_n$):

$$\underset{f \in \mathcal{H}_n^k}{\text{minimize}} \sum_{i=1}^n \big(y_i - f(x_i)\big)^2 + \lambda \cdot \mathrm{TV}(D^k f)$$

- It restricts the domain in locally adaptive spline problem to $\mathcal{H}_n^k$
- In practice, we need to restrict to the space $\mathcal{G}_n^k$, of kth degree splines with knots in $x_1, \ldots, x_n$, to "solve" this anyway
- Both $\mathcal{G}_n^k, \mathcal{H}_n^k$ are rich enough (the approximation error is small enough) that restricted solutions retain minimax rate
- When $k = 0$ or 1, something special happens: original solution lies in $\mathcal{G}_n^k$, and $\mathcal{G}_n^k = \mathcal{H}_n^k$, so all three problems are equivalent

Where to go next?

This is from T. (2014). One important consequence: gives natural way to extrapolate discrete TF solution into a bonafide function

Can extend these ideas to new domains: additive models, lattices, graphs (Wang et al., 2016; Sadhanala et al., 2016, 2017, 2018)

But is there still something to be learned in the most basic setting? Questions remain:

- why do falling factorials have this property?
- what else can we say about them?
- (especially in relation to splines ...)

Summary of rest of talk

- There is a space of piecewise polynomials called discrete splines, similar to splines, but with discrete smoothness properties
- Key property: derivatives = discrete derivatives, everywhere
- (Corollary: TV = discrete TV)
- Several other notable properties:
  - Dual basis
  - Fast interpolation
  - Fast matrix operations
  - Discrete B-spline basis

  and representation and approximation properties (not covered)
- They are the reason why trend filtering "works". May be useful for discretizing other variational problems, differential equations

Divided differences

Given a function f, its divided difference at points z_1, z_2 is

$$f[z_1, z_2] = \frac{f(z_2) - f(z_1)}{z_2 - z_1}$$

and its kth order divided difference at points $z_1, \ldots, z_{k+1}$ is

$$f[z_1, \ldots, z_{k+1}] = \frac{f[z_2, \ldots, z_{k+1}] - f[z_1, \ldots, z_k]}{z_{k+1} - z_1}$$

Special case: for unit-spaced points, simply get (scaled) differences:

$$\begin{aligned} f[a-1, a] &= f(a) - f(a-1) \\ 2! \cdot f[a-2, a-1, a] &= f(a) - 2f(a-1) + f(a-2) \\ 3! \cdot f[a-3, a-2, a-1, a] &= f(a) - 3f(a-1) + 3f(a-2) - f(a-3) \end{aligned}$$

Discrete differentation

There is a lot of old and beautiful literature on divided differences. We are going to use them to define a discrete derivative operator

Given $x_1 < \cdots < x_n$ and evaluations of f at these points. At any point x, given $f(x)$, how can we approximate $(D^k f)(x)$?

Let i be largest index so that $x_{i+k-1} < x$. We define kth discrete derivative of f at x by

$$(\Delta_n^k f)(x) = k! \cdot f[x_i, \ldots, x_{i+k-1}, x]$$

Notes:

- Centers used for divided difference vary with x
- Δ_n^k depends on design points x_i, $i = 1, \ldots, n$
- $\Delta_n^k f$ is a function, defined over continuum x

Motivating question

Classical numerical analysis: for generic smooth f, bound the error between discrete derivatives and derivatives

Flip this around: what functions f have discrete derivatives exactly equal to their derivatives?

Preliminary answer: polynomials. For $p(x) = x^k/k!$, can check that

$$(\Delta_n^k p)(x) = (D^k p)(x) = 1, \quad \text{for all } x$$

Similar for general $p(x) = \sum_{i=0}^k a_i x^i$. Beyond this? Not easy ... For example, for truncated power function $f(x) = (x-t)_+^k/k!$ we have

$$(\Delta_n^k f)(x) \neq (D^k f)(x), \quad \text{for } x \text{ near } t$$

Discrete integration

A constructive strategy to find which functions satisfy this property: examine $S_n^k = (\Delta_n^k)^{-1}$, which turns out to be a discrete integrator, based on cumulative weighted sums

Theorem. Let $1_{x_i}(x) = 1\{x > x_i\}$. For $i \geq k+2$, we have

$$(S_n^k 1_{x_i})(x) = \underbrace{\frac{1}{k!} \prod_{\ell=i-k+1}^{i} (x - x_\ell)}_{h_i^k(x)} \cdot 1\{x > x_i\},$$

a kth degree piecewise polynomial. Note that, by construction,

$$(\Delta_n^k h_i^k)(x) = 1\{x > x_i\} = (D^k h_i^k)(x), \quad \text{for all } x.$$

Proof: based on a special recursion satisfied by falling factorials

Matching derivatives

Corollary. By linearity, any $f \in \mathcal{H}_n^k$ has the same property:

$$(\Delta_n^k f)(x) = (D^k f)(x), \quad \text{for all } x.$$

This matching derivatives property implies the previous TV result in trend filtering: D_n^k is defined to produce discrete derivatives, thus

$$\begin{aligned} \|D_n^{k+1}\theta\|_{w^k} &= \sum_{i=1}^{n-k-1} \big|(D_n^k\theta)_{i+1} - (D_n^k\theta)_i\big| \\ &= \sum_{i=k+2}^{n} \big|(D^k f)(x_{i+1}) - (D^k f)(x_i)\big| = \mathrm{TV}(D^k f) \end{aligned}$$

(Several other interesting discrete-continuous relations hold as well)

Discrete splines

Further, it turns out that functions in $\mathcal{H}_n^k$ have nice structure: this is a space of discrete splines!

Theorem. The space $\mathcal{H}_n^k$ contains all piecewise polynomials f of degree k, with knots in $x_{k+1}, \ldots, x_{n-1}$, such that

$$(\Delta_n^\ell p_{i-1})(x_i) = (\Delta_n^\ell p_i)(x_i), \quad \ell = 0, \ldots, k-1$$

at each knot x_i, where p_{i-1} and p_i are the restriction of f to the interval on the left and right of x_i, respectively.

Discrete splines were first studied in numerical analysis in the early 1970s, but viewed as discrete objects. We treat them as continuum objects (i.e., functions), truly "on par" with splines

Dual basis

The matching derivatives property also allows us to define a natural dual basis to $h_1^k, \dots h_n^k$:

$$\lambda_i^k f = (\Delta_n^{k+1} f)(x_i), \quad i = 1, \dots, k+1,$$
$$\lambda_i^k f = (\Delta_n^{k+1} f)(x_i) \cdot \frac{x_i - x_{i-k-1}}{k+1}, \quad i = k+2, \dots, n.$$

These are just (scaled) discrete derivatives of f at the design points

Proposition. The basis above is a dual basis to the kth degree falling factorial basis, in the sense that for all i, j,

$$\lambda_i^k h_j^k = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise.} \end{cases}$$

Fast interpolation

Note that if

$$f = \sum_{i=1}^{n} \alpha_i h_i^k$$

then $\alpha_i = \lambda_i^k f$, a scaled version of $(\Delta_n^k f)(x_i)$

Corollary. For $f \in \mathcal{H}_n^k$, given $f(x_i)$, $i = 1, \ldots, n$,

$$f(x) = \sum_{i=1}^{k+1} (\Delta_n^{k+1} f)(x_i) \cdot h_i^k(x) + \sum_{i=k+2}^{n} (\Delta_n^{k+1} f)(x_i) \cdot \frac{x_i - x_{i-k-1}}{k+1} \cdot h_i^k(x).$$

This is an extension of Newton's divided difference formula (actually, even faster interpolation is possible: constant-time!)

Matrix computations

Recall that we define the matrix D_n^k to produce discrete derivatives. We can also define a basis matrix, with entries:

$$(H_n^k)_{ij} = h_j^k(x_i)$$

Dual basis result says matrices D_n^k, H_n^k are effectively inverses. This, and key recursion obeyed by falling factorials implies the following

> Proposition. Multiplication by $H_n^k, (H_n^k)^\mathsf{T}, (H_n^k)^{-1}$ or $(H_n^k)^{-\mathsf{T}}$ takes at most $4nk$ flops each.

Computations are simple (based on cumulative sums and iterated differences) and in-place (no extra memory required)

Discrete B-splines

Abbreviate $H = H_n^k$, and let H_T denote a submatrix whose columns correspond to knot set T. Consider the least squares projection:

$$\hat{y} = H_T H_T^\dagger y$$

onto discrete spline space with knots in T. Dual basis result implies:

$$\hat{y} = (I - B_{-T}^\dagger B_{-T}) y$$

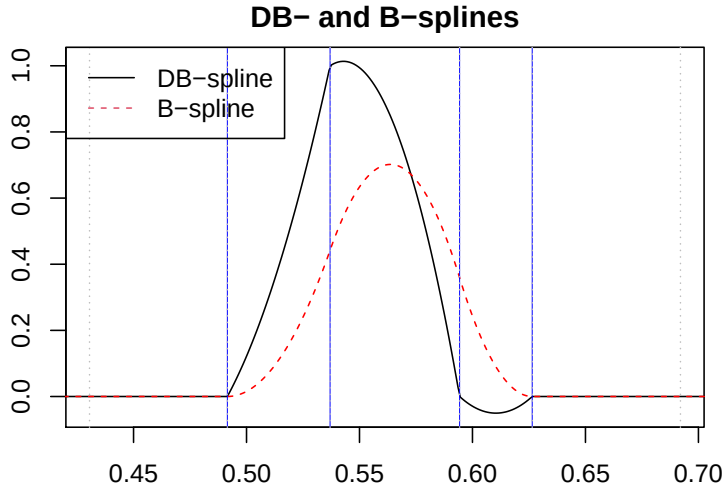
where B is basically a scaled version of D_n^{k+1}, and B_{-T} denotes an appropriate submatrix of rows. Third way:

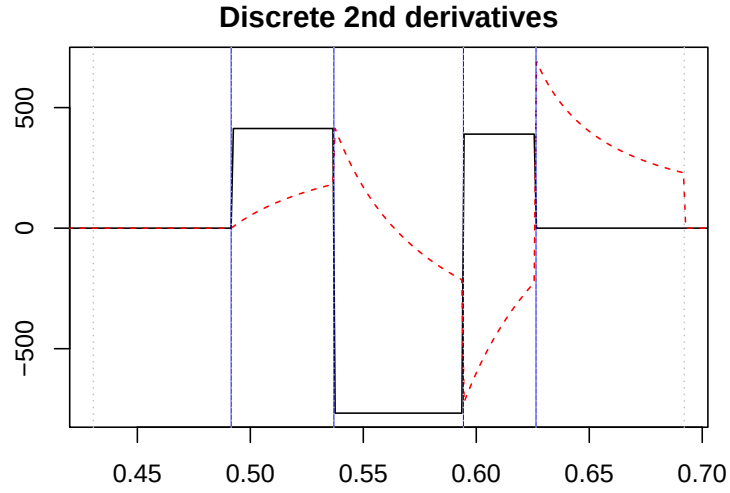
$$\hat{y} = N_T N_T^\dagger y$$

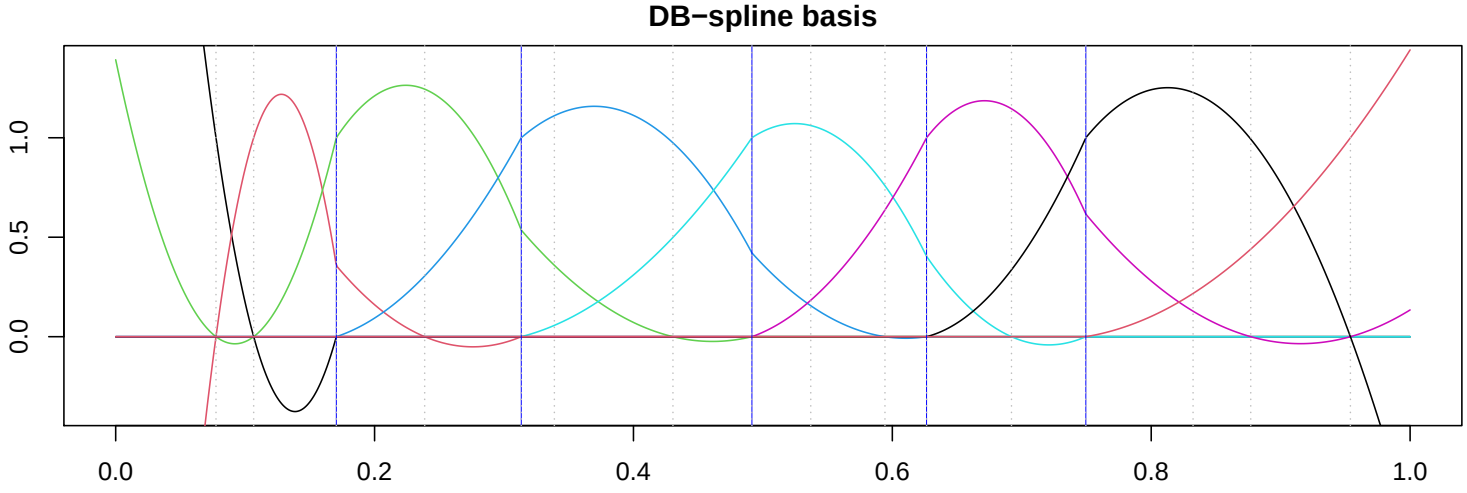
where columns of N_T are evaluations of discrete B-spline basis that spans the space of discrete splines with knots in T

Discrete B-splines (cont.)

Discrete B-splines have minimal support: each one has $k + 2$ knots

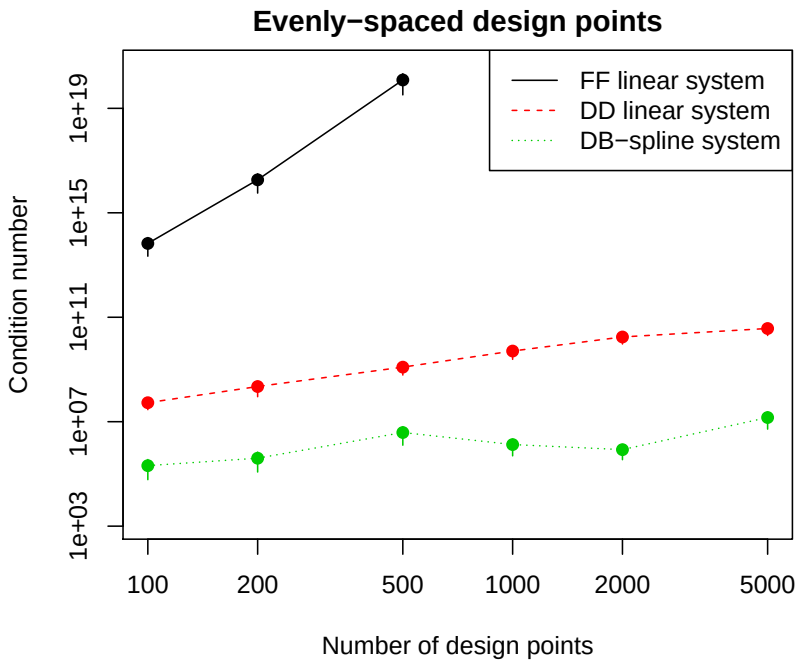


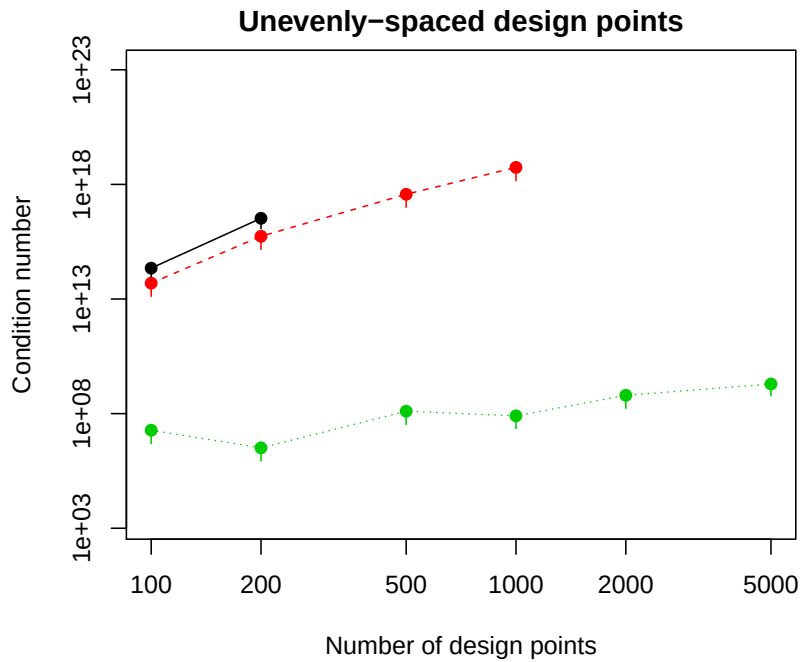




Discrete B-splines (cont.)

Thus solving a linear system in N_T is linear-time: $O(nk^2)$. Same as linear system in H_T, B_T, but discrete B-splines are more stable





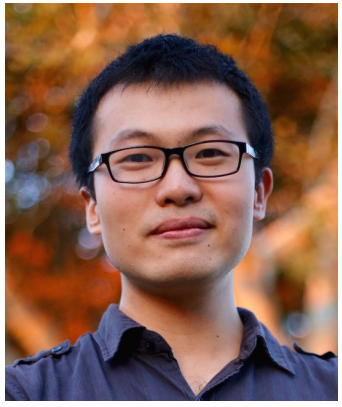
Summary of this talk

- There is a space of piecewise polynomials called discrete splines, similar to splines, but with discrete smoothness properties
- Key property: derivatives = discrete derivatives, everywhere
- (Corollary: TV = discrete TV)
- Several other notable properties:
  - Dual basis
  - Fast interpolation
  - Fast matrix operations
  - Discrete B-spline basis

  and representation and approximation properties (not covered)
- They are the reason why trend filtering "works". May be useful for discretizing other variational problems, differential equations

Acknowledgements

Veeranjaneyulu Sadhanala



Yu-Xiang Wang

`https://www.stat.cmu.edu/~ryantibs/papers/dspline.pdf`

`https://www.stat.cmu.edu/~ryantibs/talks/dspline-2021.pdf`

Thank you for listening!

Appendix

Illustration of theory for $d = 1$, general k

(Donoho & Johstone, 1998; T., 2014)

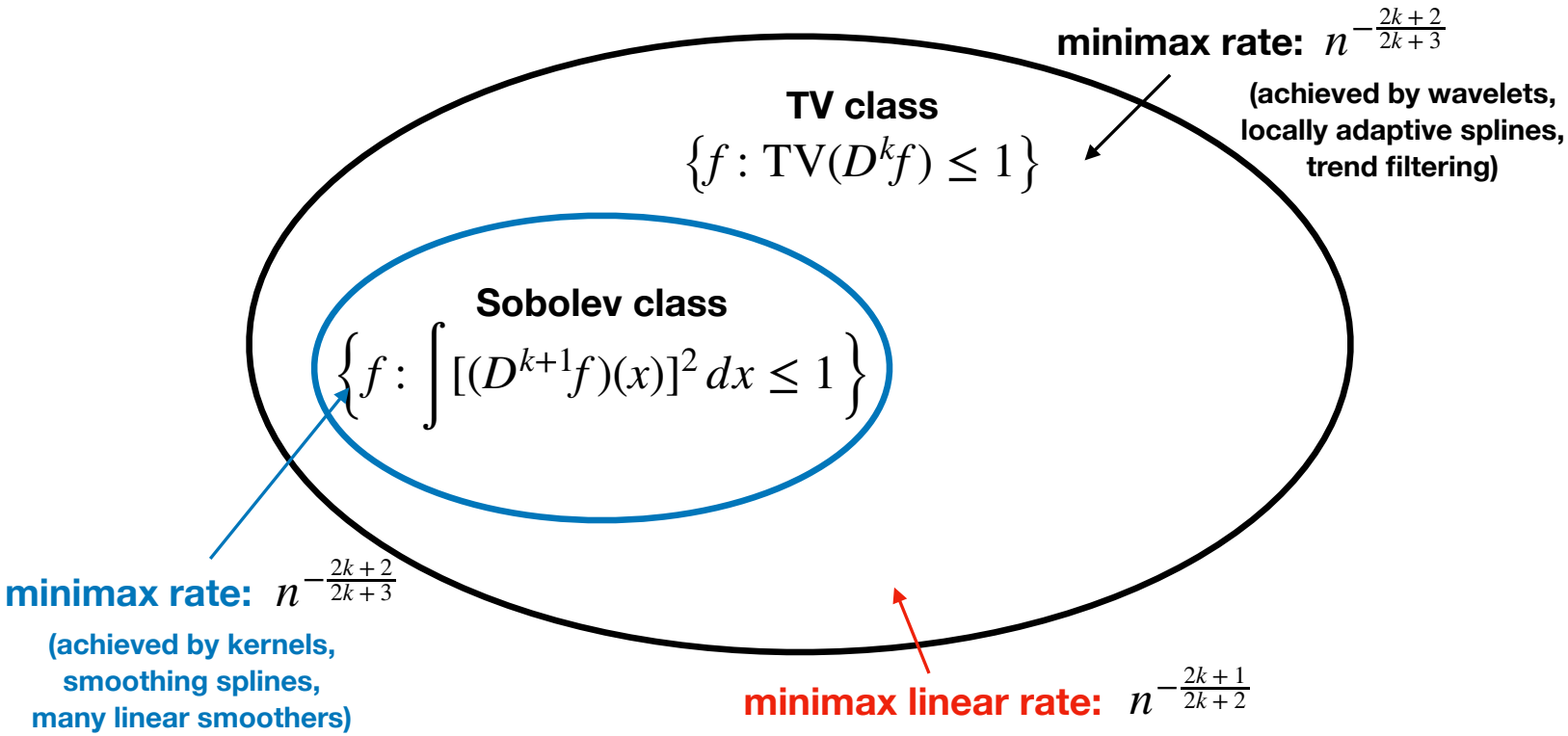


Illustration of theory for $k = 0$, general d

(Sadhanala, Wang, T., 2016)

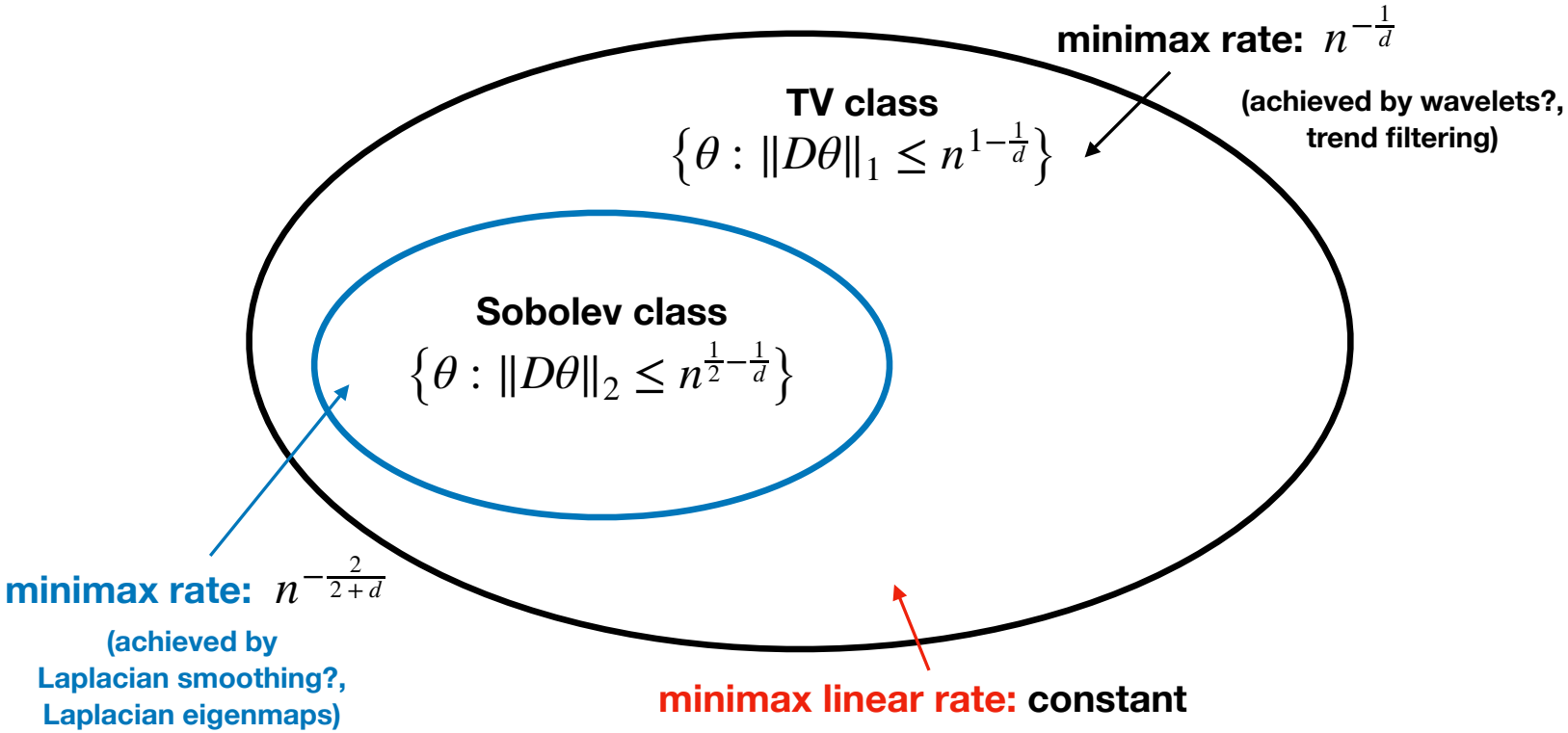
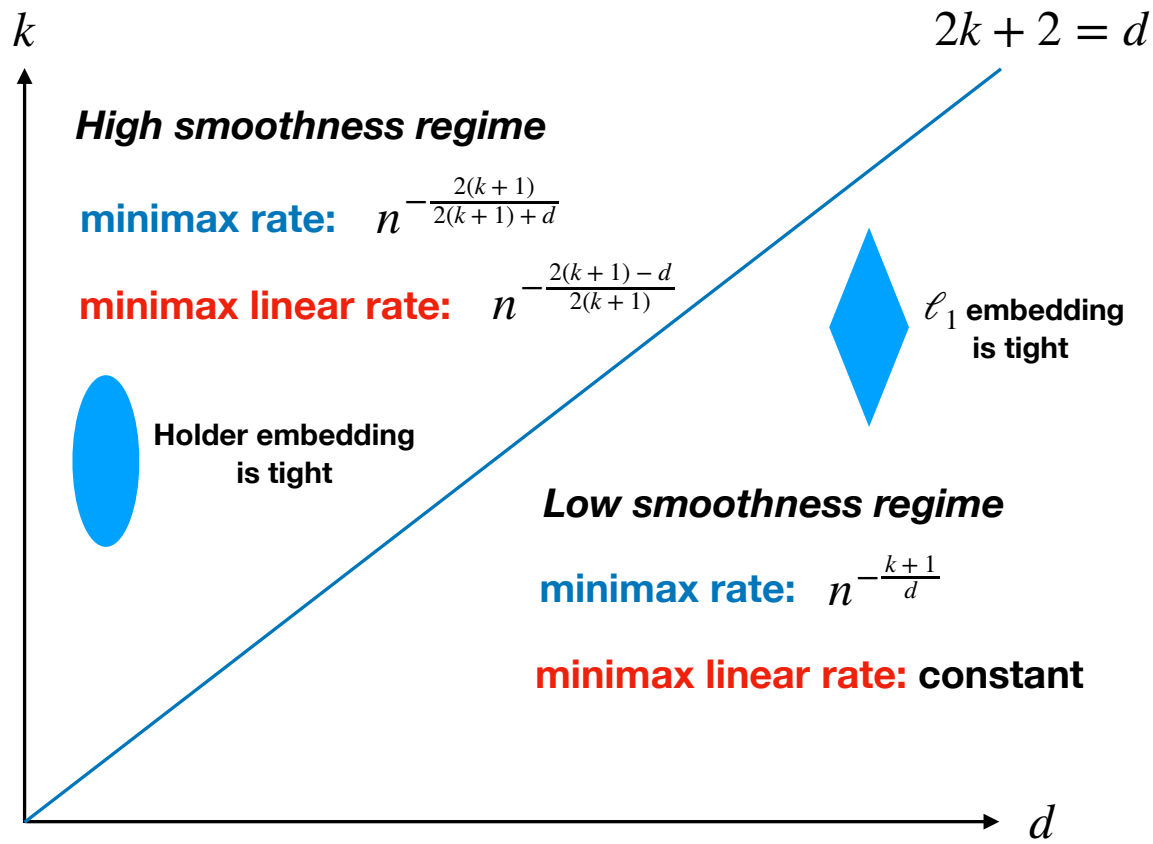


Illustration of theory for general k, d

(Sadhanala, Wang, Sharpnack T., 2017, 2021)



Representation of Sobolev functionals

Theorem. For any $f \in \mathcal{H}_n^{2m-1}$, with $\theta_i = f(x_i)$, $i = 1, \ldots, n$,

$$\int (D^m f)(x)^2 \, dx = \|V^{\frac{1}{2}} D\theta\|_2^2,$$

where $D = D_n^m$, and $V = V_n^m$ is symmetric and banded (and depends only on $x_1, \ldots, x_n$), with bandwidth $2m - 1$.

Remarks:

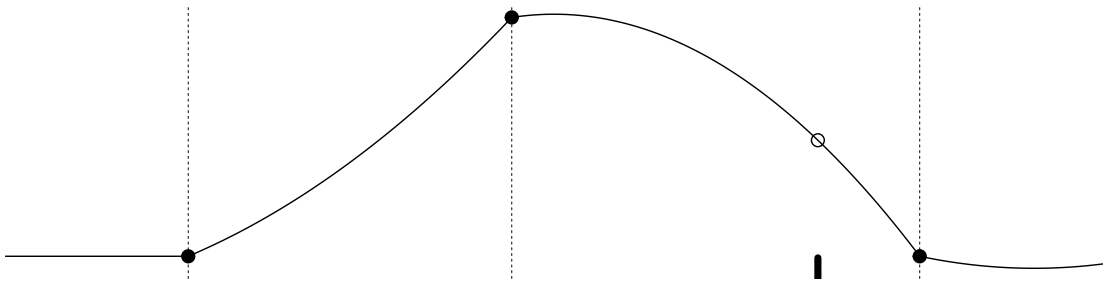
- Notably, the matrix V here is not diagonal (unlike in TV case)
- For splines, analogous result holds for a matrix V that is dense, but its inverse is banded, with bandwidth $2m - 1$
- Discrete spline result is actually more general: applies with any linear functional in place of integration (V will depend on L)

Implicit-form interpolation

Theorem. For $f \in \mathcal{H}_n^k$, given $f(x_i)$, $i = 1, \ldots, n$, we can compute $f(x)$ at any point x as follows. Let i be the smallest index such that $x_i > x$. Then $f(x)$ is the unique solution of the linear system:

$$f[x_1, \ldots, x_{k+1}, x] = 0, \quad \text{if } i \leq k; \text{ or}$$
$$f[x_{i-k}, \ldots, x_i, x] = 0, \quad \text{if } i \geq k+1.$$

Intuition: find $f(x)$ that sets $(k+1)$st discrete derivative equal to 0, hence makes f a local kth degree polynomial



Note that $f(x)$ can be computed is (nearly) constant-time!

Natural trend filtering