On the Connections and Equivalences between Gaussian Processes and Kernel Methods in Nonparametric Regression

Motonobu Kanagawa

Assistant Professor, EURECOM, Sophia Antipolis, France

Meeting in Mathematical Statistics, December 2020

Outline

Learning with Positive Definite Kernels

- In machine learning, positive definite kernels have been widely used in modeling nonlinear relationships between variables.

- There are two major approaches:

Bayesian learning with Gaussian Processes (GP)
[Rasmussen and Williams, 2006]:

- Positive definite kernels appear as covariance kernels
- e.g., Gaussian process regression, Bayesian optimization, probabilistic numerics, etc.

Learning with Positive Definite Kernels

Kernel methods using Reproducing Kernel Hilbert Spaces (RKHS) [Schölkopf and Smola, 2002]:

- Positive definite kernels appear as reproducing kernels.
- e.g., kernel ridge regression, support vector machines, kernel mean embedding, etc.

- These two approaches have been studied extensively in the literature (also in the context of neural networks).

The Aim of the Talk

- I will discuss the connections between the GP– and RKHS–based approaches, focusing on the problem of nonparametric regression.

- There is a well-known result for the equivalence between the two approaches to regression [Kimeldorf and Wahba, 1970].

- On the other hand, it is known that a sample path of the GP prior does not belong to the corresponding RKHS with probability 1 [Driscoll, 1973].

- This apparent difference in the "hypothesis spaces" may give an impression that the connection between the two approaches is rather superficial.

— How can we explain this difference? Does it contradict the equivalence?

The Aim of the Talk

- The aim of the talk is to clarify how these two approaches are related by studying the connections in their theoretical convergence properties.

- It will be shown that the above "apparent difference" does not lead to any contradiction, and there is a clear correspondence in the convergence results.

- The key is to understand that the RKHS approach has two ways of controlling the complexity of a model; the RKHS itself and regularization.

Contents and Collaborators

- The contents of the talk is based on Section 5 of

Kanagawa et al. (2018) Gaussian Processes and Kernel Methods: A Review on Connections and Equivalences, *arXiv preprint* (under revision)

- Collaborators:

- Philipp Hennig (University of Tuebingen and Max Planck Institue)
- Dino Sejdinovic (University of Oxford)
- Bharath K. Sriperumbudur (Penn State University)

Outline

Positive Definite Kernels

- Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a symmetric function on a set $\mathcal{X}$.
- The function $k(x, x')$ is called **positive definite kernel**, if

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j k(x_i, x_j) \geq 0 \quad \text{holds}$$

$$\text{for all} \quad n \in \mathbb{N}, \quad c_1, \ldots, c_n \in \mathbb{R}, \quad x_1, \ldots, x_n \in \mathcal{X}.$$

- Examples of positive definite kernels on $\mathcal{X} \subset \mathbb{R}^d$:

$$\begin{aligned}
\text{Gaussian} \quad & k(x, x') &&= \exp(-\|x - x'\|^2/\gamma^2). \\
\text{Laplace} \quad & k(x, x') &&= \exp(-\|x - x'\|/\gamma). \\
\text{Linear} \quad & k(x, x') &&= \langle x, x' \rangle. \\
\text{Polynomial} \quad & k(x, x') &&= (\langle x, x' \rangle + c)^m.
\end{aligned}$$

- In this talk, I will simply say kernels to indicate positive definite kernels.

Matérn Kernels

- In this talk, Matérn kernels will play the key role (Let $\mathcal{X} \subset \mathbb{R}^d$).

Example (Matern Kernels)

- For constants $\alpha > 0$ and $h > 0$, the Matérn kernel is defined by

$$k_{\alpha,h}(x, x') = \frac{1}{2^{\alpha-1}\Gamma(\alpha)} \left(\frac{\sqrt{2\alpha}\|x - x'\|}{h} \right)^{\alpha} K_\alpha \left(\frac{\sqrt{2\alpha}\|x - x'\|}{h} \right)$$

where Γ is the Gamma function, and K_α is the modified Bessel function of the second kind of order α.

Matérn Kernels

- If α is a half integer, an analytic form is known: e.g.,

$$k_{1/2,h}(x,x') = \exp\left(-\frac{\|x-x'\|}{h}\right) \quad \text{(Laplace kernel)},$$

$$k_{3/2,h}(x,x') = \left(1+\frac{\sqrt{3}\|x-x'\|}{h}\right)\exp\left(-\frac{\sqrt{3}\|x-x'\|}{h}\right),$$

$$k_{5/2,h}(x,x') = \left(1+\frac{\sqrt{5}\|x-x'\|}{h}+\frac{5\|x-x'\|^2}{3h^2}\right)\exp\left(-\frac{\sqrt{5}\|x-x'\|}{h}\right).$$

- The Gaussian kernel is given by the limit $\alpha \to \infty$ of the Matérn kernel.

$$\lim_{\alpha\to\infty} k_{\alpha,h}(x,x') = \exp\left(-\frac{\|x-x'\|^2}{2h^2}\right), \quad x,x' \in \mathbb{R}^d.$$

- One of the most standard kernels in the literature (e.g., geostatistics, Bayesian optimization)

Gaussian Processes (GP)

- For a given kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and a function $m : \mathcal{X} \to \mathbb{R}$, we can consider a corresponding Gaussian process (GP), which we denote by

$$g \sim \mathcal{GP}(m, k).$$

- The g is a random function on $\mathcal{X}$ satisfying the following:

  - Take any finite set $X := (x_1, \ldots, x_n) \subset \mathcal{X}$ of any size $n \in \mathbb{N}$,
  - Then the corresponding random vector induced by g

$$g_X := (g(x_1), \ldots, g(x_n))^\intercal \in \mathbb{R}^n$$

  follows the Gaussian distribution $\mathcal{N}(m_X, k_{XX})$ with
  — mean vector $m_X = (m(x_1), \ldots, m(x_n))^\intercal \in \mathbb{R}^n$
  — covariance matrix $k_{XX} = (k(x_i, x_j))_{i,j=1}^n \in \mathbb{R}^{n \times n}$

- $g \sim \mathcal{GP}(m, k)$ is a GP with mean function m and covariance kernel k.

Reproducing Kernel Hilbert Spaces (RKHS)

- For a given kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, there exists a uniquely associated Hilbert space $\mathcal{H}$ consisting of functions on $\mathcal{X}$ such that

$$\text{(i)} \quad k(\cdot, x) \in \mathcal{H} \quad \text{for all } x \in \mathcal{X}$$

where $k(\cdot, x)$ is the function of the first argument with x fixed:

$$x' \in \mathcal{X} \to k(x', x).$$

$$\text{(ii)} \quad f(x) = \langle f, k(\cdot, x) \rangle_{\mathcal{H}} \quad \text{for all } f \in \mathcal{H} \text{ and } x \in \mathcal{X},$$

which is called the **reproducing property**.

– $\mathcal{H}$ is called the RKHS of k.

– $\mathcal{H}$ can be written as

$$\mathcal{H} = \overline{\text{span}\,\{k(\cdot, x) \mid x \in \mathcal{X}\}}$$

Outline

Regression Problem

- Let $(x, y) \in \mathcal{X} \times \mathbb{R}$ be population random variables ($\mathcal{X} \subset \mathbb{R}^d$), such that

$$y = f_0(x) + \varepsilon,$$

where

— $f_0 : \mathcal{X} \to \mathbb{R}$ is the regression function (the "unknown true function");

— ε is an independent, zero-mean noise variable.

- Assume that n i.i.d. observations of (x, y)

$$(x_1, y_1), \ldots, (x_n, y_n)$$

are given as training data: let $D_n := (x_i, y_i)_{i=1}^n$.

- The task of regression is to estimate the unknown true function f_0 using D_n.

Gaussian Process Regression (GPR)

- For the regression problem, GPR takes the Bayesian approach.

- For the true unknown function f_0, we first define a prior distribution as a Gaussian process:

$$f_0 \sim \mathcal{GP}(0, k). \tag{1}$$

(For simplicity, we assume the zero-mean function, $m(x) = 0$).

- Given training data $D_n = (x_i, y_i)_{i=1}^n$, we also define a likelihood function

$$p(D_n|f_0) := \prod_{i=1}^{n} p_{\text{gauss}}(y_i; f_0(x_i), \sigma^2)$$

assuming the i.i.d. Gaussian noise model

$$y_i = f_0(x_i) + \varepsilon_i, \quad \text{where} \quad \varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2), \quad i = 1, \ldots, n.$$

- Here, $\sigma^2 > 0$ is the noise variance.

Gaussian Process Regression (GPR)

- Then Bayes' rule is applied to obtain the posterior distribution of the true function f_0:

$$P(f_0|D_n) \propto p(D_n|f_0)P(f_0),$$

where $P(f_0)$ indicates the GP prior

$$f_0 \sim \mathcal{GP}(0, k).$$

Gaussian Process Regression (GPR)

- Importantly, the posterior $P(f_0|D_n)$ is also given as a Gaussian process:

- Let $X := (x_1, \ldots, x_n) \in \mathcal{X}^n$ and $Y := (y_1, \ldots, y_n)^\intercal \in \mathbb{R}^n$. Then

$$P(f_0|D_n) = \mathcal{GP}(\bar{m}, \bar{k}),$$

where $\bar{m} : \mathcal{X} \to \mathbb{R}$ and $\bar{k} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ are given by

$$\begin{aligned} \bar{m}(x) &= k_X(x)^\top (k_{XX} + \sigma^2 I_n)^{-1} Y, \quad x \in \mathcal{X}, \\ \bar{k}(x, x') &= k(x, x') - k_X(x)^\top (k_{XX} + \sigma^2 I_n)^{-1} k_X(x'), \quad x, x' \in \mathcal{X}, \end{aligned}$$

where $I_n \in \mathbb{R}^{n \times n}$ is the identity matrix, and

$$k_X(x) := (k(x_1, x), \ldots, k(x_n, x))^\top \in \mathbb{R}^n, \quad k_{XX} := (k(x_i, x_j))_{i,j=1}^n \in \mathbb{R}^{n \times n}.$$

— $\bar{m} : \mathcal{X} \to \mathbb{R}$ is called the posterior mean function;

— $\bar{k} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called the posterior covariance function.

Kernel Ridge Regression (KRR)

- On the other hand, KRR deals with regression by solving the following optimization problem in the RKHS $\mathcal{H}_k$:

$$\hat{f}_\lambda := \arg\min_{f \in \mathcal{H}_k} \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2 + \lambda \|f\|^2_{\mathcal{H}_k},$$

where $\lambda > 0$ is a regularization constant.

- This is a regularized least-squares problem.

— The first term quantifies the fit to the training data $D_n = (x_i, y_i)_{i=1}^n$:

— The second term controls the complexity of the estimate $\hat{f}_\lambda$:

i.e., a larger λ makes $\hat{f}_\lambda$ smoother.

Kernel Ridge Regression (KRR)

- By the representor theorem, the solution $\hat{f}_\lambda$ is given in a closed form.

- Let $X := (x_1, \ldots, x_n) \in \mathcal{X}^n$ and $Y := (y_1, \ldots, y_n)^\intercal \in \mathbb{R}^n$. Then

$$\hat{f}_\lambda(x) = k_X(x)^\top (k_{XX} + n\lambda I_n)^{-1} Y, \quad x \in \mathcal{X},$$

where $I_n \in \mathbb{R}^{n \times n}$ is the identity matrix, and

$$k_X(x) := (k(x_1, x), \ldots, k(x_n, x))^\top \in \mathbb{R}^n, \quad k_{XX} := (k(x_i, x_j))_{i,j=1}^n \in \mathbb{R}^{n \times n}.$$

Outline

Formal Equivalence between GPR and KRR

- Let's compare the posterior mean function of GPR

$$\bar{m}(x) = k_X(x)^\top (k_{XX} + \sigma^2 I_n)^{-1} Y, \quad x \in \mathcal{X},$$

and the estimate of KRR

$$\hat{f}_\lambda(x) = k_X(x)^\top (k_{XX} + n\lambda I_n)^{-1} Y, \quad x \in \mathcal{X},$$

- From this, it immediately follows that the GPR and KRR estimates are equal

$$\bar{m}_n = \hat{f}_\lambda \quad \text{provided that} \quad \sigma^2 = n\lambda.$$

- This is a well known result in the literature [Kimeldorf and Wahba, 1970].

Sample Paths of a Gaussian Process

- On the other hand, it is known that a sample path $g \sim \mathcal{GP}(0, k)$ does not lie in the corresponding RKHS $\mathcal{H}_k$ (when $\mathcal{H}_k$ is infinte dimensional) [Driscoll, 1973]:

$$g \notin \mathcal{H}_k \quad \text{with probability} \quad 1.$$

- Intuitively, the sample path $g \sim \mathcal{GP}(0, k)$ becomes "rougher" than the functions in $\mathcal{H}_k$.

- For instance, assume that the kernel k is a Matern kernel of order $\alpha > 0$.

  - Then $\mathcal{H}_k$ consists of functions having smoothness $s = \alpha + d/2$ (in the sense of Sobolev; $\approx s$-times weakly differentiable), while
  - Sample path $g \sim \mathcal{GP}(0, k)$ has the smoothness α with probability 1.

— i.e., the RKHS $\mathcal{H}_k$ is $d/2$-smoother than the sample path $g \sim \mathcal{GP}(0, k)$.

Apparent Difference in Hypothesis Spaces

- Recall that GPR uses $\mathcal{GP}(0, k)$ to define a prior distribution (or as a hypothesis space) of the unknown true function f_0:

$$f_0 \sim \mathcal{GP}(0, k).$$

- Given that this prior is correct, we have

$$f_0 \notin \mathcal{H}_k \quad \text{with probability} \quad 1.$$

- Does this contradict the fact that KRR uses $\mathcal{H}_k$ as a "hypothesis space"?

$$\hat{f}_\lambda := \arg\min_{f \in \mathcal{H}_k} \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2 + \lambda \|f\|_{\mathcal{H}_k}^2,$$

Apparent Difference in Hypothesis Spaces

- The above fact may give an impression that the formal equivalence between GPR and KRR is rather superficial.

- However, we argue below that the apparent difference between the GP and RKHS as "hypothesis spaces" does not lead to any contradiction.

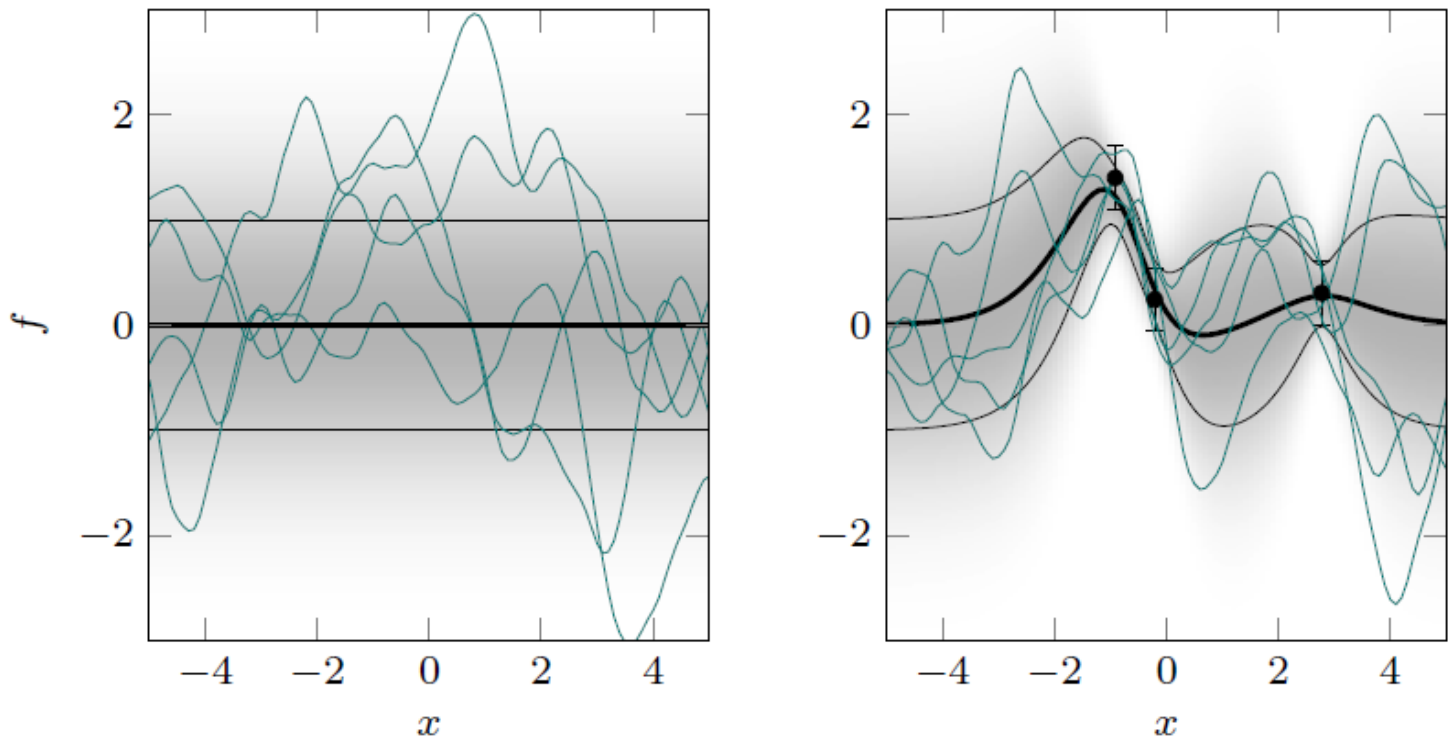
- We do this by studying the correspondence between the convergence properties of GPR and KRR.

- The key point is that KRR does not require $f_0 \in \mathcal{H}_k$ to achieve minimax optimal rates.

Illustration of GPR with a Matérn kernel order $\alpha = 5/2$

- Left: sample paths from the prior $\mathcal{GP}(0, k)$; the thick line is the prior mean function $= 0$.

- Right: sample paths from the posterior $\mathcal{GP}(\bar{m}_n, \bar{k}_n)$; the thick curve is the posterior mean function $\bar{m}_n \in \mathcal{H}_k$. The three points are training data.



Outline

Common Setting

- We assume that the true unknown function f_0 has smoothness $\beta > 0$ (in the sense of Sobolev; $\approx \beta$-times weakly differentiable).

- For simplicity we consider $\mathcal{X} := [0,1]^d \subset \mathbb{R}^d$ as an input space.

- The quality of the estimate $\hat{f}$ (either $\bar{m}_n$ in GPR or $\hat{f}_\lambda$ in KRR) is measured with the L_2 distance with respect an input distribution $P_{\mathcal{X}}$:

$$\|\hat{f} - f_0\|^2_{L_2(P_{\mathcal{X}})} = \int \left(\hat{f}(x) - f(x)\right)^2 dP_{\mathcal{X}}(x)$$

Convergence Rates for Gaussian Process Regression

- We first present a convergence result for GPR that follows from [van der Vaart and van Zanten, 2011].

- Assume that (x, y) are population random variables generated as

  - $x \sim P_{\mathcal{X}}$ for an input distribution $P_{\mathcal{X}}$ on $\mathcal{X} = [0,1]^d$.
  - $y = f_0(x) + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ is independent.

- Note that f_0 is the true unknown function.

- Assume that the data $D_n := (x_i, y_i)_{i=1}^n$ are i.i.d. with (x, y).

(Note: the Gaussian noise assumption may be relaxed [Kleijn and van der Vaart, 2006])

Convergence Rates for Gaussian Process Regression

- Assume that we perform GPR with a Matérn kernel $k_{\alpha,h}$ of order $\alpha > 0$.

— This means that a sample path from the GP prior has smoothness α.

- Assume that the true function f_0 has smoothness $\beta > 0$ and that $\min(\alpha, \beta) > d/2$.

- Then, the posterior mean $\bar{m}_n$ of GPR with data $D_n = (x_i, y_i)_{i=1}^n$ satisfies

$$\|\bar{m}_n - f_0\|_{L_2(P_X)}^2 = O_p(n^{-2\min(\alpha,\beta)/(2\alpha+d)}) \quad (n \to \infty)$$

under additional minor technical contidions.

Convergence Rates for Gaussian Process Regression

- Let's study the exponent of the rate $n^{-2\min(\alpha,\beta)/(2\alpha+d)}$:

$$\frac{2\min(\alpha,\beta)}{2\alpha+d}$$

- This becomes the largest when

$$\alpha = \beta$$

i.e., when the smoothness α of the GP prior $\mathcal{GP}(0, k_{\alpha,h})$ is equal to the smoothness β of the true function f_0.

- In this case, the rate becomes minimax optimal [Stone, 1980]:

$$\|\bar{m}_n - f_0\|^2_{L_2(P_{\mathcal{X}})} = O_p(n^{-2\beta/(2\beta+d)}) \quad (n \to \infty).$$

- This result indicates that, for GPR to perform well, the GP prior should have the same smoothness as the target function f_0

Convergence Rates for Kernel Ridge Regression

- Next, let's look at a convergence result for KRR that follows from [Steinwart et al., 2009].

- Assume that input-output variables (x, y) are given as

  - $x \sim P_{\mathcal{X}}$ for a probability distribution $P_{\mathcal{X}}$ on $\mathcal{X} = [0, 1]^d$.
  - $y = f_0(x) + \varepsilon$, where ε is an independent noise such that $\varepsilon \in [-M, M]$ for some constant $M > 0$ almost surely.

(Note: the latter boundedness assumption can be relaxed [Fischer and Steinwart, 2020])

Convergence Rates for Kernel Ridge Regression

- Assume that we perform KRR using a Matérn kernel $k_{\alpha,h}$ of order $\alpha > 0$.

— This means that the RKHS $\mathcal{H}_{k_{\alpha,h}}$ has the smoothness $s = \alpha + d/2$.

- Set the regularization constant $\lambda := \lambda_n$ as

$$\lambda_n = cn^{-2s/(2\beta+d)}$$

where $c > 0$ is an arbitrary constant independent of n.

- Assume that the true function f_0 has smoothness β and that $s \geq \beta$.

- Then for (a truncated version) of the KRR estimator $\hat{f}_{\lambda_n}$, we have

$$\|\hat{f}_{\lambda_n} - f_0\|^2_{L_2(P_{\mathcal{X}})} = O_p(n^{-2\beta/(2\beta+d)}) \quad (n \to \infty).$$

under additional minor technical conditions.

Remarks on the Rates for Kernel Ridge Regression

- The rate $n^{-2\beta/(2\beta+d)}$ is minimax optimal in nonparametric regression of a function f_0 on $\mathcal{X} \subset \mathbb{R}^d$ with smoothness β [Stone, 1980].

- The condition $s = \alpha + d/2 \geq \beta$ means that smoothness β of the true function f_0 can be **smaller** than the smoothness s of the RKHS $\mathcal{H}_{k_{\alpha,h}}$.

— i.e., f_0 can be "rougher" than the functions in $\mathcal{H}_{k_{\alpha,h}}$.

— thus, the misspecified setting $f_0 \notin \mathcal{H}_k$ is permitted.

Remarks on the Rates for Kernel Ridge Regression

- Even in the misspecified setting $f_0 \notin \mathcal{H}_k$, KRR can achieve the optimal convergence rate $n^{-2\beta/(2\beta+d)}$.

- This is enabled by an appropriately chosen regularization constant:

$$\lambda_n = cn^{-2s/(2\beta+d)}$$

which should be smaller if β is much smaller than $s = \alpha + d/2$.

i.e., if the true function f_0 is rougher than the RKHS $\mathcal{H}_{k_{\alpha,h}}$, regularization should be weaker.

Correspondence between the Rates for GPR and KRR

- Let's study how the results for GPR and KRR are related.

- Recall that the optimal rate in GPR is achieved by setting $\alpha = \beta$, where

  - β is the smoothness of the true function f_0.
  - α is the smoothness of the GP prior $\mathcal{GP}(0, k_{\alpha,h})$.

- For KRR, this corresponds to setting the smoothness $s = \alpha + d/2$ of the RKHS as

$$s = \beta + d/2 \quad \text{since} \quad \alpha = \beta$$

and the regularization constant λ_n as

$$\lambda_n = cn^{-2(\beta+d/2)/(2\beta+d)} = cn^{-1}$$

which results in the optimal rate $n^{-2\beta/(2\beta+d)}$.

Correspondence between the Rates for GPR and KRR

- In particular, by setting $c := \sigma^2$ (= the noise variance in GPR), the regularization schedule becomes

$$\lambda_n = \sigma^2 n^{-1}.$$

- Recall that this is exactly the condition

$$\sigma^2 = n\lambda_n$$

for the equivalence between GPR and KRR:

$$\bar{m}_n = \hat{f}_{\lambda_n}.$$

Summary of the Discussion

- We started the discussion from the fact that

if the GP prior $f_0 \sim \mathcal{GP}(0, k)$ is correct, we have $f_0 \notin \mathcal{H}_k$ with probability 1.

- This should not be regarded as a contradiction, since KRR allows for the misspecified setting $f_0 \notin \mathcal{H}_k$.

— This is because KRR controls the complexity of the regressor $\hat{f}_{\lambda_n}$ by regularization;

— and the optimal regularization constant λ_n when the GP prior $f_0 \sim \mathcal{GP}(0, k)$ is correct can be given by

$$\lambda_n = \sigma^2 / n$$

which implies the equivalence of GPR and KRR.

Practical Implication of the Correspondence

- Suppose that the kernel $k_{\alpha,h}$ is well-specified (in the sense that the GP sample path has the correct smoothness, $\alpha = \beta$).

- Then, the interpretation of KRR as GPR suggests the following:

- If we set the regularization constant $\lambda_n = \sigma^2/n$ very small, this implies either that

  - we know/believe that the noise variance σ^2 is very small, or that
  - the sample size n is very large.

- If neither of the above points is correct, then we shouldn't set the regularization constant λ_n too small.

Outline

Conclusions

- In this talk, I discussed the correspondence between the GPR and KRR.

- The aim was to resolve the apparent contradiction in their way of defining hypothesis spaces.

- The contradiction is shown to be superficial: the formal equivalence between GPR and KRR is essential, carrying over into their optimal convergence results.

Further Topics

- There are several other connections between the GP and RKHS approaches:

  - RKHS interpretation of GP posterior variance.
  - Powers of an RKHS as GP sample spaces
  - GP interpretations of kernel mean embeddings (e.g., MMD, HSIC)

If you are interested, please look at [Kanagawa et al., 2018].

Contents and Collaborators

- The contents of the talk is based on Section 5 of

Kanagawa et al. (2018) Gaussian Processes and Kernel Methods: A Review on Connections and Equivalences, *arXiv preprint* (under revision)

- Collaborators:

- Philipp Hennig (University of Tuebingen and Max Planck Institue)
- Dino Sejdinovic (University of Oxford)
- Bharath K. Sriperumbudur (Penn State University)


Driscoll, M. F. (1973).
The reproducing kernel Hilbert space structure of the sample paths of a gaussian process.
Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete, 26(4):309–316.


Fischer, S. and Steinwart, I. (2020).
Sobolev norm learning rates for regularized least-squares algorithms.
Journal of Machine Learning Research, 21(205):1–38.


Kanagawa, M., Hennig, P., Sejdinovic, D., and Sriperumbudur, B. K. (2018).
Gaussian processes and kernel methods: A review on connections and equivalences.
arXiv preprint arXiv:1807.02582.


Kimeldorf, G. S. and Wahba, G. (1970).
A correspondence between Bayesian estimation on stochastic processes and smoothing by splines.

The Annals of Mathematical Statistics, 41(2):495–502.


Kleijn, B. J. K. and van der Vaart, A. W. (2006).
Misspecification in infinite-dimensional Bayesian statistics.
Annals of Statistics, 34(2):837–877.


Rasmussen, C. and Williams, C. (2006).
Gaussian Processes for Machine Learning.
MIT Press.


Schölkopf, B. and Smola, A. J. (2002).
Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.
MIT press.


Steinwart, I., Hush, D., and Scovel, C. (2009).
Optimal rates for regularized least squares regression.
In Dasgupta, S. and Klivans, A., editors, *Proceedings of the 22nd Annual Conference on Learning Theory*, pages 79–93.


Stone, C. J. (1980).

Optimal rates of convergence for nonparametric estimators.
The Annals of Statistics, 8(6):1348–1360.

van der Vaart, A. and van Zanten, H. (2011).
Information rates of nonparametric Gaussian process methods.
Journal of Machine Learning Research, 12:2095–2119.