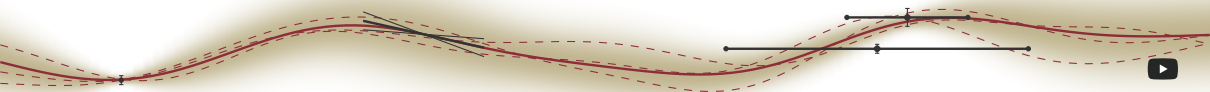
Probabilistic Inference and Learning
Lecture 10
Understanding Kernels

Philipp Hennig

19 May 2020

Eberhard Karls
Universität Tübingen

Faculty of Science
Department of Computer Science
Chair for the Methods of Machine Learning



#	date	content	Ex	#	date	content	Ex
1	20.04.	Introduction	1	14	09.06.	Logistic Regression	8
2	21.04.	Reasoning under Uncertainty		15	15.06.	Exponential Families	
3	27.04.	Continuous Variables	2	16	16.06.	Graphical Models	9
4	28.04.	Monte Carlo		17	22.06.	Factor Graphs	
5	04.05.	Markov Chain Monte Carlo	3	18	23.06.	The Sum-Product Algorithm	10
6	05.05.	Gaussian Distributions		19	29.06.	Example: Topic Models	
7	11.05.	Parametric Regression	4	20	30.06.	Mixture Models	11
8	12.05.	Learning Representations		21	06.07.	EM	
9	18.05.	Gaussian Processes	5	22	07.07.	Variational Inference	12
10	19.05.	Understanding Kernels		23	13.07.	Example: Topic Models	
11	25.05.	An Example for GP Regression	6	24	14.07.	Example: Inferring Topics	13
12	26.05.	Gauss-Markov Models		25	20.07.	Example: Kernel Topic Models	
13	08.06.	GP Classification	7	26	21.07.	Revision	

What we've seen:

- Inference in models involving linear relationships between Gaussian random variables only requires *linear algebra* operations
- features can be used to learn *nonlinear* (real-valued) functions on various domains
- feature representations can be *learned* using *type-II-maximum likelihood*
- *Gaussian process* models allow utilizing *infinitely many features* in finite time

Some questions you may have:

- What are kernels? Can I think of them as "infinitely large matrices"?
- I've heard of kernel machines. What's the connection to GPs?
- If GP's / kernel machines use infinitely many features, can they learn *every* function?

Warning

Results shown here are often simplified.
Some regularity assumptions have been dropped for easier readability.
If you don't like math, wait for the next lecture.

For deeper introductions, check out

M. Kanagawa, P. Hennig, D. Sejdinovic, and B.K. Sriperumbudur

Gaussian Processes and Kernel Methods: A Review on Connections and Equivalences

https://arxiv.org/abs/1807.02582

(*still* in review)

and

I. Steinwart, A. Christmann

Support Vector Machines

Springer SBM, 2008

What are kernels? Can I think of them as "infinitely large matrices"?

Definition (Eigenvalue)

Let $A \in \mathbb{R}^{n \times n}$ be a matrix. A scalar $\lambda \in \mathbb{C}$ and vector $v \in \mathbb{C}^n$ are called **eigenvalue** and corresponding **eigenvector** if

$$[Av]_i = \sum_{j=1}^{n} [A]_{ij}[v]_j = \lambda [v]_i.$$

Theorem (spectral theorem for symmetric positive-definite matrices)

The eigenvectors of symmetric matrices $A = A^\intercal$ are real, and form the basis of the image of A. A symmetric positive definite matrix A can be written as a Gramian (outer product) of the eigenvectors:

$$[A]_{ij} = \sum_{a=1}^{n} \lambda_a [v_a]_i [v_a]_j \qquad \text{and } \lambda_a > 0 \ \forall a = 1, \ldots, n.$$

image: The Royal Society

Definition (Eigenfunction)

A function $\phi : \mathbb{X} \to \mathbb{R}$ and scalar $\lambda \in \mathbb{C}$ that obeys

$$\int k(x, \tilde{x})\phi(\tilde{x})\, d\nu(\tilde{x}) = \lambda\phi(x)$$

are called an ***eigenfunction*** *and* ***eigenvalue*** *of k with respect to* ν.

Theorem (Mercer, 1909)

Let $(\mathbb{X}, \nu)$ *be a finite measure space and* $k : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ *a continuous (Mercer) kernel. Then there exist eigenvalues/functions* $(\lambda_i, \phi_i)_{i \in I}$ *w.r.t.* ν *such that I is countable, all* λ_i *are real and non-negative, the eigenfunctions can be made orthonormal, and the following series converges absolutely and uniformly* ν^2*-almost-everywhere:*

$$k(a, b) = \sum_{i \in I} \lambda_i \phi_i(a)\phi_i(b) \qquad \forall\, a, b \in \mathbb{X}.$$

James Mercer (1883–1932)

Are Kernels Infinitely Large Positive Definite Matrices?

Kind of ...

$$k(a,b) = \sum_{i \in I} \lambda_i \phi_i(a) \phi_i(b) =: \Phi(a) \Sigma \Phi(b)^\intercal \qquad \forall\, a, b \in \mathbb{X}.$$

- In the sense of Mercer's theorem, one may think vaguely of a kernel $k : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ evaluated at $k(a,b)$ for $a, b \in \mathbb{X}$ as the "element" of an "infinitely large" matrix k_{ab}.
- However, this interpretation is only relative to the measure $\nu : \mathbb{X} \to \mathbb{R}$.
- In general, it is not straightforward to find the eigenfunctions
- The better question is: Why do you want to think about infinite matrices?

- What are the eigenfunctions?
- Do they eigenfunctions *span a space* like the eigenvectors of a matrix?
- What's that space? Is it the sample space of a GP?

image: Rice University, 1970, CC-BY 3.0

A kernel $k(a, b)$ is called **stationary** if it can be written as

$$k(a, b) = k(\tau) \qquad \text{with} \qquad \tau := a - b$$

Theorem (Bochner's theorem (simplified))

*A complex-valued function k on $\mathbb{R}^D$ is the covariance function of a weakly **stationary** mean square continuous complex-valued random process on $\mathbb{R}^D$ if, and only if, its Fourier transform is a probability (i.e. finite positive) meausure μ:*

$$k(\tau) = \int_{\mathbb{R}^D} e^{2\pi i \boldsymbol{s}^\intercal \boldsymbol{\tau}} \, d\mu(\boldsymbol{s}) = \int_{\mathbb{R}^D} \left(e^{2\pi i \boldsymbol{s}^\intercal \boldsymbol{a}}\right) \left(e^{2\pi i \boldsymbol{s}^\intercal \boldsymbol{b}}\right)^* \, d\mu(\boldsymbol{s})$$

Note, though: Mercer's theorem described a *countable* representation!
One way to use such insights: linear-time approximations to Gaussian process regression (Rahimi & Recht, NeurIPS 2008)

Salomon Bochner
1899–1982

What are kernels? Can I think of them as "infinitely large matrices"?

- kernels have *eigenfunctions*, like matrices have eigenvectors
- eigenfunctions, though, are only defined relative to a base measure
- Mercer's theorem that the eigenfunctions "generate" the kernel
- but finding the eigenfunctions can be tricky

I've heard of kernel machines. What's the connection to GPs?

Gaussian processes, by any other name

one of the most deeply studied models in history

Equivalent and closely related names for **Gaussian process regression**

- Kriging (in particular in the geosciences)
- kernel ridge regression
- Wiener–Kolmogorov prediction
- linear least-squares regression

The Gaussian Posterior Mean is a *Least-Squares* estimate

nonparametric formulation, at explicit locations

$$
\begin{aligned}
p(f_x \mid \boldsymbol{y}) &= \frac{p(\boldsymbol{y} \mid f_X)p(f)}{p(\boldsymbol{y})} = \frac{\mathcal{N}(\boldsymbol{y}; f_X, \sigma^2 I)\mathcal{GP}(f_{x,X}; m, k)}{\mathcal{N}(\boldsymbol{y}; m_X, k_{XX} + \sigma^2 I)} \\
&= \mathcal{GP}(f_x; m_x + k_{xX}(k_{XX} + \sigma^2 I)^{-1}(\boldsymbol{y} - m_X), k_{xx} - k_{xX}(k_{XX} + \sigma^2 I)^{-1}k_{Xx}) \\
\mathbb{E}_{p(f_X|\boldsymbol{y})}(f_X) &= \arg\max_{f_X \in \mathbb{R}^{|X|}} p(f_X \mid \boldsymbol{y}) \\
&= \arg\min_{f_X} -p(f_X \mid \boldsymbol{y}) = \arg\min_{f_X} -\log p(f_X \mid \boldsymbol{y}) \\
&= \arg\min_{f_X} \frac{1}{2\sigma^2}\|\boldsymbol{y} - f_X\|^2 + \frac{1}{2}\|f_X - m_X\|_k^2 \quad \text{where} \quad \|f_X\|_k^2 := f_X^\intercal k_{XX}^{-1} f_X
\end{aligned}
$$

The **posterior mean** estimator of Gaussian (process) regression is equal to the **regularized least-squares** estimate with the regularizer $\|f\|_k^2$. This is also known as the **kernel ridge estimate**.

200 years of data analysis

and counting

portrait: Julien-Léopold Boilly, 1820 (all other portraits show a different Legendre!)

Pour cet effet, la méthode qui me paroît la plus simple et la plus générale, consiste à rendre *minimum* la somme des quarrés des erreurs. On obtient ainsi autant d'équations qu'il y a de coëfficiens inconnus ; ce qui achève de déterminer tous les élémens de l'orbite.

Comme la méthode dont je viens de parler, et que j'appelle *Méthode des moindres quarrés*, peut être d'une grande utilité dans toutes les questions de physique et d'astronomie où il s'agit de tirer de l'observation les résultats les plus exacts qu'elle peut offrir ; j'ai ajouté, dans une *appendice*, des détails particuliers sur cette méthode, et j'en ai donné l'application à la mesure de la méridienne de France, ce qui pourra servir de complément à ce que j'ai déjà publié sur cette matière.

Nouvelles méthodes pour la détermination des orbites des comètes, 1805

Adrien-Marie Legendre
1752–1833

179.

Nun will ich entwickeln, was aus diesem Gesetze folgt. Es ist von selbst klar, dass, damit das Product $\Omega = h^{\mu}\pi^{-\frac{1}{2}\mu}e^{-hh(vv+v'v'+v''v''+\ldots)}$ ein Grösstes werde, die Summe $vv+v'v'+v''v''+$ etc. ein Kleinstes werden müsse. *Es wird daher das wahrscheinlichste System der Werthe der Unbekannten p, q, r, s* etc. *dasjenige sein, in welchem die Quadrate der Unterschiede zwischen den beobachteten und berechneten Functionenwerthen V, V′, V″* etc. *die kleinste Summe geben,* wenn nämlich bei allen Beobachtungen derselbe Grad der Genauigkeit zu präsumiren ist.

Dieser Grundsatz, welcher bei allen Anwendungen der Mathematik auf die Natur-Philosophie ausserordentlich häufig benutzt wird, muss allenthalben an Stelle eines Axioms mit demselben Rechte gelten, mit welchem das arithmetische Mittel unter mehren beobachteten Werthen derselben Grösse als der wahrscheinlichste Werth angenommen wird.

Theorie der Bewegung der Himmelskörper welche in Kegelschnitten die Sonne umlaufen, 1877

Carl-Friedrich Gauss
1777 – 1855

What about all those kernel concepts?

What's the relationship between GPs and kernel ridge regression?

slides: Ulrike v. Luxburg, 2020

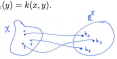
Reproducing kernel Hilbert space (RKHS)

As vector space we are going to use a space of functions:

- Consider a mapping $\Phi : \mathcal{X} \to \mathbb{R}^{\mathcal{X}}$ (where $\mathbb{R}^{\mathcal{X}}$ denotes the space of all real-valued functions from $\mathcal{X}$ to $\mathbb{R}$), defined as

$$x \mapsto \Phi(x) := k_x := k(x, \cdot)$$

That is, the point $x \in \mathcal{X}$ is mapped to the function $k_x : \mathcal{X} \to \mathbb{R}$, $k_x(y) = k(x, y)$.



Reproducing kernel Hilbert space (RKHS) (2)

- Now consider the images $\{k_x | x \in \mathcal{X}\}$ as a spanning set of a vector space. That is, we define the space $\mathcal{G}$ that contains all finite linear combinations of such functions:

$$\mathcal{G} := \{\sum_{i=1}^{r} \alpha_i k(x_i, \cdot) \mid \alpha_i \in \mathbb{R}, r \in \mathbb{N}, x_i \in \mathcal{X}\}$$

Reproducing kernel Hilbert space (RKHS) (3)

- Define a scalar product on $\mathcal{G}$ as follows:
  - For the spanning functions we define

$$\langle k_x, k_y \rangle = \langle k(x, \cdot), k(y, \cdot) \rangle := k(x, y)$$

  - For general functions in $\mathcal{G}$ the scalar product is then given as follows: If $g = \sum_i \alpha_i k(x_i, \cdot)$ and $f = \sum_j \beta_j k(y_i, \cdot)$ then

$$\langle f, g \rangle_{\mathcal{G}} := \sum_{i,j} \alpha_i \beta_j k(x_i, y_j)$$

To make sure that this is really a scalar product, we need to prove two things (EXERCISE!):

- Check that this is well-defined (not obvious because there might be several different linear combinations for the same function).
- Check that it satisfies all properties of a scalar product (crucial ingredient is the fact that k is positive definite.)

Reproducing kernel Hilbert space (RKHS) (4)

- Finally, to make $\mathcal{G}$ a proper Hilbert space we need to take its topological completion $\overline{\mathcal{G}}$, that is we add all limits of Cauchy sequences.

- The resulting space $\mathcal{H} := \overline{\mathcal{G}}$ is called the reproducing kernel Hilbert space.

- By construction, it has the property that

$$k(x, y) = \langle \Phi(x), \Phi(y) \rangle.$$

(*) RKHS, further properties

The reproducing property:

Let $f = \sum_i \alpha_i k(x_i, \cdot)$. Then $\langle f, k(x, \cdot) \rangle = f(x)$.

Proof.

$$\begin{aligned} \langle k(x, \cdot), f \rangle &= \langle k(x, \cdot), \sum_i \alpha_i k(x_i, \cdot) \rangle \\ &= \sum_i \alpha_i \langle k(x_i, \cdot), k(x, \cdot) \rangle \\ &= \sum_i \alpha_i k(x_i, x) \\ &= f(x) \end{aligned}$$

☺

(*) RKHS, further properties (2)

For those who know a bit of functional analysis:

- Let $\mathcal{H}$ be a Hilbert space of functions from $\mathcal{X}$ to $\mathbb{R}$. Then $\mathcal{H}$ is a reproducing kernel Hilbert space if and only if all evaluation functionals $\delta_x : \mathcal{H} \to \mathbb{R}$, $f \mapsto f(x)$ are continuous.
- In particular, functions in an RKHS are pointwise well defined (as opposed to, say, function in an L_2-space which are only defined almost everywhere).
- Given a kernel, the RKHS is unique (up to isometric isomorphisms). Given an RKHS, the kernel is unique.
- There is a connection to the Riesz representation theorem.

Reproducing Kernel Hilbert Spaces

Two definitions

[Schölkopf & Smola, 2002 / Rasmussen & Williams, 2006]

Definition (Reproducing kernel Hilbert space (RKHS))

Let $\mathcal{H} = (\mathbb{X}, \langle \cdot, \cdot \rangle)$ be a Hilbert space of functions $f : \mathbb{X} \to \mathbb{R}$. Then $\mathcal{H}$ is called a **reproducing kernel Hilbert space** if there exists a **kernel** $k : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ s.t.

1. $\forall x \in \mathbb{X}: \quad k(\cdot, x) \in \mathcal{H}$
2. $\forall f \in \mathcal{H}: \quad \langle f(\cdot), k(\cdot, x) \rangle_{\mathcal{H}} = f(x)$ *k reproduces* $\mathcal{H}$

Theorem [Aronszajn, 1950]: For every pos.def. k on $\mathbb{X}$, there exists a unique RKHS.

What is the RKHS? (1)

The RKHS is the space of possible posterior mean functions

[e.g. Rasmussen & Williams, 2006, Eq. 6.5]

Theorem (Reproducing kernel map representation)

Let $\mathbb{X}, \nu, (\phi_i, \lambda_i)_{i\in I}$ be defined as before. Let $(x_i)_{i\in I} \subset \mathbb{X}$ be a countable collection of points in $\mathbb{X}$. Then the RKHS can also be written as the space of linear combinations of kernel functions:

$$\mathcal{H}_k = \left\{ f(x) := \sum_{i\in I} \tilde{\alpha}_i k(x_i, x) \right\} \quad \text{with} \quad \langle f, g \rangle_{\mathcal{H}_k} := \sum_{i\in I} \frac{\tilde{\alpha}_i \tilde{\beta}_i}{k(x_i, x_i)}$$

Proof: cf. Prof. v. Luxburg's lecture

What is the RKHS? (1)

The RKHS is the space of possible posterior mean functions

[e.g. Rasmussen & Williams, 2006, Eq. 6.5]

Theorem (Reproducing kernel map representation)

Let $\mathbb{X}, \nu, (\phi_i, \lambda_i)_{i\in I}$ be defined as before. Let $(x_i)_{i\in I} \subset \mathbb{X}$ be a countable collection of points in $\mathbb{X}$. Then the RKHS can also be written as the space of linear combinations of kernel functions:

$$\mathcal{H}_k = \left\{ f(x) := \sum_{i\in I} \tilde{\alpha}_i k(x_i, x) \right\} \quad \textit{with} \quad \langle f, g\rangle_{\mathcal{H}_k} := \sum_{i\in I} \frac{\tilde{\alpha}_i \tilde{\beta}_i}{k(x_i, x_i)}$$

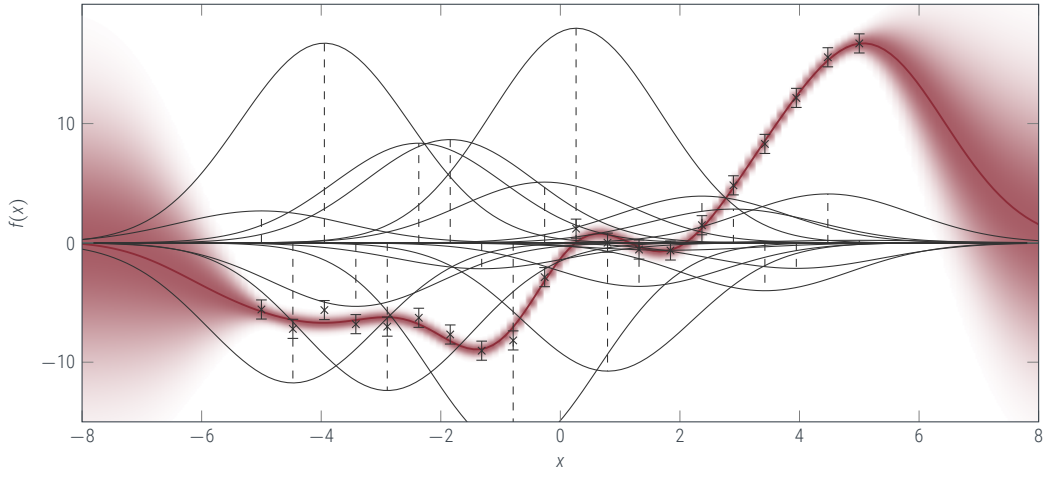
Proof: cf. Prof. v. Luxburg's lecture

Consider the Gaussian process $p(f) = \mathcal{GP}(0, k)$ with likelihood $p(\boldsymbol{y} \mid f, X) = \mathcal{N}(\boldsymbol{y}; f_X, \sigma^2 I)$. The RKHS is the space of all *possible* posterior mean functions

$$\mu(x) = k_{xX} \underbrace{(k_{XX} + \sigma^2 I)^{-1} \boldsymbol{y}}_{:= w} = \sum_{i=1}^{n} w_i k(x, x_i) \quad \text{for } n \in \mathbb{N}.$$

To understand what a GP can *learn* we have to analyze the RKHS

the connection to the statistical learning theory of RKHSs



What is the meaning of the GP point estimate?

The posterior mean is the *least-squares* estimate in the RKHS

Theorem (The Kernel Ridge Estimate)

Consider the model $p(f) = \mathcal{GP}(f; 0, k)$, $p(\boldsymbol{y} \mid f) = \mathcal{N}(\boldsymbol{y}; \boldsymbol{f}_X, \sigma^2 I)$. *The posterior mean*

$$m(x) = k_{xX}(k_{XX} + \sigma^2 I)^{-1}\boldsymbol{y}$$

is the element of the RKHS $\mathcal{H}_k$ *that minimizes the regularised* ℓ_2 *loss*

$$L(f) = \frac{1}{\sigma^2}\sum_i (f(x_i) - y_i)^2 + \|f\|^2_{\mathcal{H}_k}.$$

Proof: cf. Prof. v. Luxburg's lecture

What is the meaning of uncertainty?

Frequentist interpretation of the posterior variance

How far could the posterior mean be from the truth, assuming noise-free observations?

$$\sup_{f\in\mathcal{H},\|f\|\leq 1} (m(x)-f(x))^2 = \sup_{f\in\mathcal{H},\|f\|\leq 1} \left(\sum_i f(x_i) \underbrace{[K_{XX}^{-1}k(X,x)]_i}_{w_i} - f(x) \right)^2$$

reproducing property:

$$= \sup \left\langle \sum_i w_i k(\cdot,x_i) - k(\cdot,x), f(\cdot) \right\rangle_{\mathcal{H}}^2$$

Cauchy-Schwartz: $(|\langle a,b\rangle| \leq \|a\| \cdot \|b\|)$

$$= \left\| \sum_i w_i k(\cdot,x_i) - k(\cdot,x) \right\|_{\mathcal{H}}^2$$

reproducing property:

$$= \sum_{ij} w_i w_j k(x_i,x_j) - 2\sum_i w_i k(x,x_i) + k(x,x)$$

$$= k_{xx} - k_{xX}K_{XX}^{-1}k_{Xx} = \mathbb{E}_{|y}[(f_x-\mu_x)^2]$$

Bayesians expect the worst

it's not always true that "Frequentists are pessimists"

Theorem

Assume $p(f) = \mathcal{GP}(f; 0, k)$ and noise-free observations $p(y \mid f) = \delta(y - f_X)$. The GP posterior variance (the expected square error)

$$v(x) := \mathbb{E}_{p(f|y)}(f(x) - m(x))^2 = k_{xx} - k_{xX}K_{XX}^{-1}k_{Xx}$$

is a worst-case bound on the divergence between $m(x)$ and an RKHS element of bounded norm:

$$v(x) = \sup_{f \in \mathcal{H}_k, \|f\| \leq 1} (m(x) = f(x))^2$$

The GP's **expected** square error is the RKHS's **worst-case** square error for bounded norm.

Nb: $v(x)$ is not, in general, itself an element of $\mathcal{H}_k$.

What is the RKHS? (2)

Representation in terms of eigenfunctions

[I. Steinwart and A. Christmann. *Support Vector Machines*, 2008, Thm. 4.51]

Theorem (Mercer Representation)

Let $\mathbb{X}$ be a compact metric space, k be a continuous kernel on $\mathbb{X}$, ν be a finite Borel measure whose support is $\mathbb{X}$. Let $(\phi_i, \lambda_i)_{i\in I}$ be the eigenfunctions and values of k w.r.t. ν. Then the RKHS $\mathcal{H}_k$ is given by

$$\mathcal{H}_k = \left\{ f(x) := \sum_{i\in I} \alpha_i \lambda_i^{1/2} \phi_i(x) \text{ such that } \|f\|^2_{\mathcal{H}_k} := \sum_{i\in I} \alpha_i^2 < \infty \right\} \quad \text{with} \quad \langle f, g\rangle_{\mathcal{H}_k} := \sum_{i\in I} \alpha_i \beta_i$$

For $f = \sum_{i\in I} \alpha_i \lambda_i^{1/2} \phi_i$ and $g = \sum_{i\in I} \beta_i \lambda_i^{1/2} \phi_i$.

A compact space, simplified, is a space that is both bounded (all points have finite distance from each other) and closed (it contains all limits). For topological spaces, this is more generally defined by every open cover (every union C of open sets covering all of $\mathbb{X}$) having a finite subcover (i.e. a finite subset of C that also covers $\mathbb{X}$).

Simplified proof: First, show that this space matches the RKHS definition

1. $\forall x \in \mathbb{X}: \quad k(\cdot, x) = \sum_{i\in I} \lambda_i^{1/2} \phi_i(\cdot) \cdot \underbrace{\lambda_i^{1/2} \phi_i(x)}_{\alpha_i}$ and $\|k(\cdot, x)\|^2 = \sum_i \lambda_i \phi_i(x)^2 = k(x, x) < \infty$
2. $\langle f(\cdot), k(\cdot, x)\rangle = \sum_{i\in I} \alpha_i \lambda_i^{1/2} \phi(x) = f(x)$.

Then use Aronszajn's uniqueness result. □

What about the samples?

Draws from a Gaussian process

[for non-simplified version, cf. Kanagawa et al., 2018 (op.cit.), Thms. 4.3 and 4.9]

Theorem (Karhunen-Loève Expansion)

Let $\mathbb{X}$ be a compact metric space, $k : \mathbb{X} \times \mathbb{X}$, k be a continuous kernel, ν a finite Borel measure whose support is $\mathbb{X}$, and $(\phi_i, \lambda_i)_{i \in I}$ as above. Let $(z_i)_{i \in I}$ be a collection of iid. standard Gaussian random variables:

$$z_i \sim \mathcal{N}(0,1) \quad \textit{and} \quad \mathbb{E}[z_i, z_j] = \delta_{ij}, \quad \textit{for } i,j \in I.$$

Then (simplified!):

$$f(x) = \sum_{i \in I} z_i \lambda_i^{1/2} \phi_i(x) \sim \mathcal{GP}(0, k).$$

What about the samples?

Draws from a Gaussian process

[for non-simplified version, cf. Kanagawa et al., 2018 (op.cit.), Thms. 4.3 and 4.9]

Theorem (Karhunen-Loève Expansion)

Let $\mathbb{X}$ be a compact metric space, $k : \mathbb{X} \times \mathbb{X}$, k be a continuous kernel, ν a finite Borel measure whose support is $\mathbb{X}$, and $(\phi_i, \lambda_i)_{i \in I}$ as above. Let $(z_i)_{i \in I}$ be a collection of iid. standard Gaussian random variables:

$$z_i \sim \mathcal{N}(0, 1) \quad \text{and} \quad \mathbb{E}[z_i, z_j] = \delta_{ij}, \quad \text{for } i, j \in I.$$

Then (simplified!):

$$f(x) = \sum_{i \in I} z_i \lambda_i^{1/2} \phi_i(x) \sim \mathcal{GP}(0, k).$$

Corollary (Wahba, 1990. Proper proof in Kanagawa et al., Thm. 4.9)

If I is infinite, $f \sim \mathcal{GP}(0, k)$ implies almost surely $f \notin \mathcal{H}_k$. To see this, note

$$\mathbb{E}(\|f\|_{\mathcal{H}_k}^2) = \mathbb{E}\left(\sum_{i \in I} z_i^2\right) = \sum_{i \in I} \mathbb{E}[z_i^2] = \sum_{i \in I} 1 \not< \infty$$

GP samples are not in the RKHS!

But almost …

Theorem (Kanagawa, 2018. Restricted from Steinwart, 2017, itself generalized from Driscoll, 1973)

Let $\mathcal{H}_k$ be a RKHS and $0 < \theta \leq 1$. Consider the **θ-power of $\mathcal{H}_k$** *given by*

$$\mathcal{H}_k^\theta = \left\{ f(x) := \sum_{i \in I} \alpha_i \lambda_i^{\theta/2} \phi_i(x) \text{ such that } \|f\|_{\mathcal{H}_k}^2 := \sum_{i \in I} \alpha_i^2 < \infty \right\} \quad \text{with} \quad \langle f, g \rangle_{\mathcal{H}_k} := \sum_{i \in I} \alpha_i \beta_i.$$

Then,

$$\sum_{i \in I} \lambda_i^{1-\theta} < \infty \quad \Rightarrow \quad f \sim \mathcal{GP}(0, k) \in \mathcal{H}_k^\theta \text{ with prob. 1}$$

Non-representative Example: Let $k_\lambda(a, b) = \exp(-(a-b)^2/(2\lambda^2))$. Then $f \sim \mathcal{GP}(0, k_\lambda)$ is in $\mathcal{H}_{k_{\theta\lambda}}$ with prob. 1 for all $0 < \theta < 1$. The situation is more complicated for other kernels.

GP **samples are not in the RKHS**. They belong to a kind of "completion" of the RKHS (but that completion can be strictly larger than the RKHS).

- GP and Kernel Methods are very closely related
  - the RKHS is the space of all possible posterior mean functions
  - the posterior mean is the ℓ_2-least-squares estimate in the RKHS
  - the posterior variance (expected square error) is the **worst-case** error of bounded norm in the RKHS
  - GP samples are not in the RKHS

If GP's / kernel machines use infinitely many features, can they learn *every* function?

How powerful are kernel/GP models?

[Micchelli, Xu, Zhang, JMLR 7 (2006) 2651–2667]

- For some kernels, the RKHS "lies dense" in the space of all continuous functions (such kernels are known as "universal"). An example is the square-exponential / Gaussian / RBF kernel

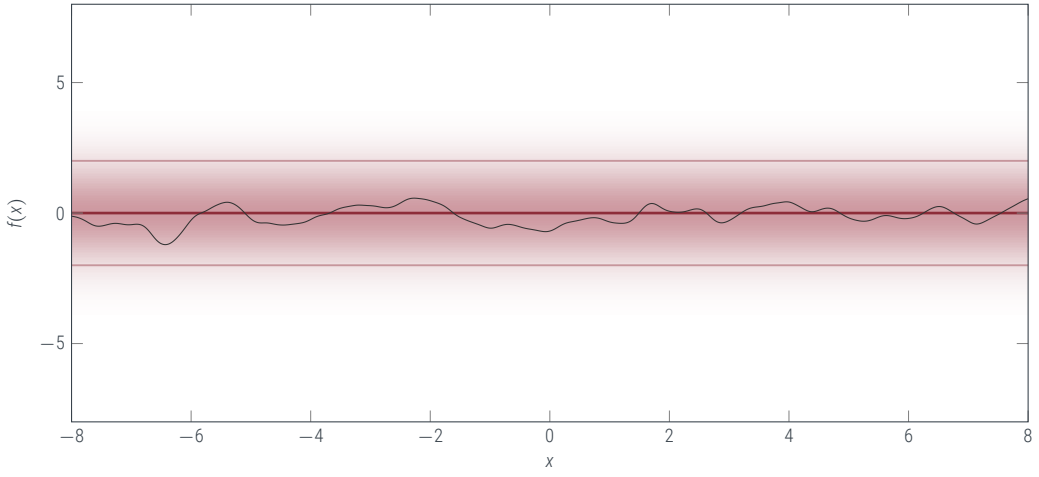
$$k(a,b) = \exp(-1/2(a-b)^2)$$

(in fact, there are many universal kernels. E.g. all stationary kernels with power spectrum of full support.)

- When using such kernels for GP / kernel-ridge regression, for any continuous functions f, for any $\epsilon > 0$ there is an RKHS element $\hat{f} \in \mathcal{H}_k$ such that $\|f - \hat{f}\| < \epsilon$ (where $\|\cdot\|$ is the maximum norm on a compact subset of $\mathbb{X}$).
- that is: Given enough data, the GP posterior mean can approximate *any function* arbitrarily well!

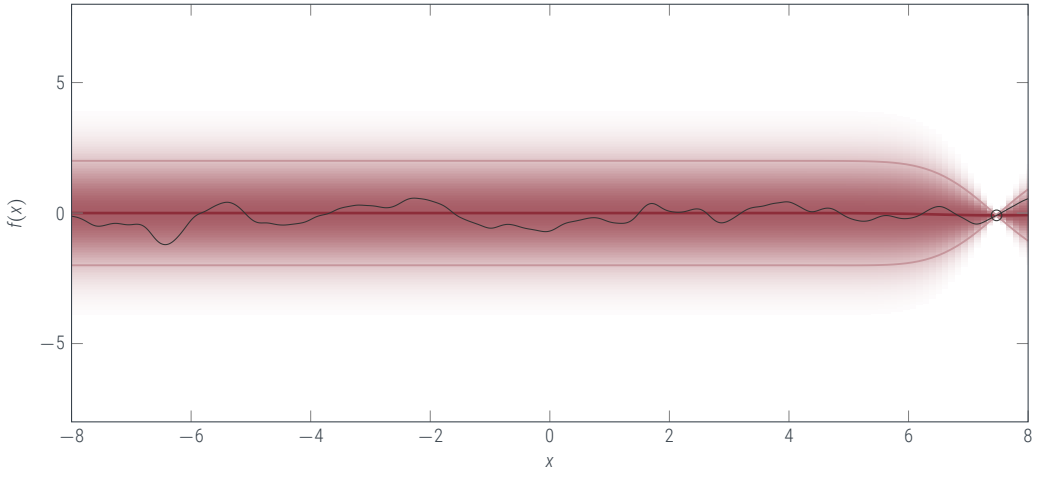
The bad news

if f is not in the RKHS – prior



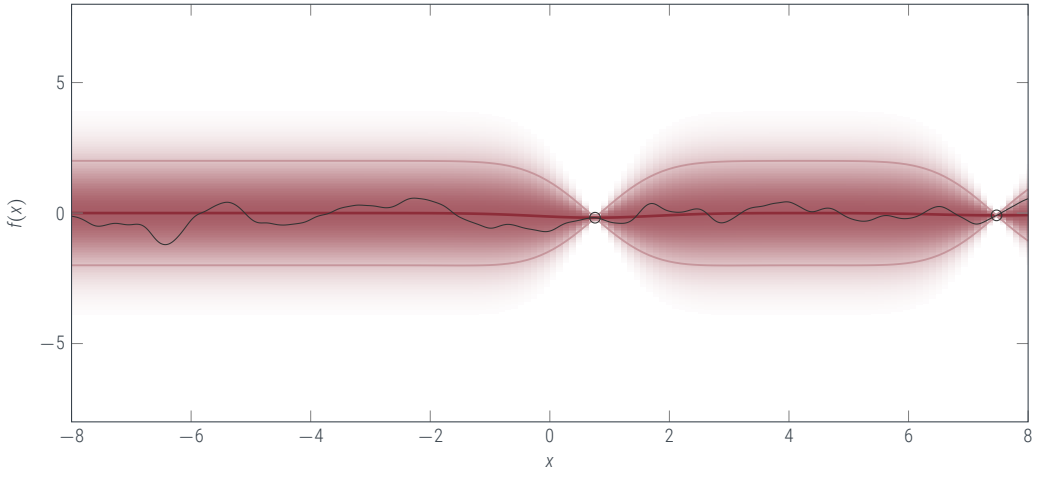
The bad news

if f is not in the RKHS – 1 evaluation



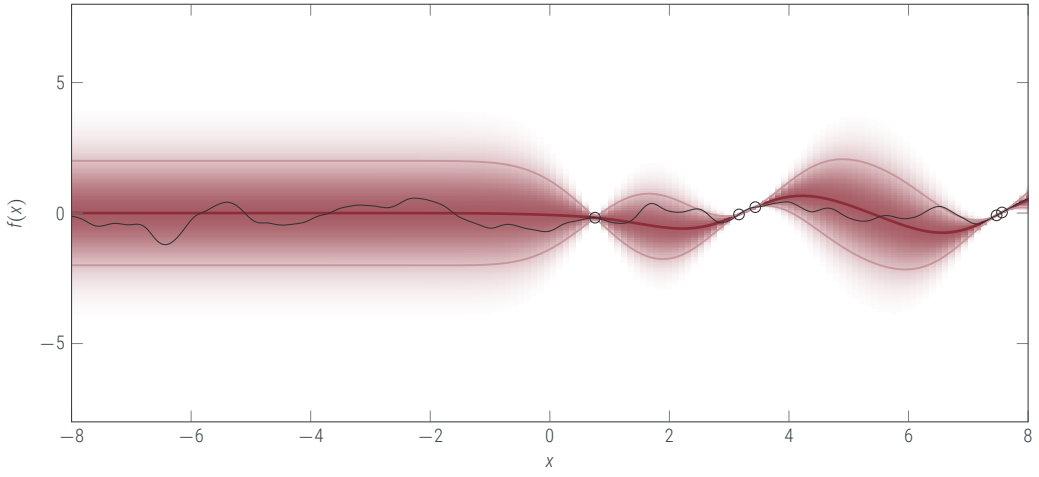
The bad news

if f is not in the RKHS – 2 evaluations



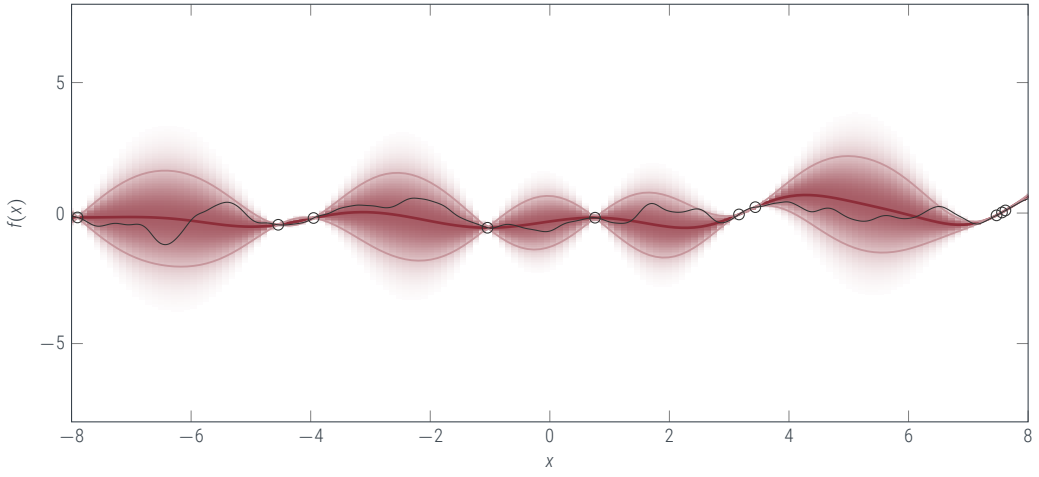
The bad news

if f is not in the RKHS – 5 evaluations



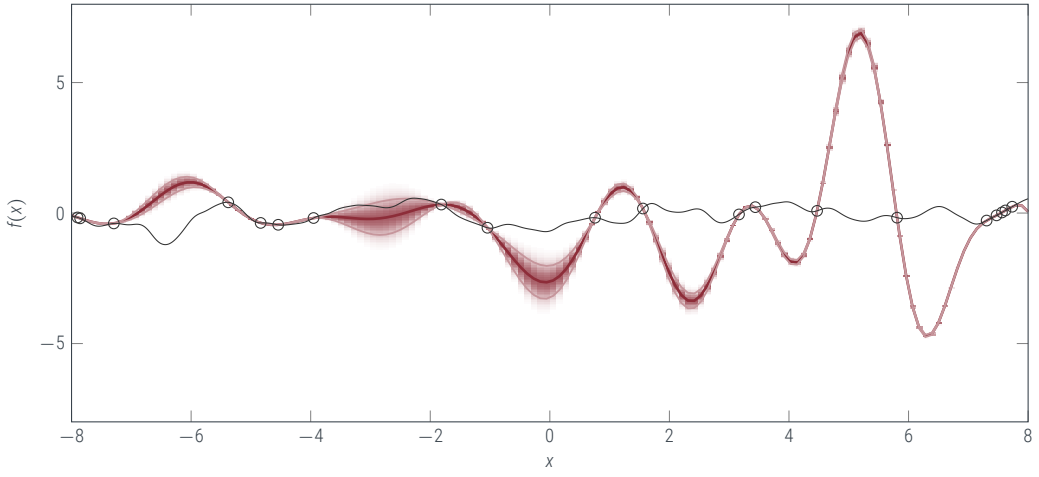
The bad news

if f is not in the RKHS – 10 evaluations



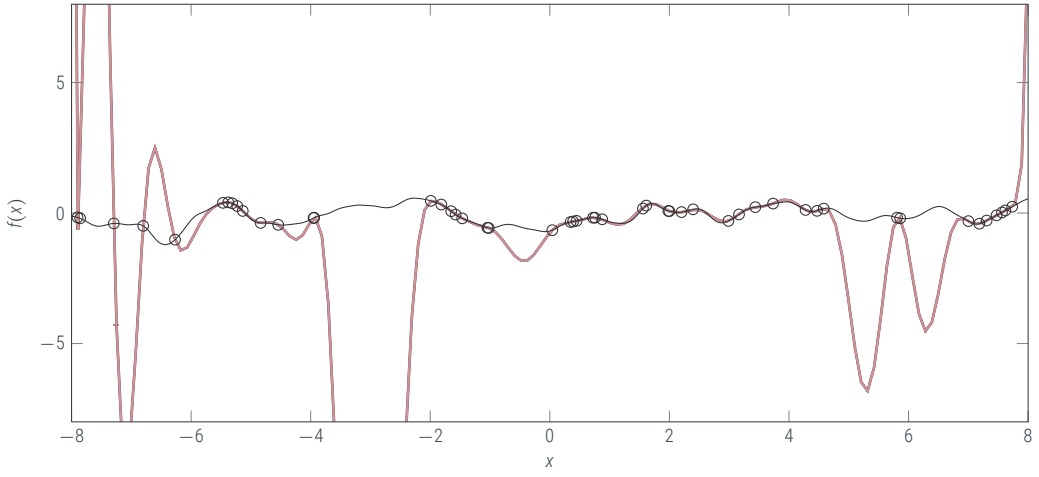
The bad news

if f is not in the RKHS – 20 evaluations



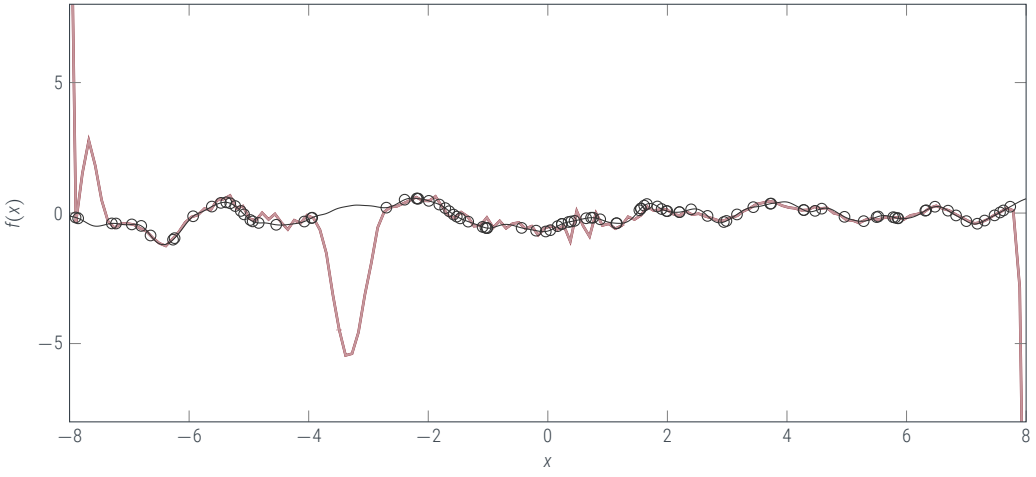
The bad news

if f is not in the RKHS – 50 evaluations



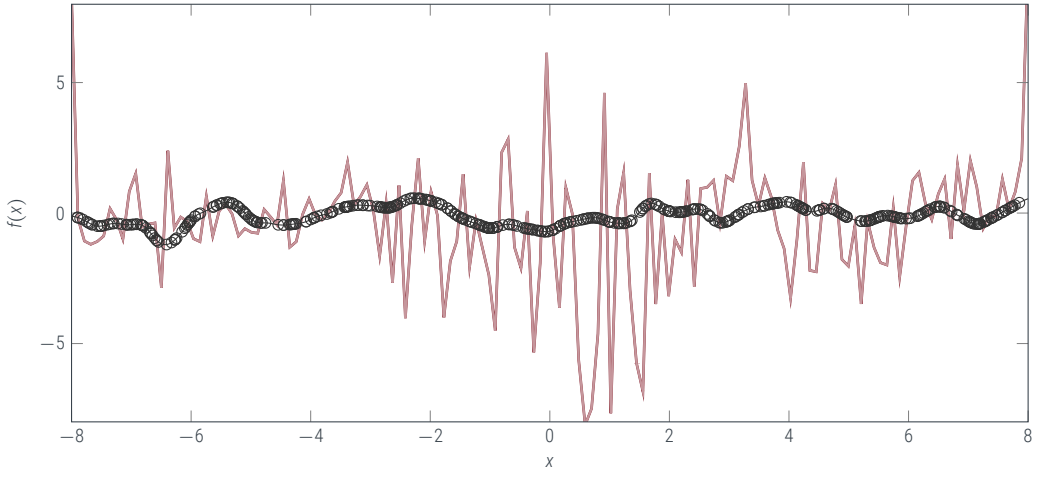
The bad news

if f is not in the RKHS – 100 evaluations



The bad news

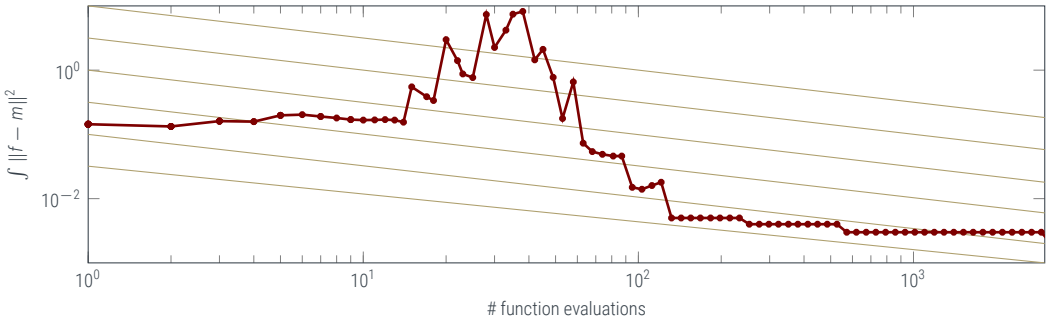
if f is not in the RKHS – 500 evaluations



Convergence Rates are Important

non-obvious aspects of f can ruin convergence

v.d.Vaart & v.Zanten. *Information Rates of Nonparametric GP models*. JMLR 12 (2011)



If f is "not well covered" by the RKHS, the number of datapoints required to achieve ϵ error can be exponential in ϵ. Outside of the observation range, there are no guarantees at all.

An Analogy

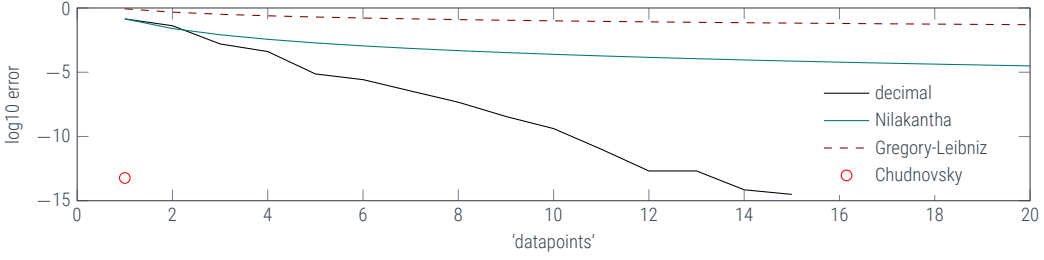
representing π in $\mathbb{Q}$

- $\mathbb{Q}$ is dense in $\mathbb{R}$

$$\pi = 3 \cdot \frac{1}{1} + 1 \cdot \frac{1}{10} + 4 \cdot \frac{1}{100} + 1 \cdot \frac{1}{1000} + \dots \quad \text{decimal}$$

$$= 4 \cdot \frac{1}{1} - 4 \cdot \frac{1}{3} + 4 \cdot \frac{1}{5} - 4 \cdot \frac{1}{7} + \dots \quad \text{Gregory-Leibniz}$$

$$= 3 \cdot \frac{1}{1} + 4 \cdot \frac{1}{2 \cdot 3 \cdot 4} - 4 \cdot \frac{1}{4 \cdot 5 \cdot 6} + 4 \cdot \frac{1}{6 \cdot 7 \cdot 8} \quad \text{Nilakantha}$$



But if you're patient, you can learn anything!

The good news.

[wording from Kanagawa et al., 2018]

Theorem (v.d. Vaart & v. Zanten, 2011)

Let f_0 be an element of the Sobolev space $W_2^\beta[0,1]^d$ with $\beta > d/2$. Let k_s be a kernel on $[0,1]^d$ whose RKHS is norm-equivalent to the Sobolev space $W_2^s([0,1]^d)$ of order $s := \alpha + d/2$ with $\alpha > 0$. If $f_0 \in C^\beta([0,1]^d) \cap W_2^\beta([0,1]^d)$ and $\min(\alpha, \beta) > d/2$, then we have

$$\mathbb{E}_{\mathcal{D}_n | f_0}\left[\int \|f - f_0\|^2_{L_2(P_\mathbb{X})} d\Pi_n(f|\mathcal{D}_n)\right] = O(n^{-2\min(\alpha,\beta)/(2\alpha+d)}) \quad (n \to \infty), \tag{1}$$

where $\mathbb{E}_{X,Y|f_0}$ denotes expectation with respect to $\mathcal{D}_n = (x_i, y_i)_{i=1}^n$ with the model $x_i \sim P_\mathbb{X}$ and $p(\boldsymbol{y} \mid f_0) = \mathcal{N}(\boldsymbol{y}; f_0(X), \sigma^2 I)$, and $\Pi_n(f|\mathcal{D}_n)$ the posterior given by GP-regression with kernel k_s.

The Sobolev space $W_2^s(\mathbb{X})$ is the vector space of real-valued functions over $\mathbb{X}$ whose derivatives up to s-th order have bounded L_2 norm. $L_2(P_\mathbb{X})$ is the Hilbert space of square-integrable functions with respect to $P_\mathbb{X}$.

If f_0 is from a sufficiently smooth space, and H_k is "covering" that space well, then the entire GP posterior (including the mean!) can contract around the true function at a **linear** rate.
GPs are "infinitely flexible": They can learn infinite-dimensional functions arbitrarily well!

- **Gaussian process regression** is closely related to **kernel ridge regression**.
  - the posterior mean is the kernel ridge / regularized kernel least-squares estimate in the RKHS $\mathcal{H}_k$.

$$m(x) = k_{xX}(k_{XX} + \sigma^2 I)^{-1} \boldsymbol{y} = \operatorname*{arg\,min}_{f \in \mathcal{H}_k} \|\boldsymbol{y} - f_X\|^2 + \|f\|^2_{\mathcal{H}_k}$$

  - the posterior variance (**expected square error**) is the **worst-case square error** for bounded-norm RKHS elements.

$$v(x) = k_{xx} - k_{xX}(k_{XX})^{-1} k_{Xx} = \operatorname*{arg\,max}_{f \in \mathcal{H}_k, \|f\|_{\mathcal{H}_k} \leq 1} \|f(x) - m(x)\|^2$$

- Similar connections apply for most **kernel methods**.
- GPs are quite powerful: They can learn any function in the RKHS (a large, generally infinite-dimensional space!)
- GPs are quite limited: If $f \notin \mathcal{H}_k$, they may converge **very** (e.g. exponentially) slowly to the truth.
- But if we are willing to be cautious enough (e.g. with a rough kernel whose RKHS is a Sobolev space of low order), then polynomial rates are achievable. (Unfortunately, exponentially slow in the dimensionality of the input space)