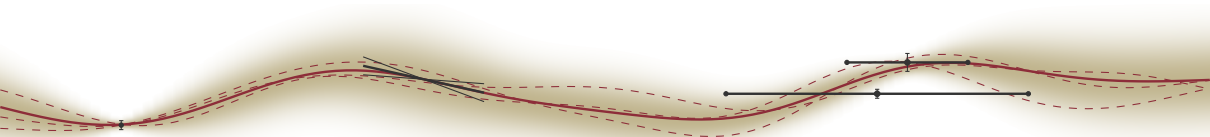
Probabilistic Inference and Learning
Lecture 12
More Theory on Gaussian Models

Philipp Hennig

26 November 2018

Eberhard Karls
Universität
Tübingen

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



Overview of Lectures so far:

0. Introduction to Reasoning under Uncertainty
1. Probabilistic Reasoning
2. Probabilities over Continuous Variables
3. Gaussian Probability Distributions
4. Gaussian Parametric Regression
5. More on Parametric Regression
6. Gaussian Processes
7. More on Kernels & GPs
8. A practical GP example
9. Markov Chains, Time Series, Filtering
10. Classification
11. Empirical Example of Classification
12. More Connections to SLT (TODAY)
13. Theory of Filtering / SDEs (NEXT MON)
14. beyond Gaussians …

Today:

- Frequentist interpretation of Gaussian process regression
- A philosophical comparison of Bayesian and Statistical Learning

Recap from Lecture 7:

- Kernels are "a bit like infinitely large matrices" in the sense of Mercer's theorem: They can be thought of as an infinite sum over orthogonal and normalizable *eigenfunctions* with positive eigenvalues. However, this interpretation is only true relative to a base measure
- Gaussian (process) posterior means are ℓ_2-regularized **least-squares** estimates, and thus connected to a deep and long history of estimation in the sciences

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 meaning of the GP point estimate?

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

Theorem (The Kernel Ridge Estimate)

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

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

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: follows on next two slides

Proof of the Kernel Ridge Estimate

Part I: The representer theorem

Theorem (Schölkopf, Herbrich, Smola, 2001)

Consider an RKHS $\mathcal{H}_k$. Any function $f \in \mathcal{H}_k$ minimizing the regularized risk functional

$$c((x_1, y, f(x_1)), \ldots, (x_n, y_n, f(x_n))) + g(\|f\|_{\mathcal{H}_k})$$

with a strictly monotonic g, admits a representation of the form

$$f(\cdot) = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i)$$

Proof:

- Write $f(\cdot) = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i) + v$ with $\langle v, k(\cdot, x_i) \rangle = 0 \; \forall x_i$
- note $f(x_j) = \left\langle \sum_{i=1}^{n} \alpha_i k(\cdot, x_i) + v, k(\cdot, x_j) \right\rangle = \sum_i \alpha_i \langle k(x_i, \cdot), k(\cdot, x_j) \rangle = \sum_i \alpha_i k(x_i, x_j)$. Thus c is independent of v.
- Because g is strictly monotonic:

$$\begin{aligned} g(\|f\|) &= g\left(\left\| \sum_{i=1}^{n} \alpha_i k(\cdot, x_i) + v \right\| \right) \\ &= g\left(\sqrt{\left\| \sum_{i=1}^{n} \alpha_i k(\cdot, x_i) \right\|^2 + \|v\|^2} \right) \\ &\geq g\left(\left\| \sum_{i=1}^{n} \alpha_i k(\cdot, x_i) \right\| \right) \end{aligned}$$

- thus $v \neq 0$ does not affect c and increases v. Hence $v = 0$. □

Proof of the Kernel Ridge Estimate

Part II: The ℓ_2 case

Theorem (The Kernel Ridge Estimate)

Consider the Bayesian regression 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

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

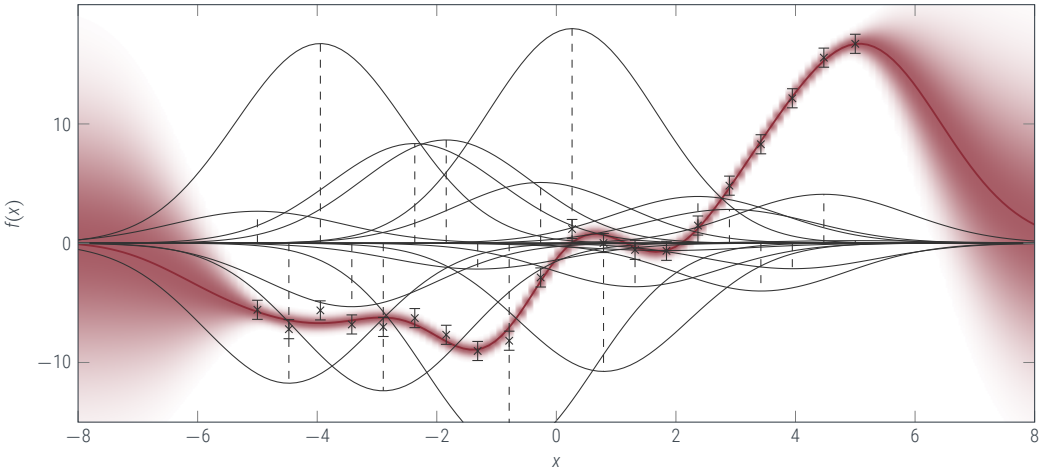
- By representer theorem, can write $\arg\min_{f \in \mathcal{H}_k} L(f) = \sum_i \alpha_i k(\cdot, x_i)$
- plug into $L(f)$, get (note $f(x_i) = \langle f(\cdot), k(\cdot, x_i)\rangle$ and $\langle k(x_i, \cdot), k(x_j, \cdot)\rangle = k(x_i, x_j)$)

$$\arg\min_{\alpha \in \mathbb{R}^n} \sigma^{-2}(\alpha^\intercal k_{XX} k_{XX} \alpha - 2\alpha^\intercal k_{XX} y + \|y\|^2) + \alpha^\intercal k_{XX} \alpha$$

- Differentiate, see that $m(x)$ gives zero gradient, not problem is convex. □

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 uncertainty?

Frequentist interpretation of the posterior variance

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

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

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 $\mathcal{H}_k$.

Reminder: Samples are tricky!

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!):

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

Reminder: Samples are tricky!

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!):

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

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,

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

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.

Technically, GP **samples are not in the RKHS**. However, in practice we can **usually** ignore the distinction and think of GP samples as RKHS functions.

- 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, but "almost"

How powerful are kernel/GP models?

first, the dangerous hope

[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

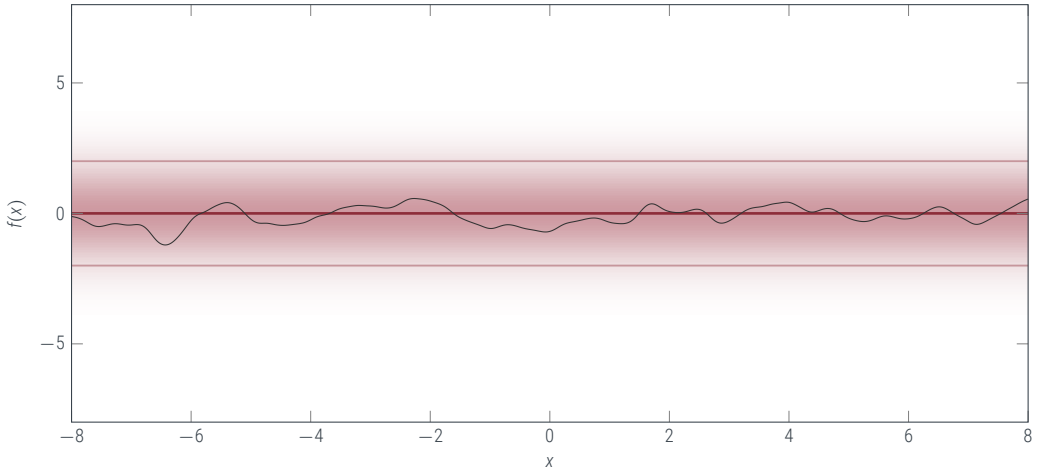
$$k(a,b) = \exp(-1/2(a-b)^2)$$

(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!

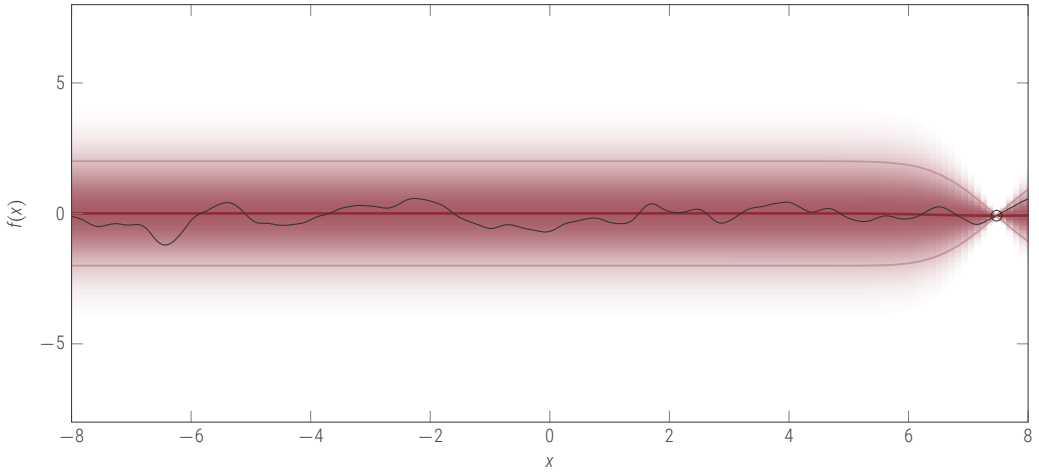
The bad news

if f is not in the RKHS – prior



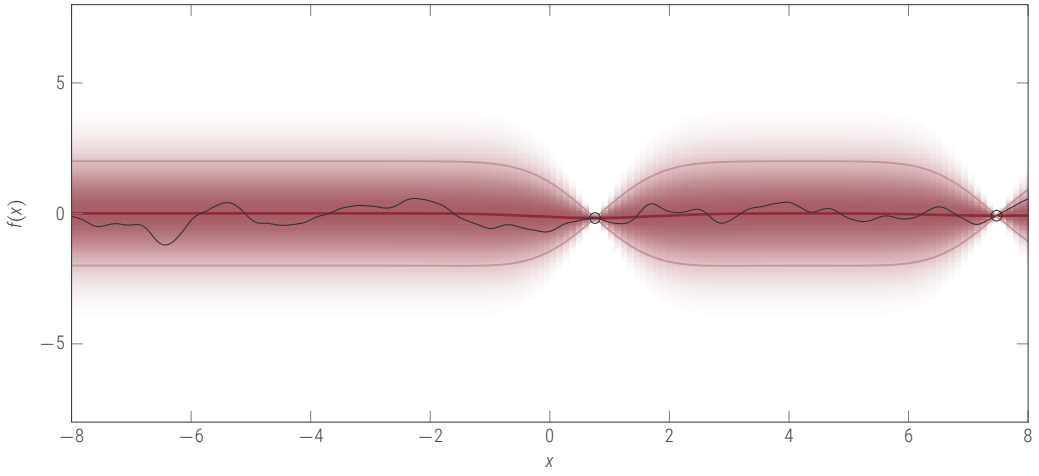
The bad news

if f is not in the RKHS – 1 evaluation



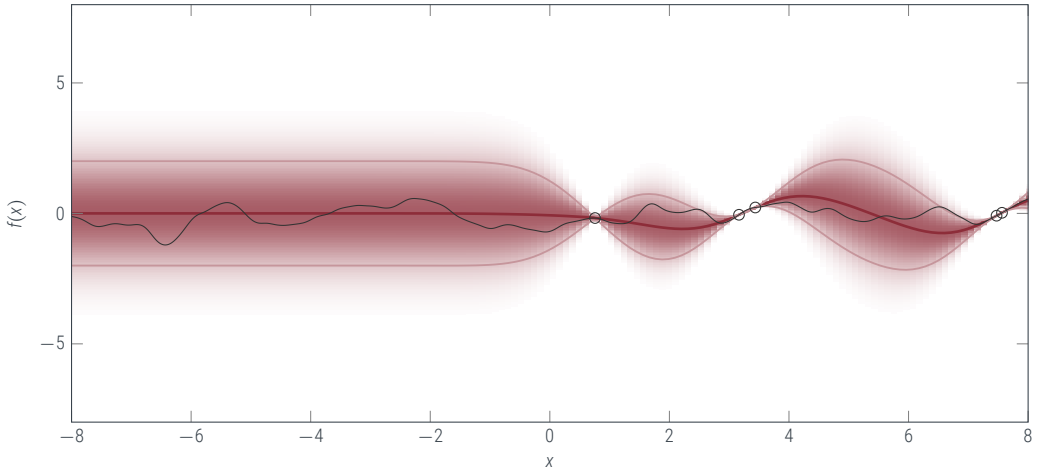
The bad news

if f is not in the RKHS – 2 evaluations



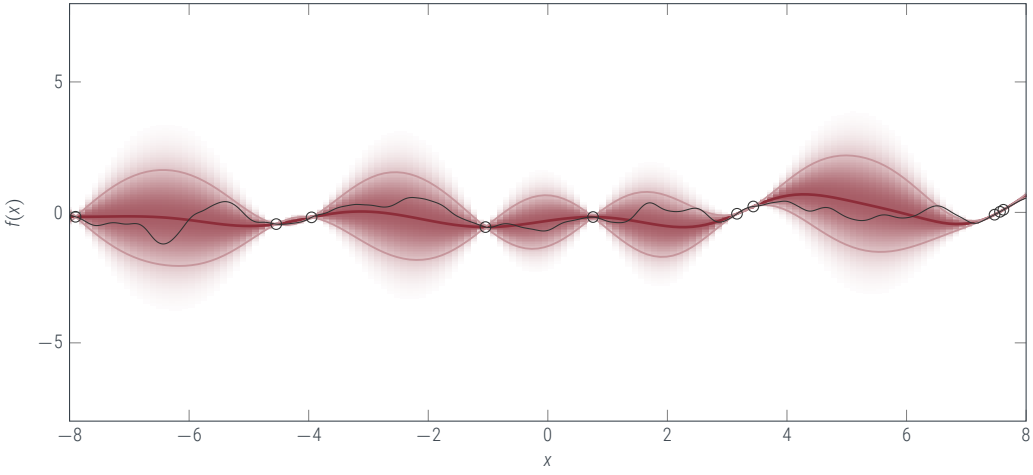
The bad news

if f is not in the RKHS – 5 evaluations



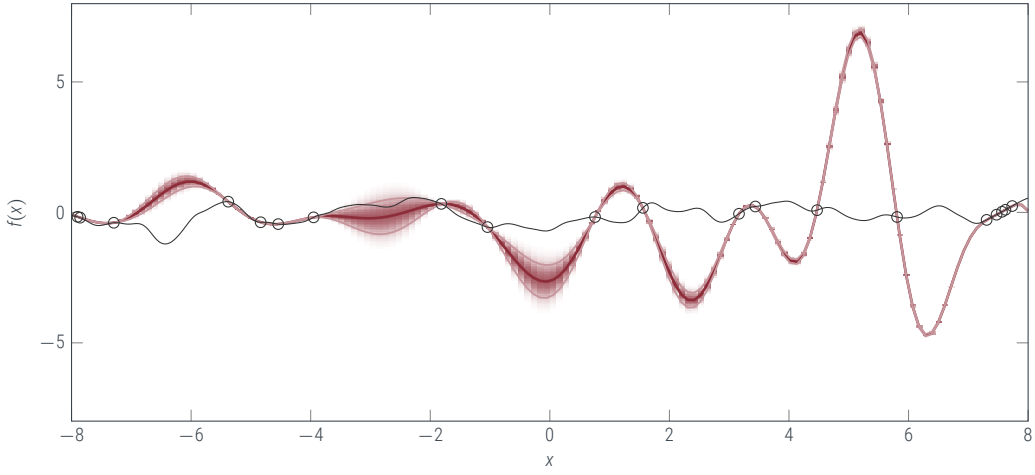
The bad news

if f is not in the RKHS – 10 evaluations



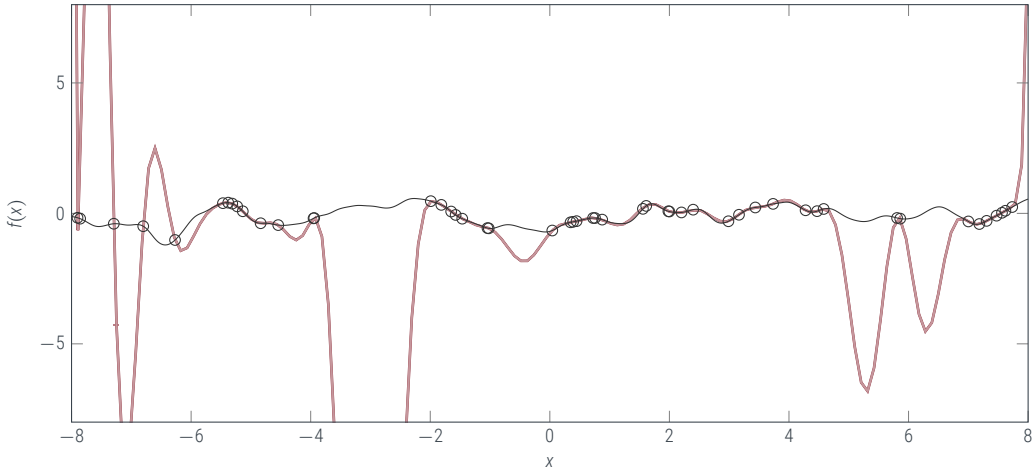
The bad news

if f is not in the RKHS – 20 evaluations



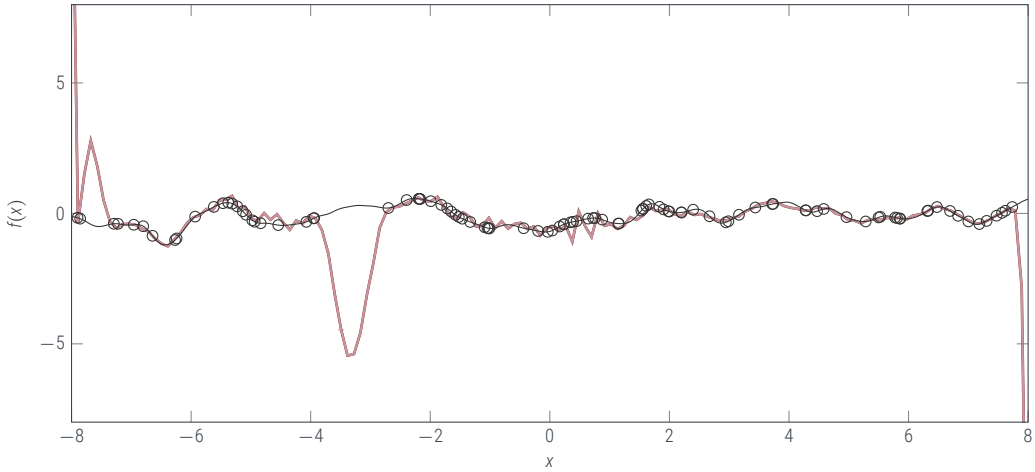
The bad news

if f is not in the RKHS – 50 evaluations



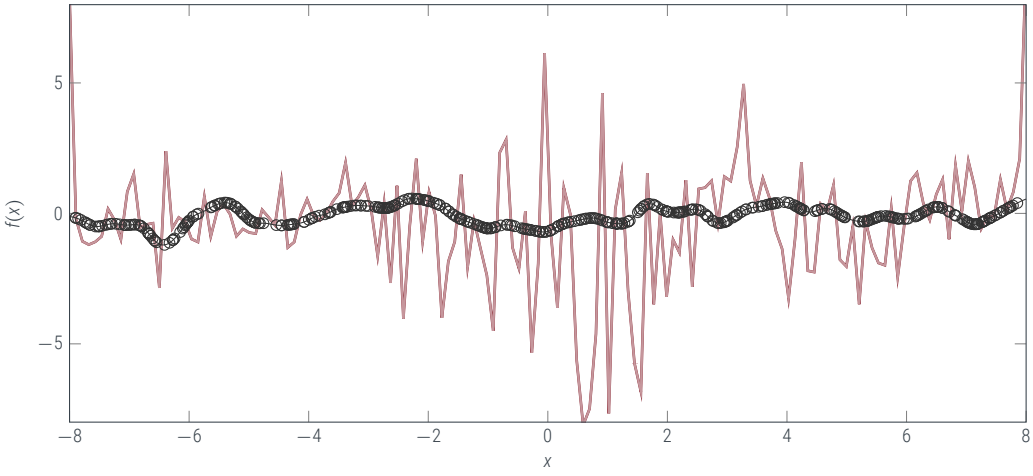
The bad news

if f is not in the RKHS – 100 evaluations



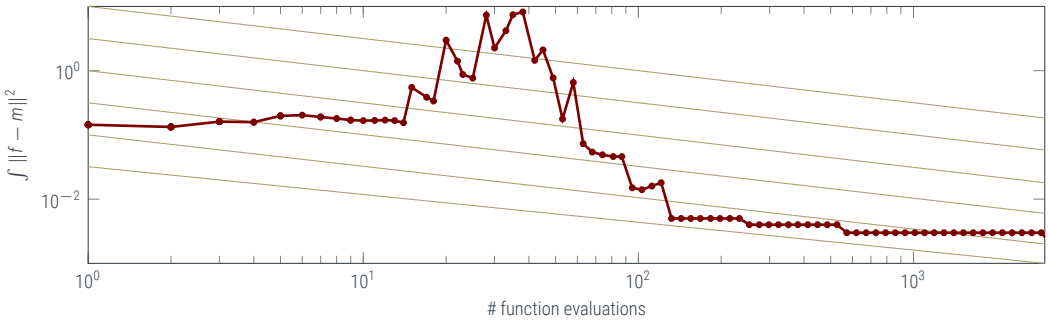
The bad news

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

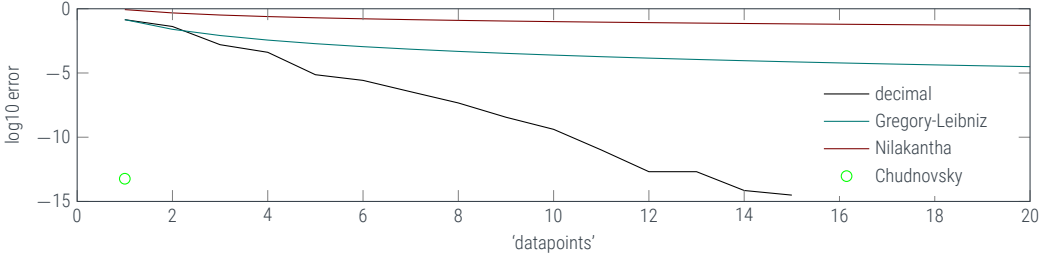
representing π in $\mathbb{Q}$

- $\mathbb{Q}$ 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 $\mathcal{H}_k$.

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

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

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

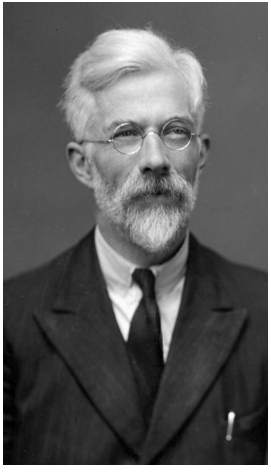
- Similar connections apply for most **kernel methods**.
- GPs are quite powerful: They can learn any function in the RKHS (a large, generally infinite-dimensional space!)
- GPs are quite limited: If $f \notin \mathcal{H}_k$, they may converge **very** (e.g. exponentially) slowly to the truth.
- But 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)

Bayesianism vs. Frequentism

not a philosophical note

The theory of inverse probability [= Bayesian inference] is founded upon an error, and must be wholly rejected.

Sir R.A. Fisher (1890-1962)



What we talk about when we talk about frequentists and Bayesians

often an irrationally emotional debate

[cf. Lecture 0]

Statistical Learning Theory (“Frequentism”)	Probabilistic Learning (“Bayesianism”)
Practical Considerations:	
formulate a loss-function *ad hoc*, mapping data to predictions/decisions. Then show that, under some external assumptions, this model has certain desirable properties.	formulate a **generative model**. Inference is then uniquely determined by Bayes’ theorem. No need to question or analyse the paradigm over and over again
mathematical **analysis** in the foreground (often in the asymptotic or large-number limit)	**numerical & computational** design in the foreground (the right model may be intractable)
statements about errors tend to focus on the **worst-case**	structured and extensive **quantification of uncertainty** by the posterior, often core motivation
Philosophical Considerations:	
probability as *frequency*	probability as *uncertainty*

A Direct Comparison

In the case of linear regression, the two frameworks are very close.

A supervised learning task:

We are given $D = (y_i, x_i)_{i=1,\dots,n} \subset \mathbb{R} \times \mathbb{X}$. Construct a function $f : \mathbb{X} \to \mathbb{R}$ that models $y_i \approx f(x_i)$.

A Direct Comparison

In the case of linear regression, the two frameworks are very close.

A supervised learning task:

We are given $D = (y_i, x_i)_{i=1,\dots,n} \subset \mathbb{R} \times \mathbb{X}$. Construct a function $f : \mathbb{X} \to \mathbb{R}$ that models $y_i \approx f(x_i)$.

Frequentist Approach:

- Make **choices:**

  choose an **empirical risk** $\quad \ell(y_i, x_i, f) = \frac{1}{2}(y_i - f(x_i))^2$

  choose a **model class** $\quad f \in \mathcal{H}_k$

  choose a **regularizer** $\quad r(f) = \frac{\sigma^2}{2}\|f\|^2_{\mathcal{H}_k}$

- choose an inference algorithm / estimator:

$$\hat{f}(\cdot) = \arg\min_{f \in \mathcal{H}_k} \sum_i^n \ell(y_i, x_i, f) + r(f) = k_{\cdot X}(k_{XX} + \sigma^2 I)^{-1} \boldsymbol{y}$$

- **analyse** the behaviour of this estimator under some **assumptions** (here, just a simple example):

  Assuming $f \in \mathcal{H}_k, \|f\|_{H_k} < c,$ and $y = f(x),$ then $\|\hat{f}(x) - f(x)\|^2 < cV(x) =: c(k_{xx} - k_{xX}K_{XX}^{-1}k_{Xx})$

  and $V(x)$ can be shown to contract as $\mathcal{O}(\dots)$

A Direct Comparison

In the case of linear regression, the two frameworks are very close.

A supervised learning task:

We are given $D = (y_i, x_i)_{i=1,\dots,n} \subset \mathbb{R} \times \mathbb{X}$. Construct a function $f : \mathbb{X} \to \mathbb{R}$ that models $y_i \approx f(x_i)$.

Bayesian Approach:

- Make **assumptions:** (ideally, use prior knowledge)

assume a likelihood	$p(\boldsymbol{y} \mid X, f) = \mathcal{N}(\boldsymbol{y}; f_X, \sigma_2 I)$
assume a prior	$p(f) = \mathcal{GP}(0, k)$

- there is no choice of inference rule! Bayes Theorem dictates:

$$p(f \mid y, x) = \mathcal{GP}(f; \hat{f}, V(x)).$$

- **question** the generative model (prior & likelihood): In which sense to the deviate from actual prior knowledge?

A Direct Comparison

In the case of linear regression, the two frameworks are very close.

For the Bayesian:

- All assumptions are spelt out in the generative model. If the generative model is known to be wrong, don't analyse the effect of this mismatch, change the model! If necessary, perform hierarchical inference.
- There is no debate over the inference rule.
- Error estimation / uncertainty are a central part of the process. That also means the error estimates are just as (un-) reliable as the point estimates.
- One may still wonder about the effect of **unknown** model mismatches. But this usually has to happen by external (statistical/frequentist) analysis.

For the Frequentist:

- Everything is a choice, even the inference rule. Thus, everything has to be analysed extensively to make sure it works under certain assumptions. But the assumptions are somewhat external: They are used as an argument in favour of an algorithm, not seen as a part of the algorithm (if the assumptions don't hold, then the analysis doesn't hold. That doesn't mean the algorithm ist wrong). Such analysis often takes place in the asymptotic limit (hopefully with rates!)
- Error estimation is an external process and can be done in many different ways. Actually estimating a concrete numerical error is usually impossible without giving up the moral high ground.

Some Stubborn Misconceptions

100 years after Fisher, the Bayesian - Frequentist divide still lives on

Don't believe the following statements!

- **Bayesian methods are expensive!**
  - Sample-efficient inference is expensive. Simple Bayesian methods are just as fast as simple frequentist methods. There are also very expensive frequentist methods.
- **Bayesian make more assumptions that Frequentists!**
  - Bayesians state their assumptions very precisely, directly in the inference algorithm. In Frequentist analysis, assumptions are often hidden in the analysis, rather than the algorithm. If assumptions are weaker than in the Bayesian form, they usually lead to vaguer statements with reduced practicaly value (e.g. asymptotic rates with unknown constant, instead of explicit error estimates).
- **Bayesian inference is subjective, thus non-scientific**
  - Bayesian inference is the mathematical formalization of the scientific method. All human knowledge is fundamentally subjective. But Bayesian assumptions can be questioned (inside and outside of the framework) simply because they are explicit in the formulation. There is no inference framework that makes nontrivial statements without prior assumptions.

The value of probabilistic formulations

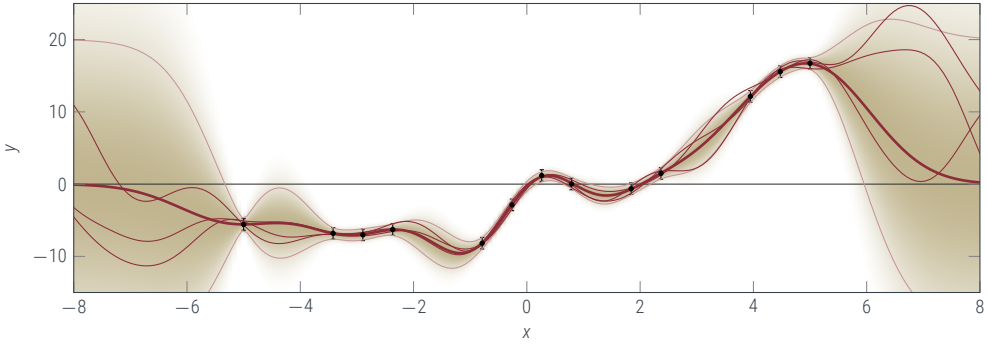
with caveats

When [there are few parameters, sufficient statistics are present, there is no important prior information, and no nuisance parameters,] and we have a reasonably large amount of data, orthodox methods become essentially equivalent to the Bayesian ones, and it will make not pragmatic difference which ideology we prefer. But today we are faced with important problems in which some or all of these conditions are violated. Only Bayesian methods have the analytical apparatus capable of dealing with such problems without sacrificing much of the relevant information available to us.

E.T. Jaynes (1922-1998)
Probability Theory — The Logic of Science, §17.12, p.550

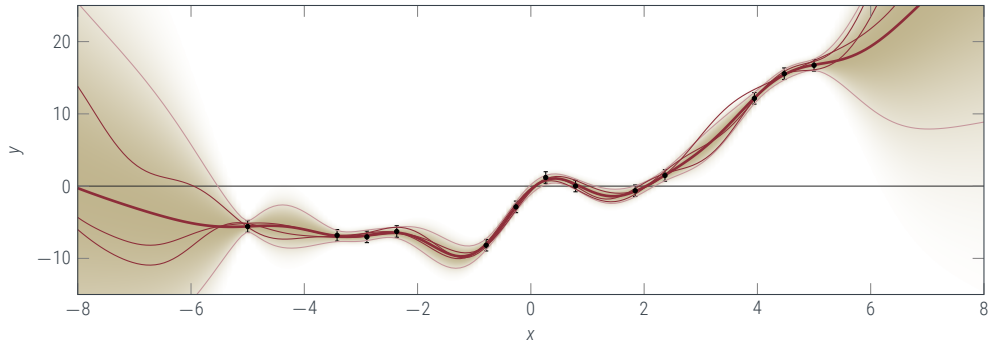
Strengths of the Bayesian formulation

generality and automation



Strengths of the Bayesian formulation

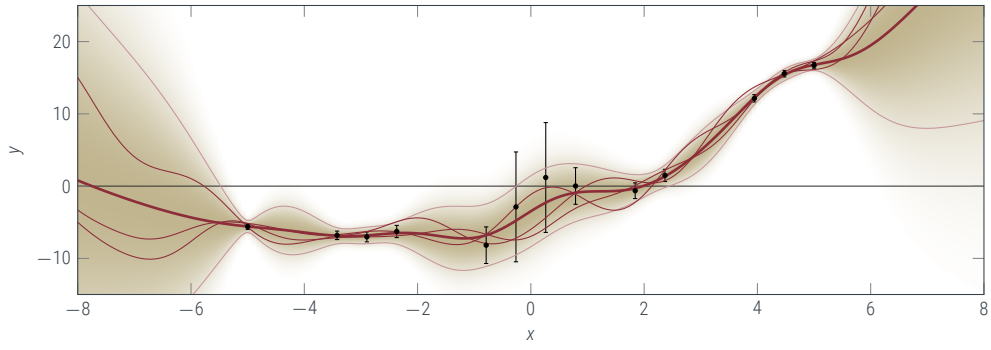
generality and automation



- If you know something unusual about the function, just change the prior!

Strengths of the Bayesian formulation

generality and automation



- If you know something unusual about the function, just change the prior!
- If you know something unusual about the data, just change the likelihood!
- All this may cause computational overhead. Solving this is up to you.

- **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 $\mathcal{H}_k$.

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

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

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

- 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. logarithmically) slowly to the truth
- Bayesians v. Frequentists
  - Probabilistic formulations can be very similar to "frequentist" ones
  - Some kinds of analysis are more natural in the frequentist language
  - Modeling, on the other hand, is often easier in the probabilistic language