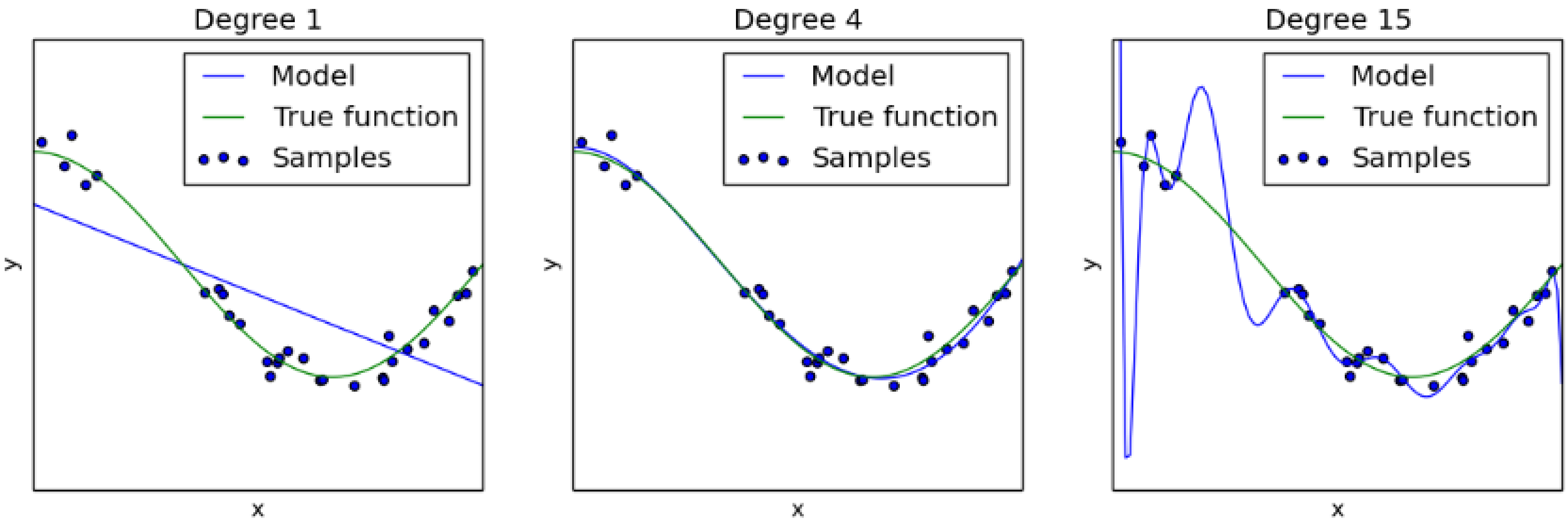
Polynomial regression without overfitting

NSM 2024

Jonas Moss (jonas.moss@bi.no)

Polynomial regression



Source: sklearn

- Polynomials with high degree have a reputation for overfitting.
- Usually, we're advised to use local polynomials or smoothing splines.
- Why? Runge's phenomenon, inherent unsophistication of polynomials.
- But polynomials are great!

jonas.moss@bi.no

Setting

- Polynomials with high degree – think 10000 if you will.
- Ridge regression on polynomial basis functions, but other objectives and regularization methods will work.

My claims

1. High-order polynomial (ridge) regression doesn't work unless you choose your polynomials wisely.
2. Well-chosen polynomials may outperform smoothing splines. And may be easier to interpret, implement, and compute.
3. I propose to use a polynomial basis for $H^2([0,1])$ in practice.

Ridge regression with polynomials

Let $\{q_k\}_{k=1}^{\infty}$ be a sequence of linearly independent polynomials.

$$\hat{\beta}_{\lambda,p} = \underset{\beta}{\operatorname{argmin}} \left[\sum_{i=1}^{n} \left(\sum_{k=0}^{p} \beta_k q_k(x_i) - y_i \right)^2 + \lambda \beta^T \beta \right]. \quad (1)$$

Alternatively, let H_p be the span of $\{q_k\}_{k=1}^{p}$ with the inner product $\langle q_i, q_j \rangle = \delta_{ij}$.

$$\hat{f}_{\lambda,p} = \underset{f \in H_p}{\operatorname{argmin}} \left[\sum_{i=1}^{n} (f(x_i) - y_i)^2 + \lambda \|f\|_{H_p}^2 \right]. \quad (2)$$

1. Formulation (1) and (2) are equivalent.
2. The space H_p is a reproducing kernel Hilbert space.
3. Computed in $O(np^2)$ time.

jonas.moss@bi.no

Kernel regression

$$\hat{f}_\lambda = \underset{f \in H}{\operatorname{argmin}} \left[\sum_{i=1}^{n} (f(x_i) - y_i)^2 + \lambda \|f\|_H^2 \right].$$

- Let H be a reproducing kernel Hilbert space.
- Then H has a unique positive semi-definite kernel k.
- The solution has closed form, computed in $O(n^3)$ time. ($O(np^2)$ for polynomials.)

$$\hat{f}_\lambda(x) = \sum_{i=1}^{n} \alpha_i k(x, x_i),$$

$$K(i, j) = k(x_i, x_j).$$

$$\alpha = (K + \lambda I)^{-1} y,$$

Kernel of H_p

- Let H_p be the span of $\{q_k\}_{k=1}^{p}$ with the inner product $\langle q_i, q_j\rangle = \delta_{ij}$.
- The kernel of H_p is $k_p(x,y) = \sum_{k=0}^{p} q_k(x)q_k(y)$.
- Define $k(x,y) = \lim_{p\to\infty} \sum_{k=0}^{p} q_k(x)q_k(y)$.

The behaviour of $\hat{f}_{\lambda,p}$ as $p\to\infty$ is decided by $\lim_{p\to\infty} k_p(x,y)$.

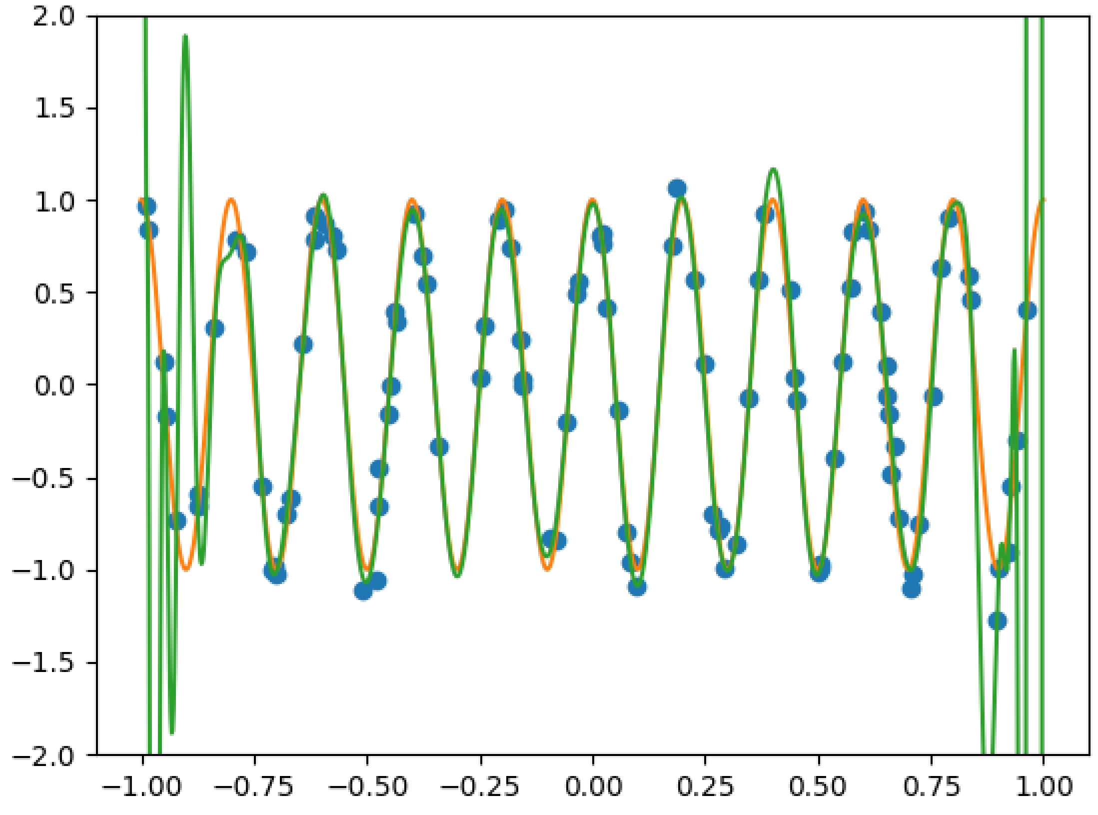
jonas.moss@bi.no

When $k_p(x, y)$ diverges

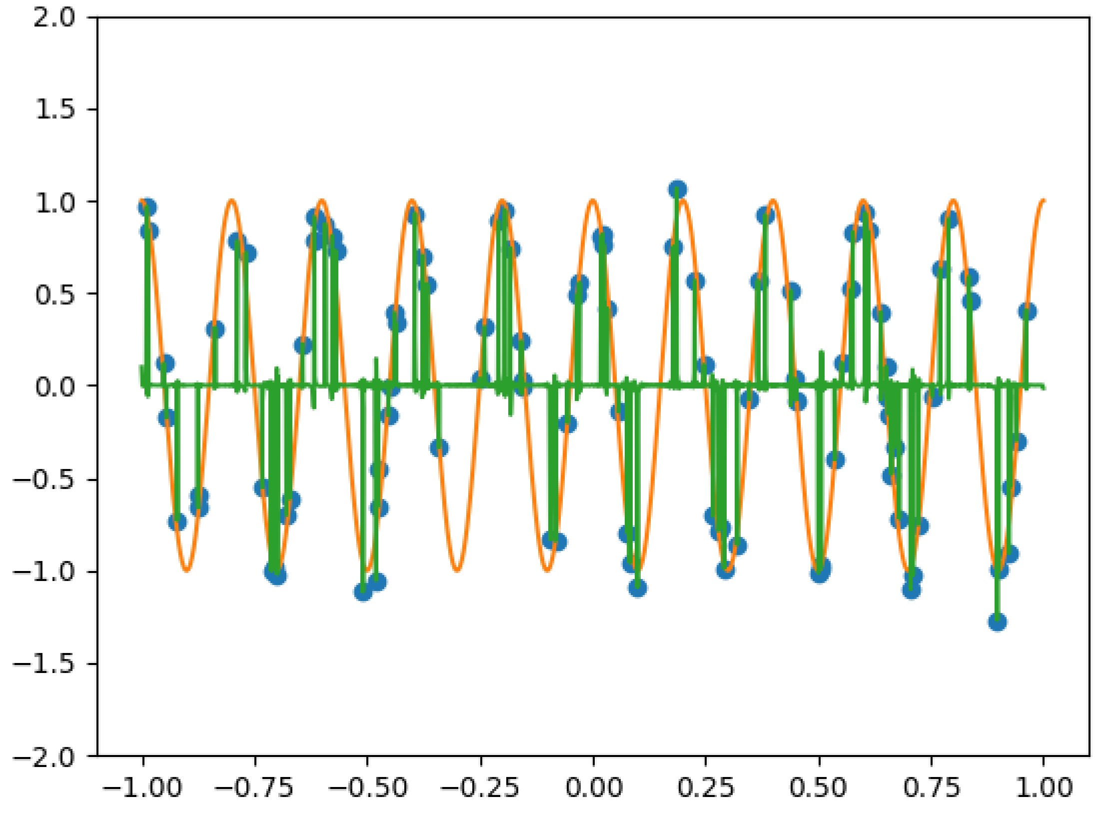
- Legendre polynomials, orthonormal on $L_2[-1,1]$.
  - $k_p(x, y) \to \begin{cases} \infty, & x = y \\ 0, & \text{otherwise.} \end{cases}$
  - Solution converges to $\sum y_\mathrm{i} 1[x = x_i]$ irrespective of λ.
- Discrete orthogonal polynomials have similar behaviour.
- Standard polynomials $(1, x, x^2, \ldots)$ on $[a, b]$ with $a \leq -1$ or $b \geq 1$ diverge too.

jonas.moss@bi.no

Legendre polynomials ($\lambda = 10^{-7}; n = 100$)



$p = 50$. Looks excellent away from the boundary.



$p = 10000$. Catastrophic overfitting.

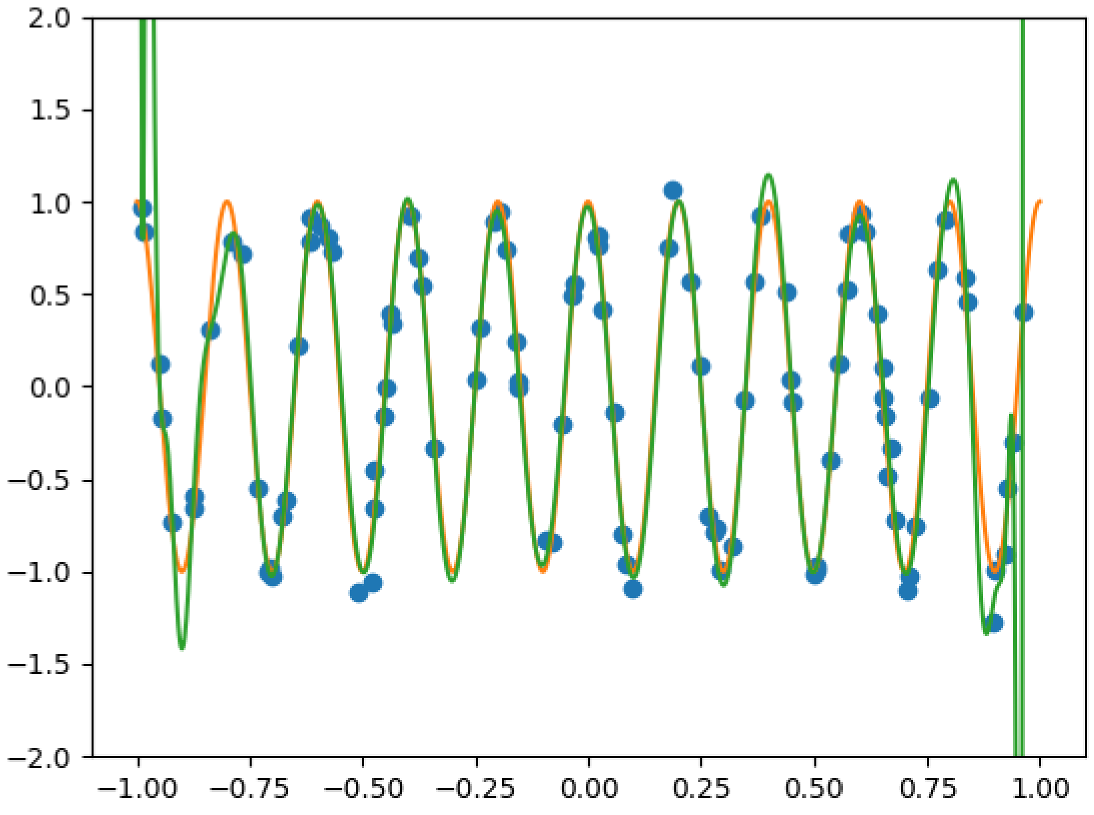
jonas.moss@bi.no

When $k_p(x, y)$ converges

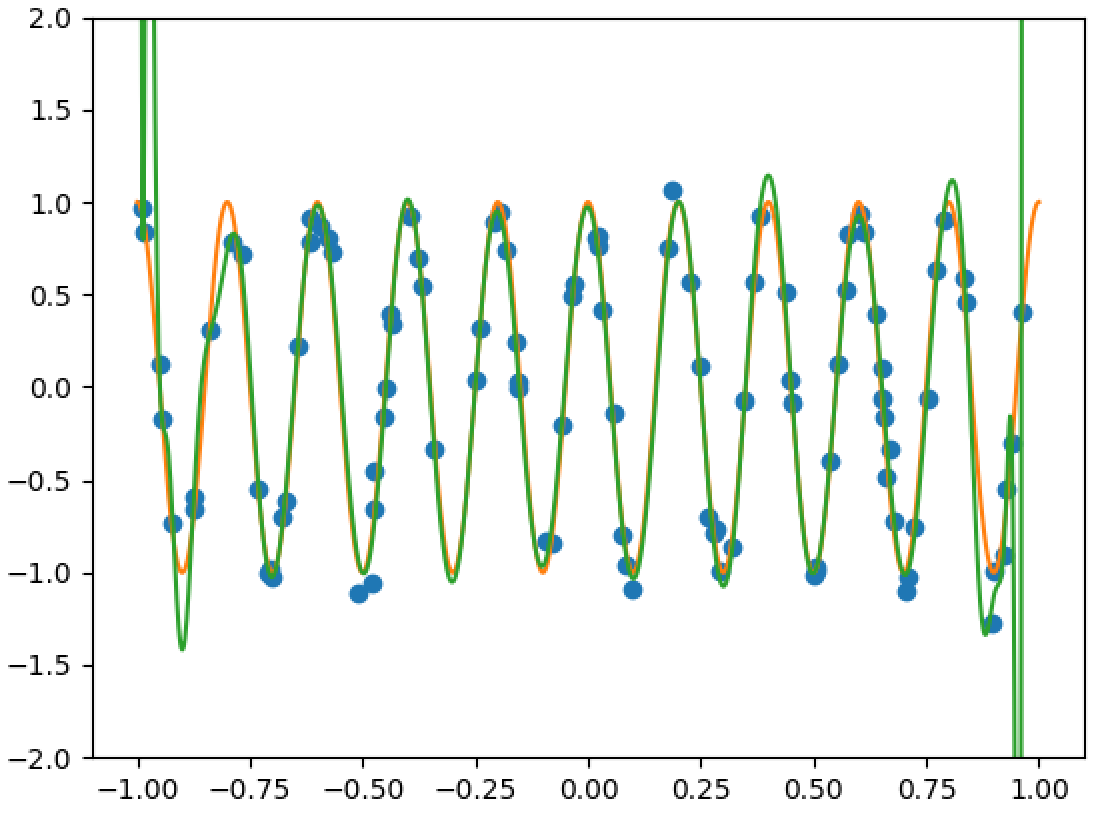
- Define $\hat{f}_{\lambda,\infty}$ as the kernel regression estimator of $k(x, y)$.
- Then $\hat{f}_{\lambda,p} \rightarrow \hat{f}_{\lambda,\infty}$.
- If the polynomials are defined on a compact set, e.g., $[0,1]$, they attain the minimax rate for functions living the reproducing kernel Hilbert space with kernel $k(x, y)$. (++)
- Standard polynomials have limit kernel $\frac{1}{1-xy}$, $x, y \in (-1,1)$.

jonas.moss@bi.no

Standard polynomials ($\lambda = 10^{-23}$)



$p = 50$. Poor boundary behaviour, but good fit in the middle.



$p = 10000$. Yes, these are different plots!

jonas.moss@bi.no

Smoothing splines

- Let the Sobolev space $H^m = \left\{ f\colon [0,1] \to \mathbb{R} \mid \int \left(f^{(m)}(x)\right)^2 dx < \infty \right\}$ be equipped with the inner product
- $\|f\| = \sum_{k=0}^{m-1} \left(f^{(k)}(0)\right)^2 + \int_0^1 \left(f^{(m)}(x)\right)^2 dx$.

The smoothing spline estimator (Wahba, 1990) is obtained by

$$\hat{f}_\lambda = \underset{f \in H}{\operatorname{argmin}} \left[\sum_{i=1}^{n} (f(x_i) - y_i)^2 + \lambda \|Pf\|_H^2 \right],$$

Where P is the orthogonal projection on $\{ f \mid f^{(k)} = 0,\ k = 0, \dots, m \}$.

We drop the projection when talking about smoothing splines.

jonas.moss@bi.no

Constructing a basis

Smoothing splines are kernel regressions on H^m.

If $\{q_{m,k}(x)\}_{k=0}^{\infty}$ is an orthonormal polynomial basis for H^m, then $\hat{f}_{\lambda,p} \to \hat{f}_{\lambda,\infty}$, the smoothing spline.

jonas.moss@bi.no

Transforming bases of $L_2[0,1]$

Let be $q_k(x)$ an orthonormal basis for $L_2([0,1])$.

$$q_{m,k}(x) = \begin{cases} \dfrac{x^k}{k!}, & k < m, \\ \dfrac{1}{(m-1)!}\displaystyle\int_0^x (x-t)^m q_{k-m}(t)dt, & k \geq m. \end{cases}$$

Then $q_{m,k}$ is an orthonormal basis for $H^m([0,1])$. (Sharapudinov, 2018; van der Vaart and Zanten, 2008).

- **Polynomial basis**: Transform the normalized and shifted Legendre polynomials.
- Cosine basis: Transform the $\{1, \cos \pi x, \cos 2\pi x\ , \ldots\}$
- Sine basis: $\{\sqrt{2}\sin \pi x\ , \sqrt{2}\sin 2\pi x\ , \ldots\}$
- Wavelet bases etc.

jonas.moss@bi.no

A polynomial basis for $H^2([0,1])$

$$p_{2,0} = 1, \qquad p_{2,1} = x,$$

$$p_{2,k} = \int_0^x (x-t)p_{k-m}(t)dt.$$

Unique orthonormal polynomial basis for $H^2([0,1])$.

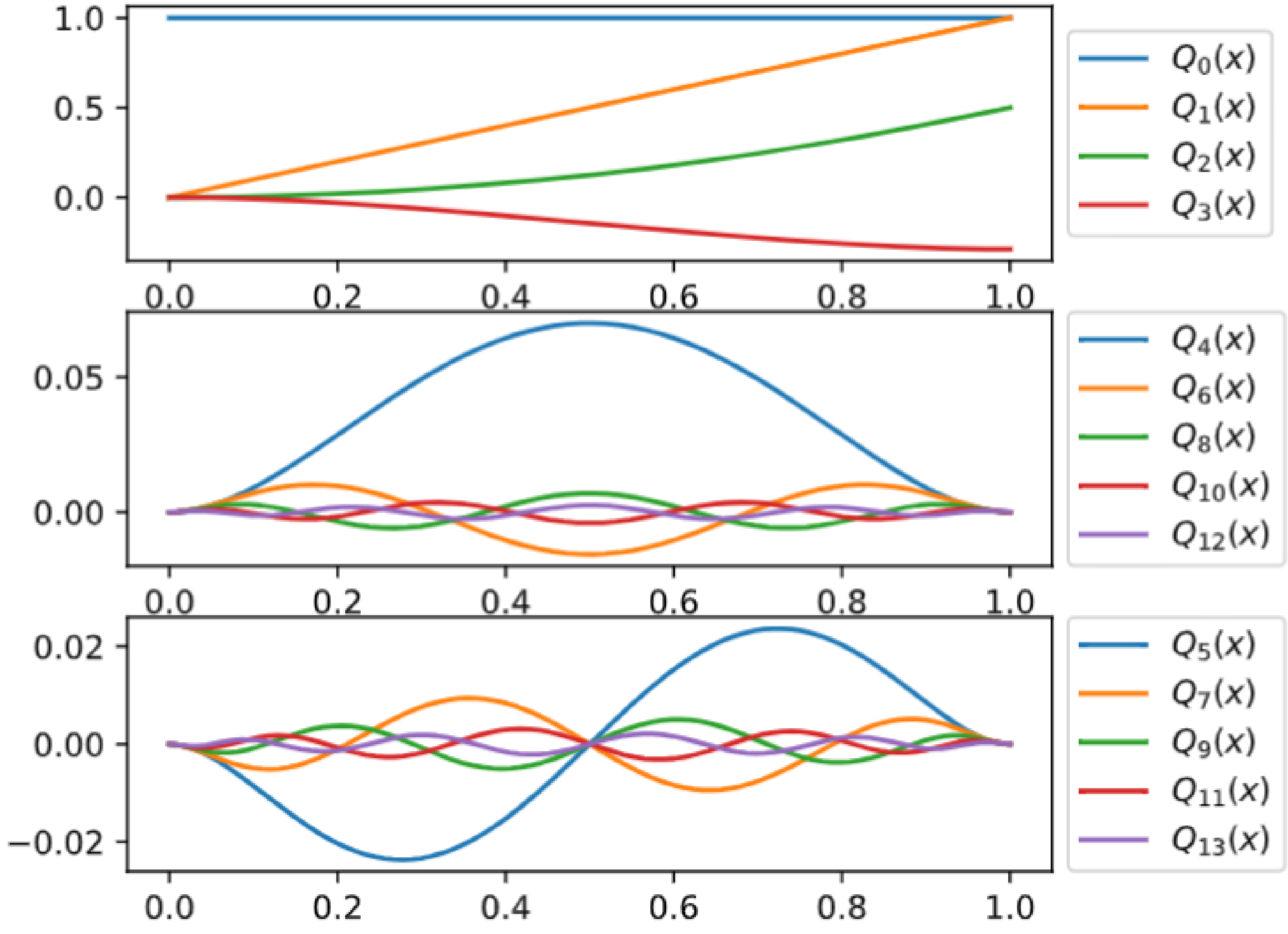
Observe: The term $x^2(1-x)^2$ occurs for all polynomials with $k \geq 4$.

Can be computed efficiently using a recursive formula.

k	Polynomial
0	1
1	x
2	$\frac{1}{4}x^2$
3	$\frac{1}{4\sqrt{3}}x^2(2x-3)$
4	$\frac{\sqrt{5}}{4}(\mathbf{1}-\boldsymbol{x})^{\mathbf{2}}\boldsymbol{x}^{\mathbf{2}}$
5	$\frac{\sqrt{7}}{4}(\mathbf{1}-\boldsymbol{x})^{\mathbf{2}}\boldsymbol{x}^{\mathbf{2}}(2x-1)$
6	$\frac{1}{4}(\mathbf{1}-\boldsymbol{x})^{\mathbf{2}}\boldsymbol{x}^{\mathbf{2}}(14x(x-1)+3)$

jonas.moss@bi.no

A polynomial basis for $H^2([0,1])$

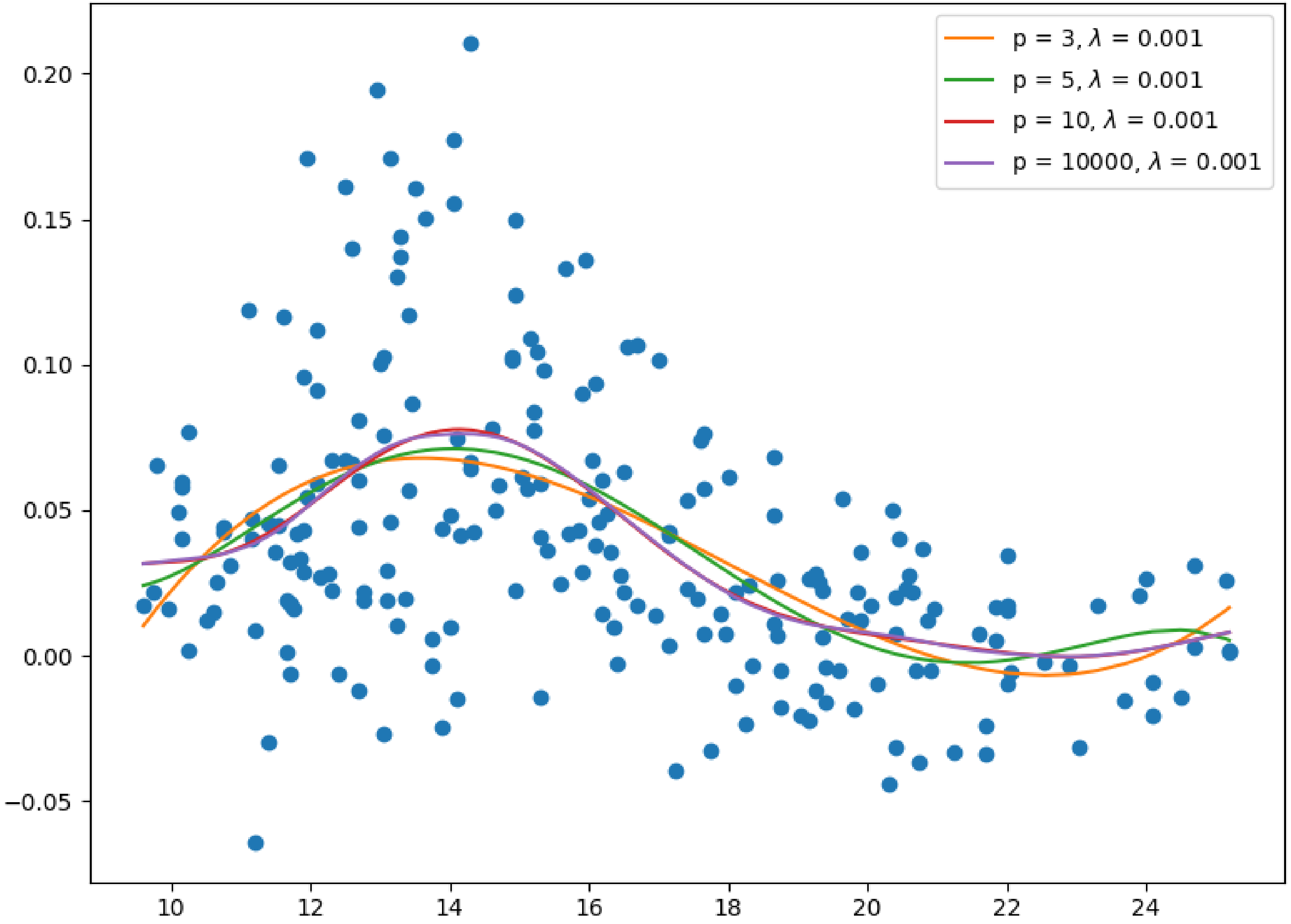


When $p \geq 4$, we have:

1. The maximum value is on the order p^{-2}.
2. Even polynomials are even (around 0.5), odd polynomials are odd.
3. The polynomials have "period" $\frac{p-3}{2}$.

jonas.moss@bi.no

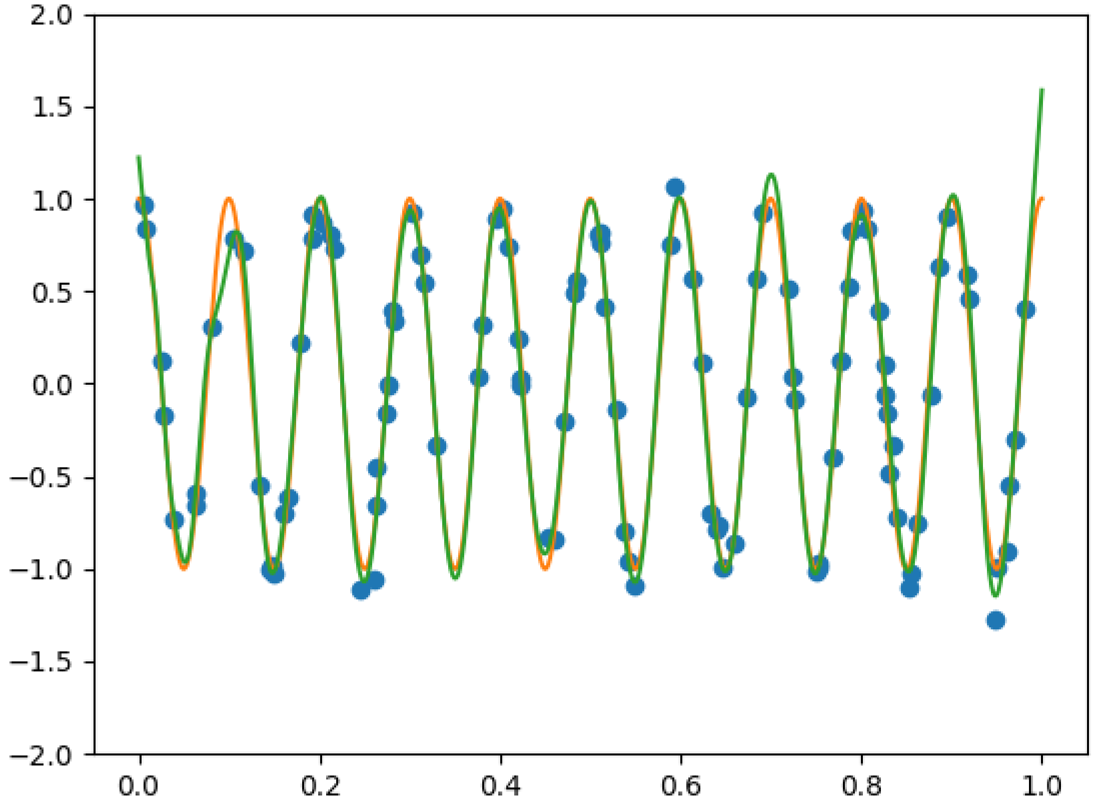
Example: Bone mineral density

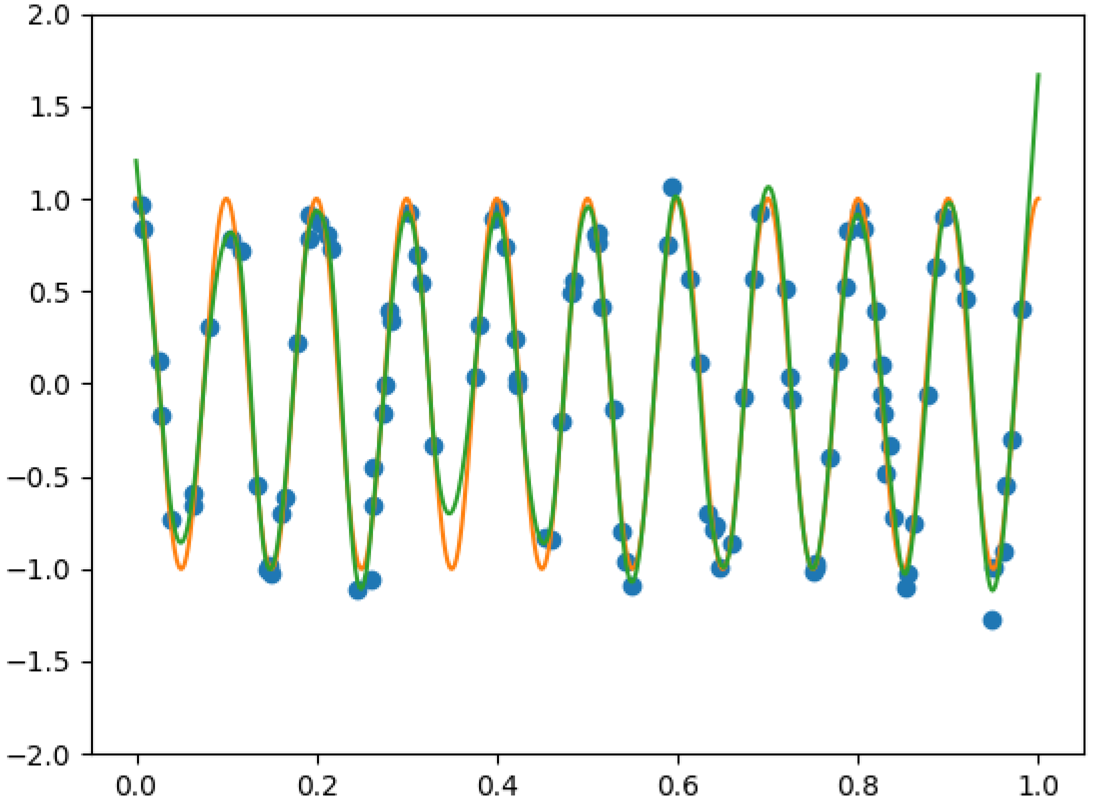


- Polynomial of degree 10000 works just fine.
- Indistinguishable from smoothing spline.
- Polynomial of degree 10 is very close.
- The other two (3,5) may not be flexible enough.
- (λ was chosen at random.)

jonas.moss@bi.no

$H^2([0,1])$ polynomials ($\lambda = 10^{-7}$)





(left) $p = 50$,(right) $p = 10000.$

jonas.moss@bi.no

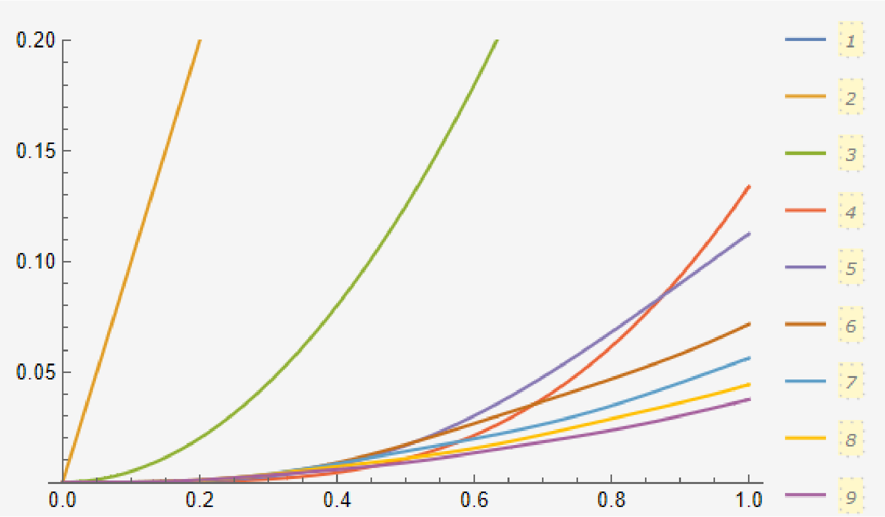
Other bases for $H^m([0,1])$

Polynomials.

- The polynomials for $H^m([0,1])$, $m \neq 2$, have similar properties.
- They are decreasing with order p^{-m} instead.

Sine basis.

- Decreasing with order p^{-m},
- Have non-negative derivatives of exactly $m - 1$ orders.
- Promising basis for shape-constrained estimation. (?)

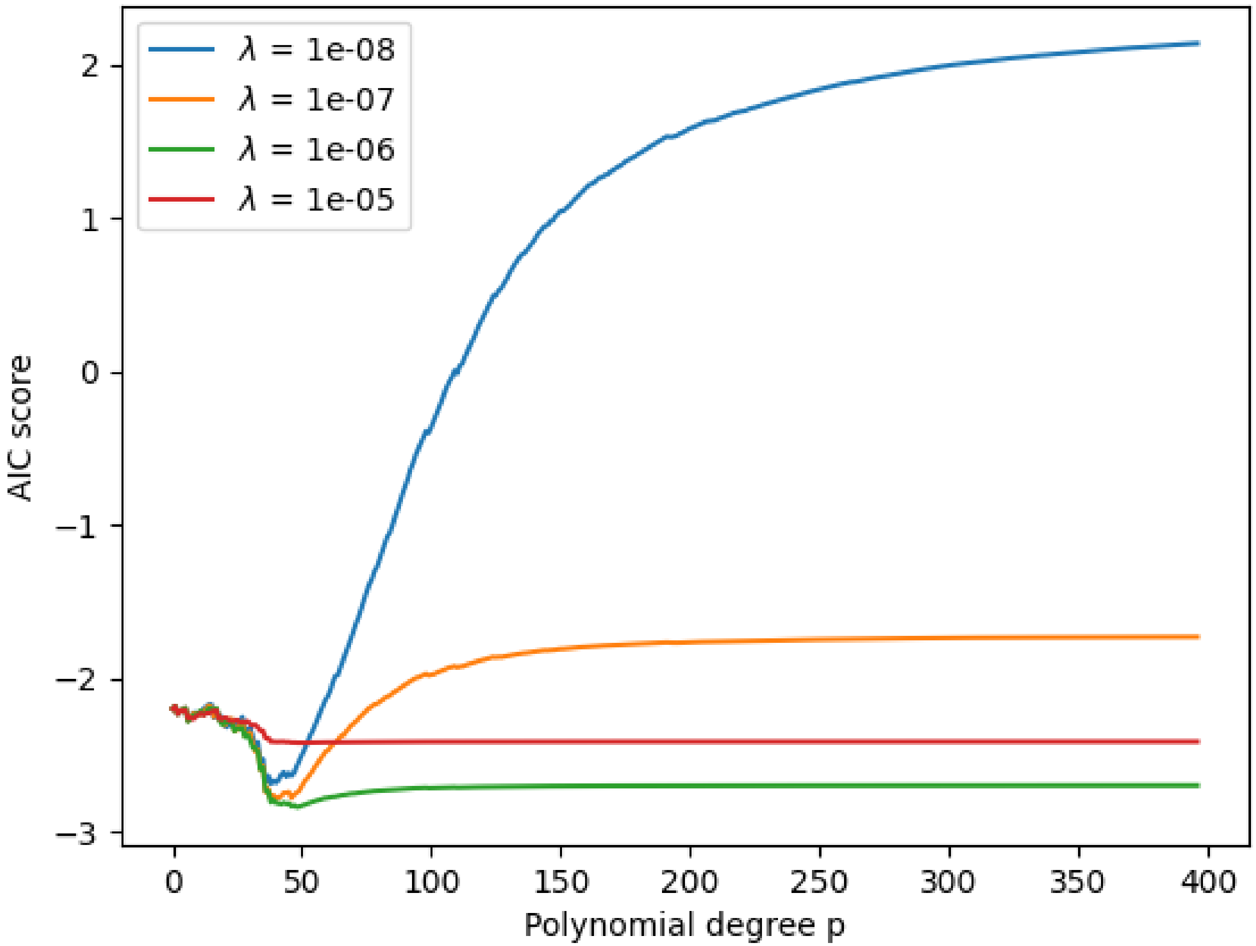


Transformed sine basis for H^3.

jonas.moss@bi.no

The AIC_{C_1} as a function of λ and p

- Hurvich et al. (1998) proposed AIC variants for smoothed regression models.
- Let's us choose λ, p.
- Asymptote corresponds to smoothing spline.
- Smaller p allows smaller λ, which may increase performance.
- Good p allows many good λs and vice versa.



jonas.moss@bi.no

Conclusion

1. High-order polynomial (ridge) regression doesn't work unless you choose your polynomials wisely by looking at the limiting kernel function.
2. Well-chosen polynomials may outperform smoothing splines. And may be easier to interpret, implement, and compute.
3. I propose to use a polynomial basis for $H^2([0,1])$ in practice.

jonas.moss@bi.no

BI