Probabilistic Inference and Learning
Lecture 07
Understanding Kernels: Connections to Statistical Learning

Philipp Hennig

07 November 2018

Eberhard Karls
Universität Tübingen

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

A few remarks on the exercises

or how to get full points

- Make sure you **write clearly**, if we can't read it, we can't give points for it.
- If you only write the solution, but do not **explain** how you arrived there, we can't give you full points.
- It might be a good idea to collect all your corrected exercise sheets. We might make mistakes when keeping track of all of your points, so **keep the corrected exercise sheets** as proof. This will also make sure, we did not miss your exercise sheet by accident.

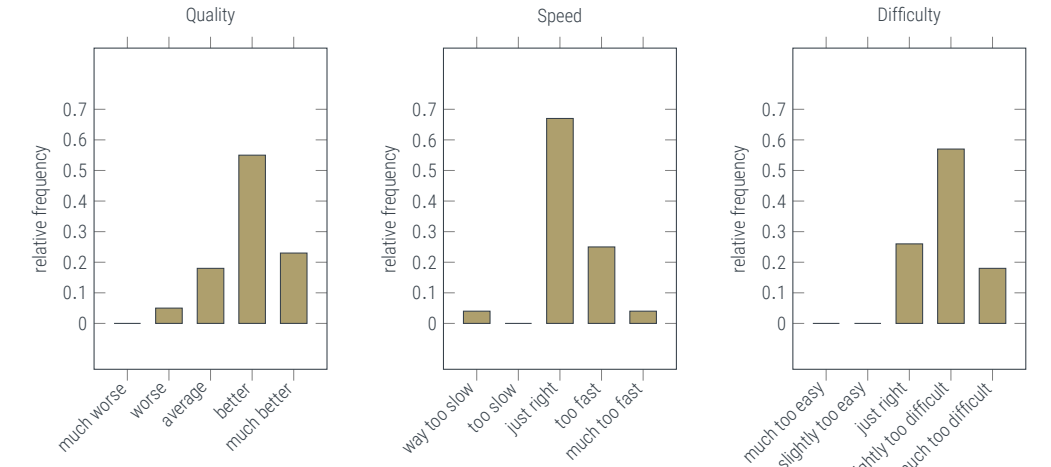
A few remarks on the exercises

regarding Ilias and uploading

- Upload your solutions to Ilias **before the end of the lecture** (before 10:00). We start correcting immediately after, and will not include late-comers in the future.
- **Put your name on all sheets and files**. This includes the sheets you hand in in the lecture, but also electronically, as well as code/jupyter notebooks (include a comment in the first line).
- Uploads to Ilias should be in the following format ***lastname_firstname_sheet_number***, e.g. mustermann_max_04.pdf for this weeks sheet. If you add code, just add _code to the end of the filename. Preferably, but everything in a zip file with the same name.
- The folder *Upload your Exercise Sheets here* has subfolders for the groups and the sheets (using the new numbering system, so this weeks sheet will be sheet 4). Make sure you **upload to the correct subfolder**, or we will ignore it.
- We prefer LaTeX**pdfs** and **Jupyter Notebooks**.

Last Lecture: Debrief

Feedback dashboard



Last Lecture: Debrief

Detailed Feedback

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Things you did not like:

- "This is all incredibly fascinating, but I can not wrap my head around those formulas"
- "That is it 8am on a Monday"
- "The lecture hall is too cold!"

Things you did not understand:

- if there are ∞ features, what are the weights now? What are we optimizing?
- "How do I evaluate new data points with an existing kernel?"
- **the 5-th point in the toolbox**
- learning the kernel
- pls. explain GP again
- *xxXXxXxxxX*
- what does $\oint \sum$ mean?
- variable names in code and on slides are not identical

Things you enjoyed:

- **pointing out relation to other lecture content / Summaries**
- **insight about closed-form integrals** (interesting new angle rel. to v. Luxburg's lecture)
- **code**

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 — Connections to Deep Learning
6. Gaussian Processes

Today:

- Theory of Kernels and Kernel Methods
- Gaussian inference from the statistical perspective
- Limits of nonparametric learning

Plan for the coming weeks: Gaussian Models

Wed, 7 November	✦ some theory of GPs ✦ connecting to the language of kernel methods
Mon, 12 November	✦ A concrete example for a nontrivial GP model ✦ intuition for the power and the limitations of GP models
Wed, 15 November	✦ In time series, GP inference can be **linear cost** ✦ building a connection to **signal processing**
Mon, 19 November	✦ **Classification** can be done with GPs, too ✦ this requires some painful derivations, but it yields **the world's most widely used ML algorithm!**
Wed, 21 November	✦ a concrete example of GP classification ✦ some beautiful python code
Mon, 26 November	✦ summary of GP models ✦ tying up various loose ends, time for open questions

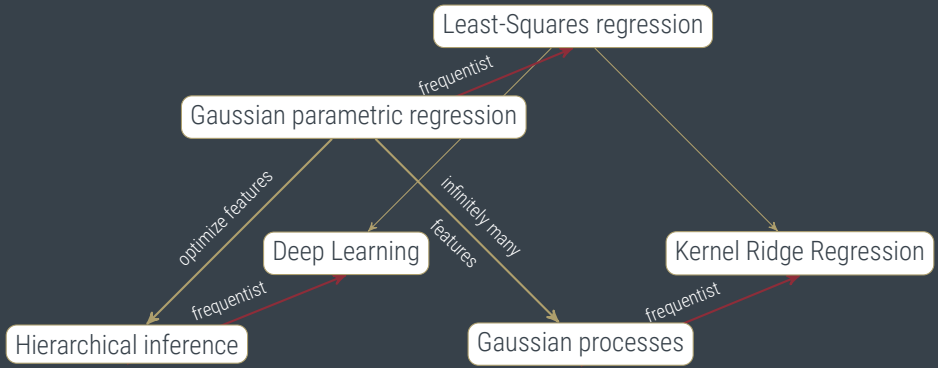
Today's lecture features advanced math.

Goals:

- Building a connection to inference and estimation in the sciences
- Building a mental connection to Statistical Learning Theory and "Frequentist" formulations of Machine Learning. Understanding that both have value, and need not mutually exclude each other
- Expressive power and limitations of kernel methods/GPs

Connections

- We continue to focus on **supervised learning**:
  Given **pairs** $(X, Y) := [(x_i, y_i)]_{i=1,\dots,n}$, infer f such that $f(x_i) \approx y_i$
- Gaussian parametric models: $f(x) = \phi(x)^\intercal \boldsymbol{w} \qquad p(w) = \mathcal{N}(w; \mu, \Sigma)$
- Gaussian process models / nonparametric regression:s $p(f) = \mathcal{GP}(f; k, m)$
- Learning the features is related to **deep learning**



Reminder: Where we left off last time

Gaussian Process Regression

$$p(f) = \mathcal{GP}(f; m, k) \qquad p(\boldsymbol{y} \mid f) = \mathcal{N}(\boldsymbol{y}; f_X, \Lambda)$$

$$\Rightarrow p(f \mid \boldsymbol{y}) = \mathcal{GP}(f_x; m_x + k_{xX}(k_{XX} + \Lambda)^{-1}(\boldsymbol{y} - m_X), \quad k_{xx} - k_{xX}(k_{XX} + \Lambda)^{-1}k_{Xx})$$

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

We will also come to find close connections to Gaussian processes in later lectures, connecting it to basic concepts in

- signal processing
- control engineering
- numerical analysis
- …

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

regularized and non-regularized least-squares

$$
\begin{aligned}
p(\boldsymbol{w} \mid \boldsymbol{y}) &= \frac{p(\boldsymbol{y} \mid X, \boldsymbol{w}) p(\boldsymbol{w})}{p(\boldsymbol{y} \mid X)} = \frac{\mathcal{N}(\boldsymbol{y}; \phi_X^\intercal \boldsymbol{w}, \sigma^2 I) \mathcal{N}(\boldsymbol{w}; \boldsymbol{\mu}, \Sigma)}{\mathcal{N}(\boldsymbol{y}; \phi_X^\intercal \boldsymbol{\mu}, \phi_X^\intercal \Sigma \phi_X + \sigma^2 I)} \\
&= \mathcal{N}(\boldsymbol{w}; \boldsymbol{\mu} + \Sigma \phi_X (\phi_X^\intercal \Sigma \phi_X + \sigma^2 I)^{-1} (\boldsymbol{y} - \phi_X \boldsymbol{\mu}), \Sigma - \Sigma \phi_X (\phi_X^\intercal \Sigma \phi_X + \sigma^2 I)^{-1} \phi_X \Sigma)
\end{aligned}
$$

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

regularized and non-regularized least-squares

$$
\begin{aligned}
p(\boldsymbol{w} \mid \boldsymbol{y}) &= \frac{p(\boldsymbol{y} \mid X, \boldsymbol{w}) p(\boldsymbol{w})}{p(\boldsymbol{y} \mid X)} = \frac{\mathcal{N}(\boldsymbol{y}; \phi_X^\intercal \boldsymbol{w}, \sigma^2 I)\mathcal{N}(\boldsymbol{w}; \boldsymbol{\mu}, \Sigma)}{\mathcal{N}(\boldsymbol{y}; \phi_X^\intercal \boldsymbol{\mu}, \phi_X^\intercal \Sigma \phi_X + \sigma^2 I)} \\
&= \mathcal{N}(\boldsymbol{w}; \boldsymbol{\mu} + \Sigma\phi_X(\phi_X^\intercal \Sigma \phi_X + \sigma^2 I)^{-1}(\boldsymbol{y} - \phi_X \boldsymbol{\mu}), \Sigma - \Sigma\phi_X(\phi_X^\intercal \Sigma \phi_X + \sigma^2 I)^{-1}\phi_X \Sigma) \\
\mathbb{E}_{p(\boldsymbol{w}|\boldsymbol{y})}(\boldsymbol{w}) &= \arg\max_{\boldsymbol{w} \in \mathbb{R}^F} p(\boldsymbol{w} \mid \boldsymbol{y}) \\
&= \arg\min_{\boldsymbol{w}} -p(\boldsymbol{w} \mid \boldsymbol{y}) = \arg\min_{\boldsymbol{w}} -\log p(\boldsymbol{w} \mid \boldsymbol{y})
\end{aligned}
$$

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

regularized and non-regularized least-squares

$$
\begin{aligned}
p(\boldsymbol{w} \mid \boldsymbol{y}) &= \frac{p(\boldsymbol{y} \mid X, \boldsymbol{w}) p(\boldsymbol{w})}{p(\boldsymbol{y} \mid X)} = \frac{\mathcal{N}(\boldsymbol{y}; \phi_X^\intercal \boldsymbol{w}, \sigma^2 I)\mathcal{N}(\boldsymbol{w}; \boldsymbol{\mu}, \Sigma)}{\mathcal{N}(\boldsymbol{y}; \phi_X^\intercal \boldsymbol{\mu}, \phi_X^\intercal \Sigma \phi_X + \sigma^2 I)} \\
&= \mathcal{N}(\boldsymbol{w}; \boldsymbol{\mu} + \Sigma\phi_X(\phi_X^\intercal \Sigma \phi_X + \sigma^2 I)^{-1}(\boldsymbol{y} - \phi_X \boldsymbol{\mu}), \Sigma - \Sigma\phi_X(\phi_X^\intercal \Sigma \phi_X + \sigma^2 I)^{-1}\phi_X \Sigma) \\
\mathbb{E}_{p(\boldsymbol{w}|\boldsymbol{y})}(\boldsymbol{w}) &= \arg\max_{\boldsymbol{w} \in \mathbb{R}^F} p(\boldsymbol{w} \mid \boldsymbol{y}) \\
&= \arg\min_{\boldsymbol{w}} -p(\boldsymbol{w} \mid \boldsymbol{y}) = \arg\min_{\boldsymbol{w}} -\log p(\boldsymbol{w} \mid \boldsymbol{y}) \\
&= \arg\min_{\boldsymbol{w}} \frac{1}{2\sigma^2}\|\boldsymbol{y} - \phi_X^\intercal \boldsymbol{w}\|^2 + \frac{1}{2}(\boldsymbol{w} - \boldsymbol{\mu})^\intercal \Sigma^{-1}(\boldsymbol{w} - \boldsymbol{\mu})
\end{aligned}
$$

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

regularized and non-regularized least-squares

$$
\begin{aligned}
p(\boldsymbol{w} \mid \boldsymbol{y}) &= \frac{p(\boldsymbol{y} \mid X, \boldsymbol{w}) p(\boldsymbol{w})}{p(\boldsymbol{y} \mid X)} = \frac{\mathcal{N}(\boldsymbol{y}; \phi_X^\intercal \boldsymbol{w}, \sigma^2 I)\mathcal{N}(\boldsymbol{w}; \boldsymbol{\mu}, \Sigma)}{\mathcal{N}(\boldsymbol{y}; \phi_X^\intercal \boldsymbol{\mu}, \phi_X^\intercal \Sigma \phi_X + \sigma^2 I)} \\
&= \mathcal{N}(\boldsymbol{w}; \boldsymbol{\mu} + \Sigma\phi_X(\phi_X^\intercal \Sigma \phi_X + \sigma^2 I)^{-1}(\boldsymbol{y} - \phi_X \boldsymbol{\mu}), \Sigma - \Sigma\phi_X(\phi_X^\intercal \Sigma \phi_X + \sigma^2 I)^{-1}\phi_X \Sigma) \\
\mathbb{E}_{p(\boldsymbol{w}|\boldsymbol{y})}(\boldsymbol{w}) &= \arg\max_{\boldsymbol{w} \in \mathbb{R}^F} p(\boldsymbol{w} \mid \boldsymbol{y}) \\
&= \arg\min_{\boldsymbol{w}} -p(\boldsymbol{w} \mid \boldsymbol{y}) = \arg\min_{\boldsymbol{w}} -\log p(\boldsymbol{w} \mid \boldsymbol{y}) \\
&= \arg\min_{\boldsymbol{w}} \frac{1}{2\sigma^2}\|\boldsymbol{y} - \phi_X^\intercal \boldsymbol{w}\|^2 + \frac{1}{2}(\boldsymbol{w} - \boldsymbol{\mu})^\intercal \Sigma^{-1}(\boldsymbol{w} - \boldsymbol{\mu}) \\
&= (\Sigma^{-1} + \sigma^{-1}\phi_X \phi_X^\intercal)^{-1}(\Sigma^{-1}\boldsymbol{\mu} + \sigma^{-2}\phi_X \boldsymbol{y}) \quad \xrightarrow{\Sigma^{-1} \to 0} \quad (\phi_X \phi_X^\intercal)^{-1}\phi_X \boldsymbol{y}
\end{aligned}
$$

The **posterior mean** estimator of Gaussian regression is equal to the **regularized least-squares** estimate with the ($\boldsymbol{\mu}$-centered) regularizer $\|\boldsymbol{w} - \boldsymbol{\mu}\|_\Sigma^2$.

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

nonparametric formulation

more precise formulation to follow

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

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

200 years of data analysis

and counting

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

- 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

Kernels

Definition from last time

Definition (kernel)

$k : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ is a (Mercer / positive definite) **kernel** if, for any finite collection $X = [x_1, \ldots, x_N]$, the matrix $k_{XX} \in \mathbb{R}^{N \times N}$ with $[k_{XX}]_{ij} = k(x_i, x_j)$ is **positive semidefinite**.

```
def kernel (f) : λ (a,b) -> [[f(a[i],b[j]) for j=1:length(b)] for i=1:length(a) ]
```

actually, in python:

```
def kernel (f) : return lambda a,b : array([ [float64(f(a[i],b[j])) for j in range(b.size) ] for i in range(a.size) ])
```

Definition (positive definite matrix)

A symmetric matrix $A \in \mathbb{R}^{N \times N}$ is called **positive (semi-) definite** if $v^\intercal A v \geq 0 \; \forall \, v \in \mathbb{R}^N$. Equivalently:

- All eigenvalues of the symmetric matrix A are non-negative
- A is a Gram matrix — the outer product of N vectors $[\phi_i]_{i=1,\ldots,N}$

So, is a kernel "like an infinitely large symmetric positive definite matrix" ?

Gaussian processes

definition from last time

Definition

Let $\mu : \mathbb{X} \to \mathbb{R}$ be any function, $k : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ be a Mercer kernel. A **Gaussian process** $p(f) = \mathcal{GP}(f; \mu, k)$ is a probability distribution over the function $f : \mathbb{X} \to \mathbb{R}$, such that every finite restriction to function values $f_X := [f_{x_1}, \ldots, f_{x_N}]$ is a Gaussian distribution $p(f_X) = \mathcal{N}(f_X; \mu_X, k_{XX})$.

- So what kind of functions does a Gaussian process produce? What is the sample-space of a GP?
- Since Gaussian distributions have support over *all* real vectors, can a GP sample "every" function?
- Can GP regression **learn** "every" function?

Warning

The following are simplified expositions!
Some regularity assumptions have been dropped for easier readability.

For the full story of the relationship on GPs and kernel methods, 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
(in review)

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?

Yes, but with caveats

$$k(a,b) = \Phi(a)\,\mathrm{diag}(\boldsymbol{\lambda})\Phi(b)^\intercal =: \Phi(a)\Sigma\Phi(b)^\intercal$$

- 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
- Kernels are unwieldy objects. Properties and restrictions that are relatively straightforward for the finite-dimensional case (matrices / Gaussian inference) can be surprising and subtle in the infinite-dimensional case (kernels / Gaussian processes)

- What are the features? later today
- What is the space of functions representable by the eigenfunctions?
- Is it equal to the sample space of a GP?

Bochner's Theorem

sometimes, one can guess the eigenvalues after all

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:*

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

This insight has been used to build linear-time approximations to kernel ridge and Gaussian process regression (Rahimi & Recht, NIPS 2008)

- 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
- 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

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?

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

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

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 a 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 $\mathbb{X}$). A Borel measure on a topological space is a measure defined on all open sets. (A topological space $(\mathbb{X}, \tau)$ is a collection τ of subsets (called open sets*) of a set $\mathbb{X}$ closed under intersection and union).*

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 is the RKHS? (2)

The RKHS is the space of possible posterior mean functions

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

Corollary (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: set $\tilde{\alpha}_i = \alpha_i / (\lambda_i^{1/2} \phi(x_i))$ and use Mercer's theorem.

What is the RKHS? (2)

The RKHS is the space of possible posterior mean functions

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

Corollary (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:

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

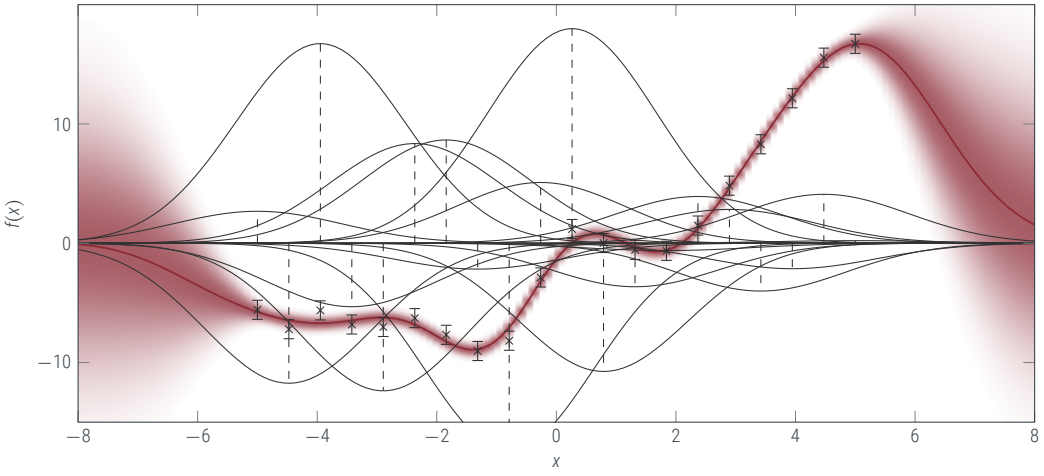
Proof: set $\tilde{\alpha}_i = \alpha_i / (\lambda_i^{1/2} \phi(x_i))$ and use Mercer's theorem.

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



- 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
- 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
- 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

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,

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

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.

- 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
- 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
- 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
  - GP samples are not in the RKHS, but "almost"