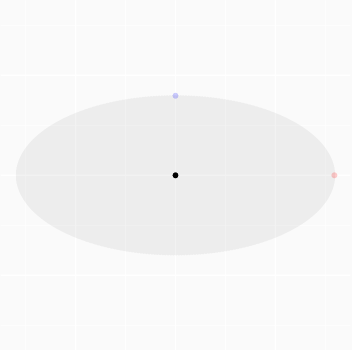
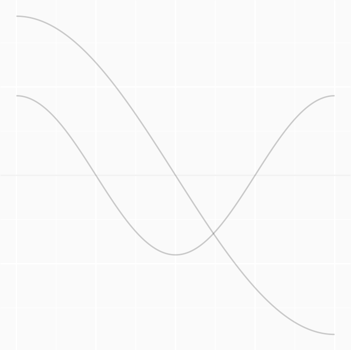
Soblev Models Review



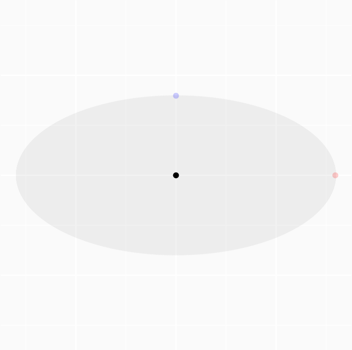


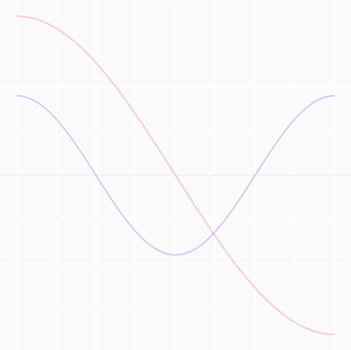
$$\mathcal{M} = \left\{ \sum_{j=0}^{\infty} m_j \phi_j(x) \;\; : \;\; \sum_{j=0}^{\infty} \lambda_j m_j^2 \leq 1 \right\}$$

for $\phi_j(x) = \sqrt{2}\cos(\pi j x)$ and $\lambda_j = \pi^2 j^2$.

Q. What's the correspondence between coefficients and curves?

Cosine Series

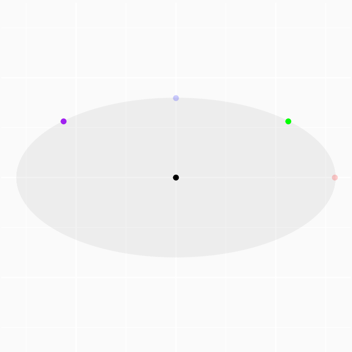


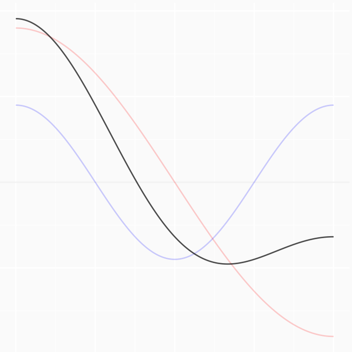


$$\mathcal{M} = \left\{ \sum_{j=0}^{\infty} m_j \phi_j(x) \;\; : \;\; \sum_{j=0}^{\infty} \lambda_j m_j^2 \le 1 \right\}$$

for $\phi_j(x) = \sqrt{2}\cos(\pi j x)$ and $\lambda_j = \pi^2 j^2$.

Q. What's the correspondence between coefficients and curves?



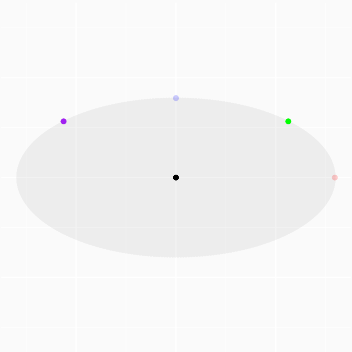


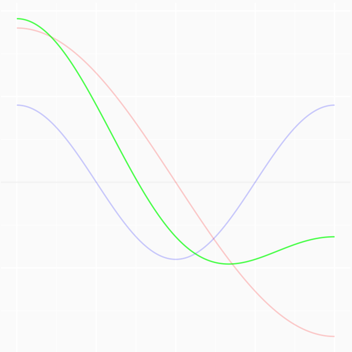
$$\mathcal{M} = \left\{ \sum_{j=0}^{\infty} m_j \phi_j(x) \ : \ \sum_{j=0}^{\infty} \lambda_j m_j^2 \le 1 \right\}$$

for $\phi_j(x) = \sqrt{2}\cos(\pi j x)$ and $\lambda_j = \pi^2 j^2$.

Q. Have I drawn the curve with the green coefficients or the purple ones?

Cosine Series

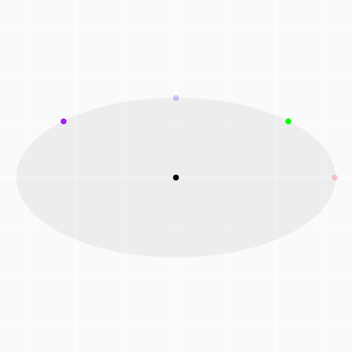


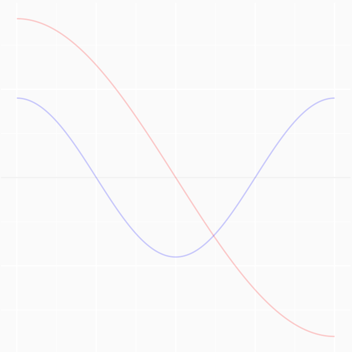


$$\mathcal{M} = \left\{ \sum_{j=0}^{\infty} m_j \phi_j(x) \; : \; \sum_{j=0}^{\infty} \lambda_j m_j^2 \le 1 \right\}$$

for $\phi_j(x) = \sqrt{2}\cos(\pi j x)$ and $\lambda_j = \pi^2 j^2$.

Q. Have I drawn the curve with the green coefficients or the purple ones?



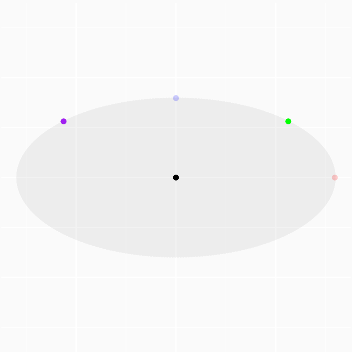


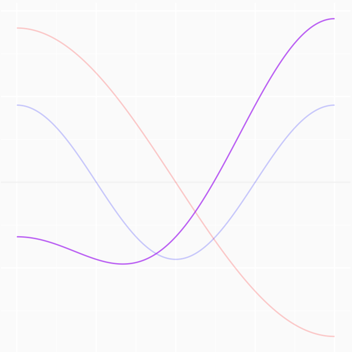
$$\mathcal{M} = \left\{ \sum_{j=0}^{\infty} m_j \phi_j(x) \ : \ \sum_{j=0}^{\infty} \lambda_j m_j^2 \le 1 \right\}$$

for $\phi_j(x) = \sqrt{2}\cos(\pi j x)$ and $\lambda_j = \pi^2 j^2$.

Exercise. Draw the curve with the purple coefficients.

Cosine Series

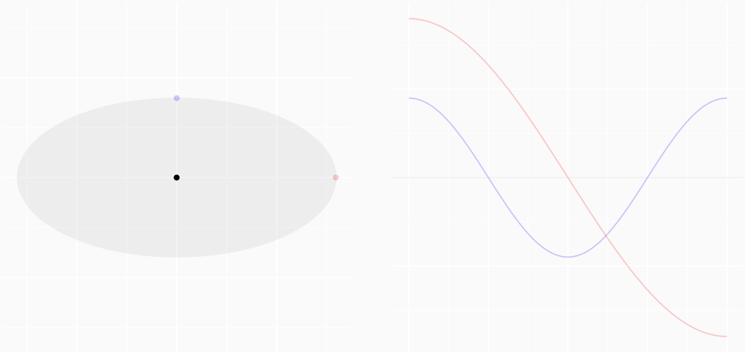




$$\mathcal{M} = \left\{ \sum_{j=0}^{\infty} m_j \phi_j(x) \quad : \quad \sum_{j=0}^{\infty} \lambda_j m_j^2 \leq 1 \right\}$$

for $\phi_j(x) = \sqrt{2}\cos(\pi j x)$ and $\lambda_j = \pi^2 j^2$.

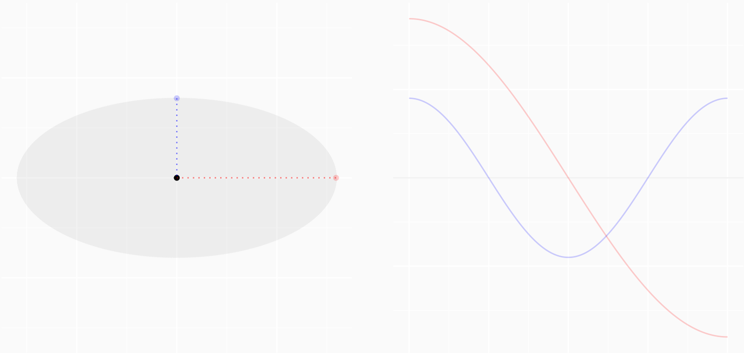
Exercise. Draw the curve with the purple coefficients.



$$\mathcal{M} = \left\{ \sum_{j=0}^{\infty} m_j \phi_j(x) \ \ : \ \ \sum_{j=0}^{\infty} \lambda_j m_j^2 \le 1 \right\}$$

for $\phi_j(x) = \sqrt{2}\cos(\pi j x)$ and $\lambda_j = \pi^2 j^2$.

Q. What's the geometric significance of $\frac{1}{\sqrt{\lambda_j}}$?



$$\mathcal{M} = \left\{ \sum_{j=0}^{\infty} m_j \phi_j(x) \quad : \quad \sum_{j=0}^{\infty} \lambda_j m_j^2 \leq 1 \right\}$$

for $\phi_j(x) = \sqrt{2}\cos(\pi j x)$ and $\lambda_j = \pi^2 j^2$.

Q. What's the geometric significance of $\frac{1}{\sqrt{\lambda_j}}$? **A.** They're ellipse radii.

Fourier Series Representation

There's an equivalent definition in terms of an *orthogonal basis* for functions on $[0, 1]$.

$$\mathcal{M} = \left\{ m : \int_0^1 m'(x)^2 \, dx \le 1 \right\} = \left\{ \sum_{j=0}^{\infty} m_j \phi_j(x) : \sum_{j=0}^{\infty} \lambda_j m_j^2 \le 1 \right\}$$

$$\text{where} \qquad \int_0^1 \phi_j(x) \phi_k(x) \, dx = 0 \quad \text{for} \quad j \ne k.$$

- We call this a *Fourier series representation*.
- It makes stuff looks a bit like what you'd see in intro classes.
- We can think of the *higher order terms* — ϕ_j where λ_j is large — much like we'd think about quadratic terms, interactions, etc., in linear regression.

In fact, these basis functions are *cosines* of increasing frequency.

Bounding the Width of Sobolev Models using Finite-dimensional Approximation

The Idea

$$\mathrm{w}(\mathcal{M}_s) \le s\sqrt{K/n} \quad \text{and therefore} \quad s = 2\sigma\sqrt{\frac{K}{n}} \quad \text{satisfies} \quad \frac{s^2}{2\sigma} \ge \sqrt{\frac{K}{n}} \ge \mathrm{w}(\mathcal{M}_s)$$

if (every vector in) $\mathcal{M}$ is — or is contained in — a subspace of dimension K.

[Gaussian Width Homework]

What if $\mathcal{M}$ is *approximately* finite-dimensional instead of exactly?

$$\mathcal{M} \subseteq \{u + v \;\; : \;\; \|u\|_{L_2(P)} \le \lambda_K^{-1/2} \;\; \text{and} \;\; v \in \mathcal{M}^K \subseteq \mathbb{R}^K\}$$

- For the Sobolev model $\mathcal{M} = \{m \;:\; \|m^{(p)}\|_{L_2(P)} \le 1\}$ ...
- ...this is true with $\lambda_K = (\pi K)^{2p}$ for every K.

[Sobolev Models Homework]

Exercise

Let's how these ideas fit together. Use the *Cauchy-Schwarz inequality* to bound the width of a neighborhood of zero in a model like this.

$$\mathrm{w}(\mathcal{M}_s) = \mathrm{E}\left[\max_{u+v \,\in\, \mathcal{M}_s} \langle g,\ u\rangle_{L_2(\mathrm{P_n})} + \langle g,\ v\rangle_{L_2(\mathrm{P_n})}\right] \quad \text{where}$$

$m \in \mathcal{M}_s$ satisfy $\|u\|_{L_2(\mathrm{P_n})} \le \lambda_K^{-1/2}$ and $v \in \mathbb{R}^K$ for $\lambda_K = (\pi K)^{2p}$

The Actual Width of Sobolev Models

A Convenient Bound on Width

Recall what gaussian width is. It's the mean of something.

$$\mathrm{w}(\mathcal{V}) = \mathrm{E}\, Z \quad \text{for} \quad Z = \max_{v \in \mathcal{V}} \frac{1}{n} \sum_{i=1}^{n} g_i v_i$$

Means are always less than root-mean-squares.

$$\mathrm{w}(\mathcal{V}) \le \mathrm{w}_2(\mathcal{V}) := \sqrt{\mathrm{E}\, Z^2} = \sqrt{(\mathrm{E}\, Z)^2 + \mathrm{Var}(Z)}$$

That root-mean-square $\mathrm{w}_2(\mathcal{V})$ is what we'll bound.

Outline

We'll bound the width of ...

1. The whole model
2. A neighborhood of zero
3. A neighborhood of an arbitrary point in the model.

Each is a small step from the last.

We'll assume X_i is uniformly distributed on $[0,1]$, i.e.,

$$\langle \phi_i, \phi_j \rangle_{L_2(P)} = \langle \phi_i, \phi_j \rangle_{L_2} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

When it's not, we can still get a similar bound. See the homework solution.

$$\mathrm{w}_2(\mathcal{M}) = \sqrt{\mathrm{E}\, Z^2} \quad \text{for} \quad Z = \max_{\substack{\text{sequences } m \\ \sum_j \lambda_j m_j^2 \le B^2}} \frac{1}{n} \sum_{i=1}^{n} g_i \left\{ \sum_j m_j \phi_j(X_i) \right\}.$$

Step 1. For all sequences m like this, via the Cauchy-Schwarz inequality,

$$\mathrm{w}_2(\mathcal{M}) = \sqrt{\mathrm{E}\, Z^2} \quad \text{for} \quad Z = \max_{\substack{\text{sequences } m \\ \sum_j \lambda_j m_j^2 \le B^2}} \frac{1}{n}\sum_{i=1}^n g_i \left\{ \sum_j m_j \phi_j(X_i) \right\}.$$

Step 1. For all sequences m like this, via the Cauchy-Schwarz inequality,

$$\begin{aligned}
\left| \frac{1}{n}\sum_{i=1}^n g_i \left\{ \sum_j m_j \phi_j(X_i) \right\} \right| &= \left| \sum_j m_j \left\{ \frac{1}{n}\sum_{i=1}^n g_i \phi_j(X_i) \right\} \right| \\
&= \left| \sum_j m_j \lambda_j^{1/2} \cdot \lambda_j^{-1/2} \left\{ \frac{1}{n}\sum_{i=1}^n g_i \phi_j(X_i) \right\} \right| \\
&\le \sqrt{\sum_j \left\{ m_j \lambda_j^{1/2} \right\}^2} \cdot \sqrt{\sum_j \left[\lambda_j^{-1/2} \left\{ \frac{1}{n}\sum_{i=1}^n g_i \phi_j(X_i) \right\} \right]^2} \\
&\le B \cdot \sqrt{\sum_j \lambda_j^{-1} \left\{ \frac{1}{n}\sum_{i=1}^n g_i \phi_j(X_i) \right\}^2}
\end{aligned}$$

The Whole Model's Width

$$\mathrm{w}_2(\mathcal{M}) = \sqrt{\mathrm{E}\, Z^2} \quad \text{where} \quad |Z| \le B \cdot \sqrt{\sum_j \lambda_j^{-1} \left\{ \frac{1}{n} \sum_{i=1}^{n} g_i \phi_j(X_i) \right\}^2}$$

Step 2. We can calculate the mean square of this bound on $|Z|$.

$$\mathrm{w}_2(\mathcal{M}) = \sqrt{\mathrm{E}\, Z^2} \quad \text{where} \quad |Z| \le B \cdot \sqrt{\sum_j \lambda_j^{-1} \left\{ \frac{1}{n} \sum_{i=1}^n g_i \phi_j(X_i) \right\}^2}$$

Step 2. We can calculate the mean square of this bound on $|Z|$.

$$\begin{aligned}
\mathrm{E}\, Z^2 &\le B^2 \sum_j \lambda_j^{-1}\, \mathrm{E} \left\{ \frac{1}{n} \sum_{i=1}^n g_i \phi_j(X_i) \right\} \left\{ \frac{1}{n} \sum_{i'=1}^n g_{i'} \phi_j(X_{i'}) \right\} \\
&= \frac{B^2}{n^2} \sum_j \lambda_j^{-1} \sum_{i=1}^n \sum_{i'=1}^n \mathrm{E}\, g_i g_{i'} \phi_j(X_i) \phi_j(X_j) \\
&\overset{(a)}{=} \frac{B^2}{n^2} \sum_j \lambda_j^{-1} \sum_{i=1}^n \sum_{i'=1}^n \mathrm{E}\, g_i g_{i'}\, \mathrm{E}\, \phi_j(X_i) \phi_j(X_j) \\
&\overset{(b)}{=} \frac{B^2}{n^2} \sum_j \lambda_j^{-1} \sum_{i=1}^n \mathrm{E}\, g_i^2\, \mathrm{E}\, \phi_j(X_i)^2 \\
&\overset{(c)}{=} \frac{B^2}{n^2} \sum_j \lambda_j^{-1} \sum_{i=1}^n 1 = \frac{B^2}{n} \sum_j \lambda_j^{-1}
\end{aligned}$$

(a) g and X are independent; (b) g_i and $g_{i'}$ are independent w/ mean zero; (c) $\mathrm{E}\, g_i^2 = \mathrm{Var}(g_i) = 1$ and $\mathrm{E}\, \phi_j(X_i)^2 = 1$.

The Whole Model's Width

$$\mathrm{w}_2(\mathcal{M}) \leq \sqrt{\frac{B^2}{n}\sum_j \lambda_j^{-1}} \quad \text{is our bound.}$$

A Neighborhood of Zero's Width

There's a trick to this. Curves in our model are in one ellipse.

$$B^2 \geq \sum_j \lambda_j m_j^2$$

The constraint that we're near zero restricts our coefficients to another ellipse.

$$s^2 \geq \left\| \sum_j m_j \phi_j \right\|_{L_2}^2 = \left\langle \sum_j m_j \phi_j, \sum_k m_k \phi_k \right\rangle_{L_2} = \sum_{j,k} m_j m_k \langle \phi_j, \phi_k \rangle_{L_2} = \sum_j m_j^2.$$

In a neighborhood of zero within our model, *both constraints* are satisfied.
And so are linear combinations of them.

$$\sum_j m_j \phi_j \in \mathcal{M}_s \implies \frac{1}{B^2} \sum_j \lambda_j m_j^2 + \frac{1}{s^2} \sum_j m_j^2 \leq 1 + 1 = 2.$$

That tells that curves in our neighborhood are contained in another ellipse.

$$\mathcal{M}_s \subseteq \tilde{\mathcal{M}}_s := \left\{ \sum_j m_j \phi_j \;\; : \;\; \sum_j \tilde{\lambda}_j m_j^2 \leq 2 \right\} \quad \text{for} \quad \tilde{\lambda}_j = \frac{\lambda_j}{B^2} + \frac{1}{s^2}$$

And we can use our 'whole model' bound on this ellipse.

$$\mathrm{w}_2(\mathcal{M}_s) \leq \mathrm{w}_2(\tilde{\mathcal{M}}_s) \leq \sqrt{\frac{2}{n} \sum_j \left(\frac{\lambda_j}{B^2} + \frac{1}{s^2} \right)^{-1}}.$$

A Neighborhood of an Arbitrary Curve

There's a trick to this too. Think of our model as a ball in the *Sobolev seminorm*.

$$\mathcal{M} = \{m : \rho(m) \le B\} \quad \text{where} \quad \rho\left(\sum_j m_j \phi_j\right) = \sqrt{\sum_j \lambda_j m_j^2}$$

Now let's think about our centered neighborhood.

$$\mathcal{M}_s - \mu^\star = \left\{m - \mu^\star \ : \ \rho(m) \le B \text{ and } \|m - \mu^\star\|_{L_2} \le s\right\}$$

This is contained in a neighborhood of zero by the triangle inequality.

$$\rho(m - \mu^\star) \le \rho(m) + \rho(\mu^\star) \le B + B$$

so

$$\mathcal{M}_s - \mu^\star \subseteq \left\{m - \mu^\star \ : \ \rho(m - \mu^\star) \le 2B \quad \text{and} \quad \|m - \mu^\star\|_{L_2} \le s\right\}.$$

This means we can use our last bound if we *double* B. Easy enough.

$$\mathrm{w}_2(\mathcal{M}_s) \le \mathrm{w}_2(\tilde{\mathcal{M}}_s) \le \sqrt{\frac{2}{n}\sum_j \left(\frac{\lambda_j}{(2B)^2} + \frac{1}{s^2}\right)^{-1}}.$$

Getting Concrete

Let's calculate this for our Sobolev model $\mathcal{M}^p$ by plugging in our eigenvalues.

$$\mathrm{w}_2(\mathcal{M}^p_s) \le \sqrt{\frac{2}{n}\sum_j \left(\frac{\lambda_j}{4B^2} + \frac{1}{s^2}\right)^{-1}} \quad \text{for} \quad \lambda_j = (\pi j)^{2p}$$

Let's calculate this for our Sobolev model $\mathcal{M}^p$ by plugging in our eigenvalues.

$$\mathrm{w}_2(\mathcal{M}^p_s) \le \sqrt{\frac{2}{n}\sum_j \left(\frac{\lambda_j}{4B^2} + \frac{1}{s^2}\right)^{-1}} \quad \text{for} \quad \lambda_j = (\pi j)^{2p}$$

Bounds and Approximations

1. Sum exceeds max.
$$\left(\frac{\lambda_j}{4B^2} + \frac{1}{s^2}\right)^{-1} \le \left\{\max\left(\frac{\lambda_j}{4B^2}, \frac{1}{s^2}\right)\right\}^{-1} = \min\left(\frac{4B^2}{\lambda_j},\ s^2\right)$$
2. Integral approximation.
$$\sum_j \min\left(\frac{4B^2}{\lambda_j},\ s^2\right) \approx \int_0^\infty \min\left(\frac{4B^2}{\lambda_x}, s^2\right) dx$$

Conclusion

$$\mathrm{w}_2(\mathcal{M}^p_s) \lesssim \sqrt{\frac{2}{n}\int_0^\infty \min\{4B^2(\pi x)^{-2p},\ s^2\}\, dx}$$
$$\le \sqrt{\frac{8B^2}{n}\int_0^\infty \min\{(\pi x)^{-2p},\ s^2\}\, dx} \quad \text{for} \quad B \ge 1/2$$

What is this really?

$$\mathrm{w}_2(\mathcal{M}_s^p) \leq \sqrt{\frac{8B^2}{n}\int_0^\infty \min\{(\pi x)^{-2p},\ s^2\}dx} \quad \text{for} \quad B \geq 1/2$$

The integral has two parts.

1. The beginning, where $(\pi x)^{-2p}$ is big and we're just integrating s^2.
2. The end, where $(\pi x)^{-2p}$ is small and we're integrating that.

When does the end start?

What is this really?

$$\mathrm{w}_2(\mathcal{M}_s^p) \leq \sqrt{\frac{8B^2}{n}\int_0^\infty \min\{(\pi x)^{-2p},\ s^2\}dx} \quad \text{for} \quad B \geq 1/2$$

The integral has two parts.

1. The beginning, where $(\pi x)^{-2p}$ is big and we're just integrating s^2.
2. The end, where $(\pi x)^{-2p}$ is small and we're integrating that.

It starts when $x > \pi^{-1} s^{-1/p}$. Let's do it.

What is this really?

$$\mathrm{w}_2(\mathcal{M}_s^p) \le \sqrt{\frac{8B^2}{n}\int_0^\infty \min\{(\pi x)^{-2p},\ s^2\}dx} \quad \text{for} \quad B \ge 1/2$$

The integral has two parts.

1. The beginning, where $(\pi x)^{-2p}$ is big and we're just integrating s^2.
2. The end, where $(\pi x)^{-2p}$ is small and we're integrating that.

It starts when $x > \pi^{-1}s^{-1/p}$. Let's do it.

$$= \int_0^{\pi^{-1}s^{-1/p}} s^2 \quad + \quad \int_{\pi^{-1}s^{-1/p}}^\infty \pi^{-2p}x^{-2p}dx$$

$$= \pi^{-1}s^{2-1/p} \quad + \quad \pi^{-2p}\,\frac{x^{1-2p}}{1-2p}\Big|_{\pi^{-1}s^{-1/p}}^{\infty} \qquad \text{if } p > 1/2\text{, otherwise } \infty$$

$$= \pi^{-1}s^{2-1/p} \quad + \quad \pi^{-2p}\,\frac{\pi^{2p-1}s^{2-1/p}}{2p-1} = c_p s^{2-1/p} \qquad \text{for} \quad c_p = \pi^{-1}\{1 + 1/(2p-1)\}.$$

Our width bound is proportional to $1/\sqrt{n}$ times the integral's square root.

$$\mathrm{w}(\mathcal{M}_s^p) \lesssim Bn^{-1/2}s^{1-1/2p}.$$

An Error Bound

To bound our least squares estimator's error, we do what we always do.
We find the smallest solution we can to this inequality.

$$\|\hat{\mu} - \mu^{\star}\| \leq s \text{ w.p } 1 - \delta \quad \text{if} \quad s^2 \geq 2\sigma c_\delta \, \mathrm{w}(\mathcal{M}_s^p) \quad \text{and therefore if} \quad s^2 \geq c'_\delta B n^{-1/2} s^{1-1/2p}$$

An Error Bound

To bound our least squares estimator's error, we do what we always do.
We find the smallest solution we can to this inequality.

$$\|\hat{\mu} - \mu^{\star}\| \le s \text{ w.p } 1-\delta \quad \text{if} \quad s^2 \ge 2\sigma c_\delta \, \mathrm{w}(\mathcal{M}_s^p) \quad \text{and therefore if} \quad s^2 \ge c_\delta' B n^{-1/2} s^{1-1/2p}$$

$$\begin{aligned} s^2 &\gtrsim n^{-1/2} s^{1-1/2p} && \text{or equivalently} \\ s^{1+1/2p} &\gtrsim n^{-1/2} && \text{or equivalently} \\ s &\gtrsim n^{-1/\{2(1+1/2p)\}} = n^{-1/(2+1/p)}. \end{aligned}$$

An Error Bound

To bound our least squares estimator's error, we do what we always do.
We find the smallest solution we can to this inequality.

$$\|\hat{\mu} - \mu^{\star}\| \le s \text{ w.p } 1-\delta \quad \text{if} \quad s^2 \ge 2\sigma c_\delta \, \mathrm{w}(\mathcal{M}_s^p) \quad \text{and therefore if} \quad s^2 \ge c'_\delta B n^{-1/2} s^{1-1/2p}$$

$$
\begin{aligned}
s^2 &\gtrsim n^{-1/2} s^{1-1/2p} && \text{or equivalently} \\
s^{1+1/2p} &\gtrsim n^{-1/2} && \text{or equivalently} \\
s &\gtrsim n^{-1/\{2(1+1/2p)\}} = n^{-1/(2+1/p)}.
\end{aligned}
$$

And this means that we gain a decimal point of precision with …

An Error Bound

To bound our least squares estimator's error, we do what we always do.
We find the smallest solution we can to this inequality.

$$\|\hat{\mu} - \mu^\star\| \le s \text{ w.p } 1-\delta \quad \text{if} \quad s^2 \ge 2\sigma c_\delta \, \mathrm{w}(\mathcal{M}_s^p) \quad \text{and therefore if} \quad s^2 \ge c'_\delta B n^{-1/2} s^{1-1/2p}$$

$$\begin{aligned} s^2 &\gtrsim n^{-1/2} s^{1-1/2p} && \text{or equivalently} \\ s^{1+1/2p} &\gtrsim n^{-1/2} && \text{or equivalently} \\ s &\gtrsim n^{-1/\{2(1+1/2p)\}} = n^{-1/(2+1/p)}. \end{aligned}$$

And this means that we gain a decimal point of precision with ...

- $10^3 = 1000$ times more data using a model with $p = 1$ bounded derivative.
- $10^{2.50} \approx 300$ times more data using a model with $p = 2$ bounded derivatives.
- $10^{2.33} \approx 200$ times more data using a model with $p = 3$ bounded derivatives.
- $10^{2.25} \approx 175$ times more data using a model with $p = 4$ bounded derivatives.

This is starting to look more possible. But getting into models we don't understand.
What do 3 or 4 bounded derivatives look like? This'll be a problem soon.

Multidimensional Sobolev Models

The Isotropic Sobolev Model

To get a multidimensional generalization of our ($p = 1$) Sobolev model, we can replace the squared derivative with the *squared norm* of the gradient.

$$\mathcal{M}^1 = \{m : \rho_{\text{-}\Delta}(m) \le B\} \quad \text{where} \quad \rho_{\text{-}\Delta}(m) = \sqrt{\int_{[0,1]^d} \|\nabla m(x)\|^2 \, dx}.$$

Much like in the univariate case, we can use integration by parts to get an equivalent definition in terms of a self-adjoint operator.

$$\mathcal{M}^1 = \{m : \rho_{\text{-}\Delta}(m) \le B\} \quad \text{where} \quad \rho_{\text{-}\Delta}(m) = \sqrt{\langle \text{-}\Delta^p \, m, m \rangle_{L_2}}.$$

That operator is the second derivative's simplest higher-dimensional generalization.

The Laplacian $\quad \text{-}\Delta \, m = -\dfrac{\partial^2}{\partial x_1^2} m(x) - \ldots - \dfrac{\partial^2}{\partial x_d^2} m(x)$

It's a self-adjoint operator on functions that are even and 2-periodic along each axis.

$$f(\pm x_1, \ldots, \pm x_d) = f(x_1 + 2j_1, \ldots, x_j + 2j_d) = f(x_1, \ldots, x_d) \quad \text{for} \quad \underbrace{j \in \mathbb{Z}^d}_{\text{integer vectors}}.$$

Eigenvectors and Eigenvalues

Because this operator self-adjoint, we know it has an orthogonal basis of eigenvectors.

The Laplacian $$-\Delta\, m = -\frac{\partial^2}{\partial x_1^2} m(x) - \ldots - \frac{\partial^2}{\partial x_d^2} m(x)$$

Anybody want to guess?

Eigenvectors and Eigenvalues

Because this operator self-adjoint, we know it has an orthogonal basis of eigenvectors.

The Laplacian $$-\Delta\, m = -\frac{\partial^2}{\partial x_1^2} m(x) - \ldots - \frac{\partial^2}{\partial x_d^2} m(x)$$

Anybody want to guess?

They're *products* of cosines.

$\phi_j(x) = \cos(\pi j_1 x_1) \cdots \cos(\pi j_d x_d)$ with eigenvalue $\lambda_j = (\pi \|j\|_2)^2$ for $\underset{\text{integer vectors}}{j \in \mathbb{Z}^d}.$

There are versions for higher order derivatives.

$$\mathcal{M}^p = \{m : \rho_{\text{-}\Delta^p}(m) \le B\} \quad \text{where} \quad \rho_{\text{-}\Delta^p}(m) = \sqrt{\langle \text{-}\Delta^p\, m, m \rangle_{L_2}}$$

And Fourier series representations.

$$\mathcal{M}^p = \left\{ \sum_{j \in \mathbb{Z}^d} m_j \phi_j \; : \sum_{j \in \mathbb{Z}^d} \lambda_j^p \; m_j^2 \le B^2 \right\} \quad \text{for} \quad \phi_j(x) = \cos(\pi j_1 x_1) \cdots \cos(\pi j_d x_d)$$

$$\text{and} \quad \lambda_j = (\pi \|j\|_2)^2.$$

You can derive all this stuff the same way as the univariate case.

The Gaussian Width of a Neighborhood

Abstractly, width is the same thing. All we used before were the eigenvalues.

$$\mathrm{w}(\mathcal{M}_s^p) \le \sqrt{\frac{8B^2}{n}\sum_j \min\left\{\lambda_j^{-1},\ s^2\right\}} \quad \text{for} \quad \lambda_j = (\pi\|j\|_2)^{2p}.$$

- But now we're summing more or them, spreading out in all d directions.
- This means we see the same value of λ_j^{-1} in the sum multiple times.
- Same $\|j\|_2$, different j.

Integral approximation makes it easy to 'count' these copies.

$$\mathrm{w}(\mathcal{M}_s^p) \lesssim \sqrt{\frac{8B^2}{n}\int_{x\in\mathbb{R}^d} \min\{(\pi\|x\|_2)^{-2p},\ s^2\}\,dx}$$

- The 'number of copies' gets larger as $\|x\|_2$ does.
- To be precise, it's the surface area of the sphere of radius $r = \|x\|_2$
- And if we change variables to polar coordinates, the integral is easy.

Step 1. Reduce it to a one-dimensional integral.

$$\mathrm{w}(\mathcal{M}_s^p)^2 \lesssim \frac{8B^2}{n} \int_{x \in \mathbb{R}^d} \min\left\{ (\pi \|x\|_2)^{-2p}, s^2 \right\} dx$$

in rectangular coordinates

The Integral

Step 1. Reduce it to a one-dimensional integral.

$$\mathrm{w}(\mathcal{M}_s^p)^2 \lesssim \frac{8B^2}{n} \int_{x \in \mathbb{R}^d} \min\left\{(\pi \|x\|_2)^{-2p}, s^2\right\} dx \qquad \text{in rectangular coordinates}$$

$$= \frac{8B^2}{n} \int \left[\int r^{d-1} \min\left\{(\pi r)^{-2p}, s^2\right\} dr \right] d\theta_1 \ldots \theta_{d-1} \qquad \text{in polar coordinates}$$

Step 1. Reduce it to a one-dimensional integral.

$$\mathrm{w}(\mathcal{M}_s^p)^2 \lesssim \frac{8B^2}{n} \int_{x \in \mathbb{R}^d} \min\{(\pi \|x\|_2)^{-2p}, s^2\} \, dx \qquad \text{in rectangular coordinates}$$

$$= \frac{8B^2}{n} \int \left[\int r^{d-1} \min\{(\pi r)^{-2p}, s^2\} \, dr \right] d\theta_1 \dots \theta_{d-1} \qquad \text{in polar coordinates}$$

$$= \frac{8B^2}{n} \left[\int r^{d-1} \min\{(\pi r)^{-2p}, s^2\} \, dr \right] \underbrace{\int 1 \, d\theta_1 \dots \theta_{d-1}}_{\substack{\text{sphere surface area} \\ 2\pi^{d/2} / \, \Gamma(d/2) \le 35}} \qquad \left[\int \dots \right] \text{ is constant in } \theta$$

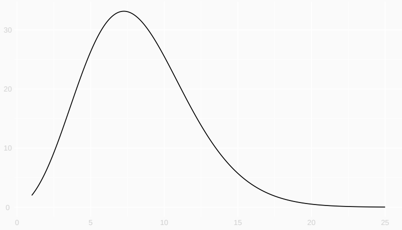


Figure 1: sphere surface area vs. dimension

The Integral

Step 2. Calculate the one-dimensional integral. This should be familiar.

$$\mathrm{w}(\mathcal{M}_s^p)^2 \lesssim \frac{8B^2}{n} \cdot \frac{2\pi^{d/2}}{\Gamma(d/2)} \cdot \int r^{d-1} \min\{(\pi r)^{-2p}, s^2\}\, dr$$

The integral has two parts.

1. The beginning, where $(\pi r)^{-2p}$ is big and we're just integrating $r^{d-1} \times s^2$.
2. The end, where $(\pi r)^{-2p}$ is small and we're integrating $r^{d-1} \times$ that.

When does the end start?

The Integral

Step 2. Calculate the one-dimensional integral. This should be familiar.

$$\mathrm{w}(\mathcal{M}_s^p)^2 \lesssim \frac{8B^2}{n} \cdot \frac{2\pi^{d/2}}{\Gamma(d/2)} \cdot \int r^{d-1} \min\{(\pi r)^{-2p}, s^2\} \, dr$$

The integral has two parts.

1. The beginning, where $(\pi r)^{-2p}$ is big and we're just integrating $r^{d-1} \times s^2$.
2. The end, where $(\pi r)^{-2p}$ is small and we're integrating $r^{d-1} \times$ that.

It starts when $r > \pi^{-1} s^{-1/p}$. Let's do it.

The Integral

Step 2. Calculate the one-dimensional integral. This should be familiar.

$$\mathrm{w}(\mathcal{M}_s^p)^2 \lesssim \frac{8B^2}{n} \cdot \frac{2\pi^{d/2}}{\Gamma(d/2)} \cdot \int r^{d-1} \min\{(\pi r)^{-2p}, s^2\}\, dr$$

The integral has two parts.

1. The beginning, where $(\pi r)^{-2p}$ is big and we're just integrating $r^{d-1} \times s^2$.
2. The end, where $(\pi r)^{-2p}$ is small and we're integrating $r^{d-1} \times$ that.

It starts when $r > \pi^{-1} s^{-1/p}$. Let's do it.

$$= \int_0^{\pi^{-1}s^{-1/p}} r^{d-1} s^2\, dr \quad + \quad \int_{\pi^{-1}s^{-1/p}}^{\infty} \pi^{-2p} r^{d-1-2p}\, dr$$

$$= s^2 \frac{r^d}{d} \Big|_0^{\pi^{-1}s^{-1/p}} \quad + \quad \pi^{-2p} \frac{r^{d-2p}}{d-2p} \Big|_{\pi^{-1}s^{-1/p}}^{\infty}$$

if $p > d/2$, otherwise ∞

$$= \frac{\pi^{-d} s^{2-d/p}}{d} \quad + \quad \frac{\pi^{-d} s^{2-d/p}}{2p-d} = c_{d,p} s^{2-d/p}$$

for $c_{d,p} = \frac{\pi^{-d}}{d}\left\{1 + \frac{1}{\frac{2p}{d} - 1}\right\}$

Summary.

Our width bound is proportional to $n^{-1/2}\ s^{1-d/2p}$.

$$\mathrm{w}(\mathcal{M}_s^p)^2 \lesssim \frac{8B^2}{n} \cdot \frac{2\pi^{d/2}}{\Gamma(d/2)} \cdot c_{d,p} s^{2-d/p}$$

An Error Bound

To bound our least squares estimator's error, we do what we always do.

$$\|\hat{\mu} - \mu^\star\| \le s \text{ w.p } 1-\delta \quad \text{if } s^2 \ge 2\sigma c_\delta \, \mathrm{w}(\mathcal{M}_s^p) \quad \text{and therefore if } s^2 \ge c'_\delta B n^{-1/2} s^{1-d/2p}$$

We've essentially solved this in the 1D case.
But now **smoothness is relative to dimension**: p/d is the new p.

$n^{-1/(2+d/p)}$ is our rate of convergence.

To bound our least squares estimator's error, we do what we always do.

$$\|\hat{\mu} - \mu^\star\| \le s \text{ w.p } 1-\delta \quad \text{if } s^2 \ge 2\sigma c_\delta \, \mathrm{w}(\mathcal{M}_s^p) \quad \text{and therefore if } s^2 \ge c'_\delta B n^{-1/2} s^{1-d/2p}$$

We've essentially solved this in the 1D case.
But now **smoothness is relative to dimension**: p/d is the new p.

$n^{-1/(2+d/p)}$ is our rate of convergence.

Derivation.

$$\begin{aligned} s^2 &\gtrsim n^{-1/2} s^{1-d/2p} && \text{or equivalently} \\ s^{1+d/2p} &\gtrsim n^{-1/2} && \text{or equivalently} \\ s &\gtrsim n^{-1/\{2(1+d/2p)\}} = n^{-1/(2+d/p)}. \end{aligned}$$

An Error Bound

To bound our least squares estimator's error, we do what we always do.

$$\|\hat{\mu} - \mu^\star\| \le s \text{ w.p } 1-\delta \quad \text{if } s^2 \ge 2\sigma c_\delta \, \mathrm{w}(\mathcal{M}_s^p) \quad \text{and therefore if } s^2 \ge c'_\delta B n^{-1/2} s^{1-d/2p}$$

We've essentially solved this in the 1D case.
But now **smoothness is relative to dimension**: p/d is the new p.

$n^{-1/(2+d/p)}$ is our rate of convergence.

Implications.
This means that we gain a decimal point of precision with …

An Error Bound

To bound our least squares estimator's error, we do what we always do.

$$\lVert\hat{\mu}-\mu^{\star}\rVert \le s \text{ w.p } 1-\delta \quad \text{if } s^2 \ge 2\sigma c_\delta \, \mathrm{w}(\mathcal{M}_s^p) \quad \text{and therefore if } s^2 \ge c'_\delta B n^{-1/2} s^{1-d/2p}$$

We've essentially solved this in the 1D case.
But now **smoothness is relative to dimension**: p/d is the new p.

$n^{-1/(2+d/p)}$ is our rate of convergence.

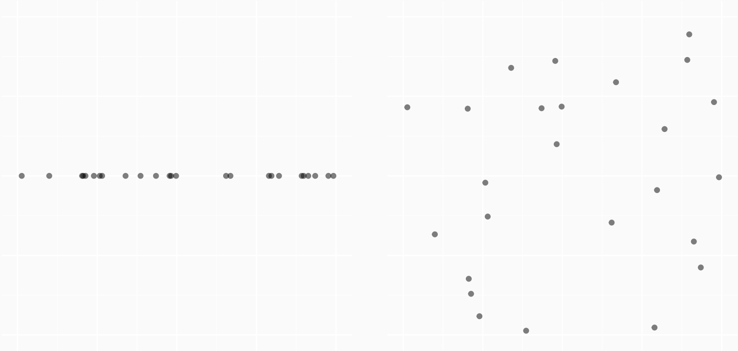
Implications.
This means that we gain a decimal point of precision with ...

- $10^4 = 10,000$ times more data using a model with $p = d/2$ bounded derivatives.
- $10^3 = 1000$ times more data using a model with $p = d$ bounded derivatives.
- $10^{2.50} \approx 300$ times more data using a model with $p = 2d$ bounded derivatives.
- $10^{2.33} \approx 200$ times more data using a model with $p = 3d$ bounded derivatives.
- $10^{2.25} \approx 175$ times more data using a model with $p = 4d$ bounded derivatives.

Smoothness doesn't count for much if it's spread over many dimensions.
Even if we've got *tons* of data, we need $3+$ derivatives in $3+$ dimensions.
That's the **curse of dimensionality**.

Intuition

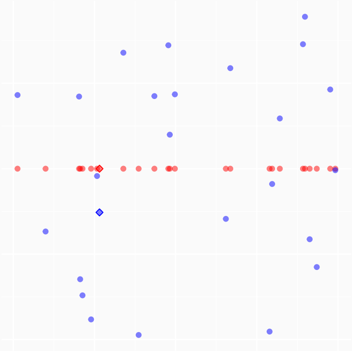
If two points are close, a smooth functions's values at them will be close.
But this isn't very useful if our observations are far apart.
And higher-dimensional observations *do* tend to be further apart.

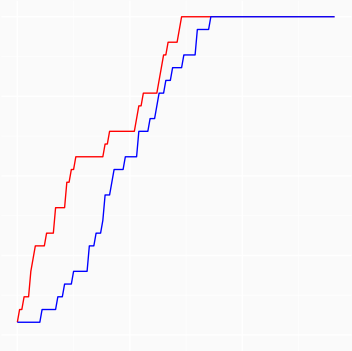


Left Uniformly distributed points in the unit interval.
Right Uniformly distributed points in the square interval.

Intuition

If two points are close, a smooth functions's values at them will be close.
But this isn't very useful if our observations are far apart.
And higher-dimensional observations *do* tend to be further apart.

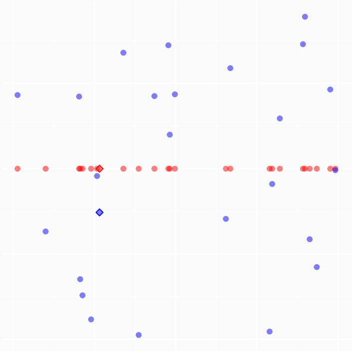


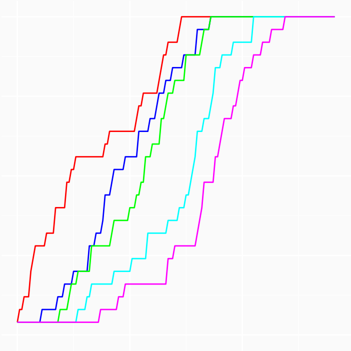


Left. As before, but overlaid.
Right. Fraction of points (y) within a distance (x) of one of them ($\diamond$).

Intuition

If two points are close, a smooth functions's values at them will be close.
But this isn't very useful if our observations are far apart.
And higher-dimensional observations *do* tend to be further apart.





Left. As before, but overlaid.
Right. Fraction of points (y) within a distance (x) of one of them ($\diamond$).
Extra curves are for the unit 3/4/5-dimensional cubes.

This looks bad

$$n^{-1/(2+d/p)} \quad \text{is our rate of convergence.}$$

The cube-root interpretation.

- With one-dimensional data, we've been getting $n^{-1/3}$ rates.
  - That's more 1 digit of precision / $1000\times$ more observations.
  - It's going from a study that enrolls the students in one intro class to everyone at Emory, UGA and Tech.
  - That's a lot, but maybe it's what we're used to and we can accept that.
  - It's what we got for monotone, bounded variation, and lipschitz regression.
- With two-dimensional data, we can do that by constraining *second derivatives*.
- With data in $3+$ dimensions, we'd need to constrain 3rd derivatives. That's bad.
  - We don't have much intution for 3rd derivatives
  - So we'd be relying on assumptions we essentially don't understand.
- People say the curse is a *high dimensional* phenomenon. It's not.
- By this standard, 3 dimensional data — most data — is high dimensional.

This looks bad

$$n^{-1/(2+d/p)} \quad \text{is our rate of convergence.}$$

The fourth-root interpretation.

- If we want to estimate something like an average treatment effect— a number rather than a curve—things aren't quite as bad.
- Clever estimators like the *R-Learner* amplify our precision.
- They make it possible to get a $n^{-1/2}$ rate estimates the effect.
    - That's more 1 digit of precision / $100\times$ more observations.
    - It's going from a study that enrolls the students in one intro class to everyone at Emory. Not terrible.
    - And there's no way to do better, even with extremely strong assumptions.
    - That's the rate at which sample averages converge.
- What we need to do that is $n^{-1/4}$ rate estimates of a few curves. π and β.
- We can do that with constrained pth derivatives for $p = d/2$.
- i.e. we can do without third derivatives until we've got $5+$-dimensional data.

This looks bad

$n^{-1/(2+d/p)}$ is our rate of convergence.

The everyone in the world interpretation

- Suppose we've run a study on a 80-student intro class.
- And we're now going to rerun it on everyone in the world.
- About 8 billion people. A hundred million (10^8) times more.
- That's a hard thing to do, so we want a big return. Two more digits.
- We can do that if we're estimating curve in K-or-fewer dimensions. What's K?

Good news?

The Isotropic Sobolev model may be the wrong model to use.
It's popular, but it's a terrible model for most things.

$$\mathcal{M} = \left\{ m : \frac{1}{2^d} \int_{[-1,1]^d} \|\nabla m(x)\|_2^2 \le B^2 \right\}$$

The problem is that it's isotropic, i.e. rotation invariant. Almost.

You can show it using the chain rule. If $m_R(x) = m(Rx)$ for a rotation matrix R,

Good news?

The Isotropic Sobolev model may be the wrong model to use.
It’s popular, but it’s a terrible model for most things.

$$\mathcal{M} = \left\{ m : \frac{1}{2^d} \int_{[-1,1]^d} \|\nabla m(x)\|_2^2 \leq B^2 \right\}$$

The problem is that it’s isotropic, i.e. rotation invariant. Almost.

You can show it using the chain rule. If $m_R(x) = m(Rx)$ for a rotation matrix R,

$$\begin{aligned} \nabla m_R(x) = R\, \nabla m(Rx) \implies \|\nabla m_R(x)\|_2^2 &= \langle R\, \nabla m(Rx),\ R\, \nabla m(Rx) \rangle_2 \\ &= \langle \underbrace{R^T R}_{=I}\, \nabla m(Rx),\ \nabla m(Rx) \rangle_2 = \|\nabla m(Rx)\|_2^2 \end{aligned}$$

And our squared Sobolev norm is this integrated over the unit cube.
That’s $\|\nabla m\|_2^2$ integrated over a rotation of that cube.

Good news?

The Isotropic Sobolev model may be the wrong model to use.
It's popular, but it's a terrible model for most things.

$$\mathcal{M} = \left\{ m : \frac{1}{2^d} \int_{[-1,1]^d} \|\nabla m(x)\|_2^2 \leq B^2 \right\}$$

The problem is that it's isotropic, i.e. rotation invariant. Almost.

Intuition.

We pay the same for variation along every unit-length combination of covariates.

$$\begin{pmatrix} \text{income74} \\ \text{income75} \end{pmatrix} \quad \text{rotates to} \quad \frac{1}{\sqrt{2}} \begin{pmatrix} \text{income74} - \text{income75} \\ \text{income74} + \text{income75} \end{pmatrix}.$$

We usually expect different amounts of variation along different combinations.
The curse hits, in part, because the model doesn't encode our assumptions.

An Overcorrection

Additive models *only* allow variation along the axes.

$$\mathcal{M} = \left\{ m(x) = m_1(x_1) + \ldots + m_d(x_d) \quad : \quad \|m_1'\|_{L_2}^2 + \ldots \|m_d'\|_{L_2}^2 \le B^2 \right\}$$

We take the contributions of each covariate and sum them up.

- What's nice is that they don't suffer from the curse of dimensionality.
- We always get error bounds comparable to what we'd get in $1D$.
- What isn't is that they can't fit all that much.

An Overcorrection

Additive models *only* allow variation along the axes.

$$\mathcal{M} = \left\{ m(x) = m_1(x_1) + \ldots + m_d(x_d) \quad : \quad \|m_1'\|_{L_2}^2 + \ldots \|m_d'\|_{L_2}^2 \le B^2 \right\}$$

We take the contributions of each covariate and sum them up.

$$\begin{pmatrix} \text{income74} \\ \text{income75} \end{pmatrix} \quad \text{rotates to} \quad \frac{1}{\sqrt{2}} \begin{pmatrix} \text{income74} - \text{income75} \\ \text{income74} + \text{income75} \end{pmatrix}.$$

- You might think average income in 74 and 75 predicts income in 76. Additive.
- Maybe you'll earn a bit more if you were on an upward trajectory. Maybe Additive.
- Maybe you'll also earn much more if you took a big dip in 75.
  e.g. you spent part of 75 unemployed. That's not additive.

Sobolev Models with Higher Order *Mixed Partials* are somewhere between these.
They penalize off-axis variation *more*, but still allow it.

This is a 2D version. We include the mixed partial.

$$\mathcal{M} = \left\{ m \quad : \quad \frac{1}{4} \int_{[-1,1]^2} \|\nabla m(x)\|^2 + \left\{ \frac{\partial^2}{\partial x_1 \partial x_2} m(x) \right\}^2 \leq B^2 \right\}$$

And this is the general case. We include *all* mixed partials.

$$\mathcal{M} = \left\{ m \quad : \quad \frac{1}{2^d} \int_{[-1,1]^d} \sum_{\substack{k \in \mathbb{Z}_+^d \\ \max_{i \leq d} k_i = 1}} \left\{ \frac{\partial^{\sum_i k_i}}{\partial x_1^{k_1} \ldots \partial x_d^{k_d}} m(x) \right\}^2 \leq B^2 \right\}$$

A Picture

Exercise

Bound the width of a neighborhood in this model.

$$\mathcal{M} = \left\{ m(x) = m_1(x_1) + \ldots + m_d(x_d) \quad : \quad \|m_1'\|_{L_2}^2 + \ldots \|m_d'\|_{L_2}^2 \le B^2 \right\}$$

Exercise

Bound the width of a neighborhood in this model.

$$\mathcal{M} = \left\{ m \quad : \quad \frac{1}{4} \int_{[-1,1]^2} \|\nabla m(x)\|^2 + \left\{ \frac{\partial^2}{\partial x_1 \partial x_2} m(x) \right\}^2 \leq B^2 \right\}$$