工学院大学研究報告第130号　令和4年3月

グリーン関数理論に基づく共分散行列のガウス過程回帰への応用

永井朋子*

The covariance matrix of Green functions and its application to Gaussian Process Regression

Tomoko NAGAI*

Abstract

2 階常微分方程式の境界値問題のグリーン関数を規格化し，ガウス過程回帰の新しいカーネル関数として提案した．ベイズ推定に基づき，観測データに対する条件付き分布として予測分布を求めた．カスプのある場合には，グリーン関数カーネルの対数尤度は広く利用されるガウスカーネルのそれより，大きなあるいは同程度の値を示し，よい回帰結果が得られた．

Keywords: グリーン関数，共分散行列，ガウス過程回帰

1. Introduction

It was shown in 2002 that there is a close relationship between machine learning and reproducing kernel theory[1]. A covariance matrix in Bayesian regression, which is composed of kernel functions, is presented as a kernel matrix in regression[1)~6)]. The Gaussian process regression and interpolation based on the Bayesian approach give the predictive distribution[1)~6)].

It is interesting to note that at almost the same time in 2005 Green functions to boundary value problem for differential equations, which are response functions for impulses, were shown to be reproducing kernels of suitable Hilbert spaces[7), 8)].

The above two results suggest the relationship between Green functions and kernel functions in machine learning.

The purpose of this paper is to clarify roles of Green functions as a Gaussian process regression algorithm. In particular, a covariance matrix composed of normalized Green functions is proposed. By applying Bayesian approach, the covariance matrix provides a predictive distribution.

2. Application of Green function to Gaussian process regression

We first survey the result of Kametaka[7), 8)] in which Green function for a simple boundary value problem is a reproducing kernel for a suitable Hilbert space. We here adopt the normalized Green function as a kernel function of Gaussian process regression and propose a new regression algorithm based on Green function theory.

2.1 Green function

We start with the following boundary value problem of 2nd order linear ordinary differential equation:

$$\begin{cases} -\frac{d^2u}{dx^2} + a^2u = w(x) \ \ (0 < x < 1) \\ \quad u(0) = u(1) = 0 \end{cases} \tag{1}$$

where a is a nonnegative constant and $w(x)$ is an external force term. This equation stands for a bending of a string supported by a uniformly distributed spring with spring constant a^2. The above problem is the simplest but also the most important example of boundary value problems[9]. The solution formula of Eq. (1) is given by

$$u(x) = \int_0^1 G(x,y)\,w(y)dy, \tag{2}$$

where $G(x,y)$ is a Green function defined by

$$G(x,y) = G(a;x,y) =$$

[*] 工学院大学学習支援センター講師

$$\begin{cases} \frac{\sinh\left(a \min(x,y)\right) \sinh\left(a\,(1-\max(x,y))\right)}{a \sinh a} & (a > 0) \\ \min(x,y)\left(1-\max(x,y)\right) & (a = 0) \end{cases}. \qquad (3)$$

Let H be a function space defined by

$$\mathrm{H}= \{u | u, u' \in L^2(0,1), u(0) = u(1) = 0\}, \qquad (4)$$

equipped with an inner product

$$(u,v)_{\mathrm{H}} = \int_0^1 (u'(x)v'(x) + a^2 u(x)v(x))dx. \qquad (5)$$

It should be noted that $(\mathrm{H},(\cdot,\cdot)_{\mathrm{H}})$ is a Hilbert space. Kametaka *et al.* showed that $G(x,y)$ is a reproducing kernel of H[7), 8)]. In other words, the following two properties hold:

(i) If one fixes $y \in [0,1]$, $G(x,y)$, as a function of x, belongs to H.

(ii) For all $u \in$H, the following reproducing relation holds:

$$(u, G(\cdot. y))_{\mathrm{H}}$$
$$= \int_0^1 \left(u'(x)\partial_x G(x,y) + a^2 u(x) G(x,y)\right)dx = u(y). \qquad (6)$$

We consider the case $a > 0$. Since $G(x,y)$ is non-negative, L^1 norm of a cross section Green function $G(x,y)$ is calculated as follows:

$$L_1 = L_1(y) = \int_0^1 |G(x,y)|\, dx = \int_0^1 G(x,y)\, dx$$
$$= \frac{1}{a^2}\left(1 - \cosh(ay) + \frac{(\cosh a - 1)\sinh(ay)}{\sinh a}\right). \qquad (7)$$

We also define $\tilde{G}(x,y)$ as Green function divided by its L^1 norm

$$\tilde{G}(x,y) = \frac{G(x,y)}{L_1(y)} \quad (0 < y < 1), \qquad (8)$$

which satisfies the relation

$$\int_0^1 \tilde{G}(x,y)\, dx = 1. \qquad (9)$$

We call the function $\tilde{G}(x,y)$ the normalized Green function hereafter. Example of $\tilde{G}(x,y)$ is shown in Fig. 1, which means the normalized response function by the impulse at point y. Note that $\tilde{G}(0,y) = \tilde{G}(1,y) = 0$ holds in accordance with the boundary condition.

Together with the relation between machine learning and reproducing kernel theory[1)], we can expect an application of Green function theory to Gaussian process regression.

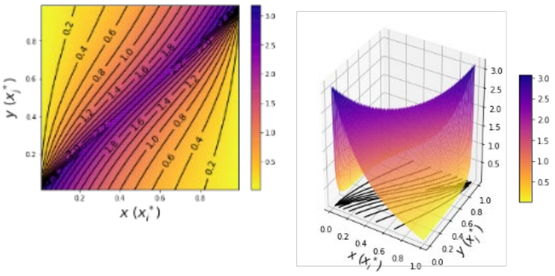

Figure 1 The plot of normalized Green function $\tilde{G}(x,y) = \tilde{G}(a; x,y)$ with $a = 2.88$. (left) 2-dimensional contour, (right) 3-dimensional plot.

2.2 Gaussian process

In this subsection we review Gaussian process[1) ~6)]. For a given set of input and output data $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_N, y_N)$, Gaussian process regression is one of the algorithms to infer a function $y = f(\boldsymbol{x})$. Generally input variable $\boldsymbol{x}$ is a multi-dimensional vector, however we deal with a one-dimensional real variable x hereafter.

In general, Gaussian process is defined as a probability distribution over function $\boldsymbol{f}(x)$ such that the set of values of $f(x_1), f(x_2), \cdots, f(x_N)$ evaluated at an arbitrary set of points $x_1,\ x_2, \cdots,\ x_N$ has a Gaussian distribution. A key point of Gaussian process is that the joint distribution of N function values $\boldsymbol{f}(X_N) = \left(f(x_1), f(x_2), \cdots, f(x_N)\right)^T$ for input variable $X_N = \{x_1,\ x_2, \cdots,\ x_N\}$ is specified by $\boldsymbol{f}(X_N) \sim N\left(\boldsymbol{\mu}(X_N), V(X_N, X_N)\right)$, where $\boldsymbol{\mu}(X_N) = (\mu(x_1), \mu(x_2), \cdots, \mu(x_N))^T$ is an N dimensional mean vector and $V(X_N, X_N)$ is an $N \times N$ covariance matrix. Each element of the covariance matrix $v(x, x')$ is equal to a kernel function[10), 11)]. We can select a variety of kernel functions. The choice of kernel function and mean $\mu(x) = \mu_0(x)$ determines the prior distribution over the function f, before observing any data. In the absence of prior knowledge, the mean is defined by $\mu_0(x) = 0$ in general.

In order to apply Gaussian process to the problem of regression, Gaussian noise is taken into account on the observed value $\boldsymbol{y} = (y_1, y_2, \cdots, y_N)^T$, which are given by $y_i = f(x_i) + \varepsilon\ (i = 1, 2, \cdots, N)$, where ε is a random noise variable whose value is chosen independently for each observation. It is assumed that the noise processes have a Gaussian distribution as, $\varepsilon \sim N(0, \rho)$.

2.3 Gaussian process regression with the Green function kernel

In this subsection we apply the normalized Green function to Gaussian process regression[1)~6)] based upon Bayesian inference. Given the data set D={$(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)$} of N observations, where $X_N=\{x_1,x_2,\cdots,x_N\}$ is an input variable and $\boldsymbol{y}=(y_1,y_2,\cdots,y_N)^T$ is an observed value, we consider the following problem:

Problem: Using Green function kernel, predict M dimensional output variable $\boldsymbol{y}^*=(y_1^*,y_2^*,\cdots,y_M^*)^T$ for a new input variable $X_M^*=\{x_1^*,x_2^*,\cdots,x_M^*\}$.

We assume the stochastic process governed by Gaussian process. First, we consider the prior distribution of f. We apply the normalized Green function as a kernel function $v(x,x')=\tilde{G}(x,x')$ and set mean $\mu(x)=\mu_0(x)$, which specify the prior distribution of Gaussian process. Next we consider the posterior distribution. We apply the normalized Green function to regression. We also assume the Gaussian noise ε on the observed N values.

In order to obtain the posterior distribution of function value $\boldsymbol{f}(X_M^*)=\left(f(x_1^*),f(x_2^*),\cdots,f(x_M^*)\right)^T$ for new input variable X_M^* as the conditional distribution $p(\boldsymbol{f}(X_M^*)|\boldsymbol{y})$, we consider the joint prior distribution of $\boldsymbol{y}$ and $\boldsymbol{f}(X_M^*)$. By adopting the Green functions as kernel functions of a covariance matrix, first the distribution of function values $\boldsymbol{f}(X_N)=(f(x_1),f(x_2),\cdots,f(x_N))^T$ is given as

$$\boldsymbol{f}(X_N)\sim N(\boldsymbol{\mu}_0(X_N),H(X_N,X_N)) \tag{10}$$

$$\boldsymbol{\mu}_0(X_N)=(\mu_0(x_1),\mu_0(x_2),\cdots,\mu_0(x_N))^T \tag{11}$$

$$H(X_N,X_N)=\begin{pmatrix}\tilde{G}(x_1,x_1) & \cdots & \tilde{G}(x_1,x_N)\\ \vdots & \ddots & \vdots\\ \tilde{G}(x_N,x_1) & \cdots & \tilde{G}(x_N,x_N)\end{pmatrix}, \tag{12}$$

where $\boldsymbol{\mu}_0(X_N)$ is a mean vector, and $H(X_N,X_N)$ is an $N\times N$ covariance matrix whose (i,j)-th element is the normalized Green function $\tilde{G}(x_i,x_j)$. Adding Gaussian noise process $\varepsilon\sim N(0,\rho)$ to $\boldsymbol{f}(X_N)$, as $y_{\boldsymbol{i}}=f(x_i)+\varepsilon\ (i=1,2,\cdots,N)$, we obtain the distribution of $\boldsymbol{y}$ governed by

$$\boldsymbol{y}\sim N(\boldsymbol{\mu}_0(X_N),H(X_N,X_N)+\rho I_N), \tag{13}$$

$$(H(X_N,X_N)+\rho I_N)_{i,j}=\tilde{G}\left(x_i,x_j\right)+\rho\delta_{i,j}$$
$$(i,j=1,2,\cdots,N) \tag{14}$$

where I_N denotes the $N\times N$ unit matrix and $\delta_{i,j}$ is the Kronecker delta.

Next the function value $\boldsymbol{f}(X_M^*)$ for a new input variable X_M^* is governed by a Gaussian distribution:

$$\boldsymbol{f}(X_M^*)\sim N(\boldsymbol{\mu}_0(X_M^*),H(X_M^*,X_M^*)) \tag{15}$$

$$\boldsymbol{\mu}_0(X_M^*)=(\mu_0(x_1^*),\mu_0(x_2^*),\cdots,\mu_0(x_M^*))^T \tag{16}$$

$$H(X_M^*,X_M^*)=\begin{pmatrix}\tilde{G}(x_1^*,x_1^*) & \cdots & \tilde{G}(x_1^*,x_M^*)\\ \vdots & \ddots & \vdots\\ \tilde{G}(x_M^*,x_1^*) & \cdots & \tilde{G}(x_M^*,x_M^*)\end{pmatrix}. \tag{17}$$

A covariance matrix between $\boldsymbol{f}(X_N)$ and $\boldsymbol{f}(X_M^*)$ is given by $H(X_N,X_M^*)$, whose (i,j)-th entry is $\tilde{G}\left(x_i,x_j^*\right)\ (i=1,2,\cdots,N,j=1,2,\cdots,M)$. Therefore, we obtain the joint prior distribution of $\boldsymbol{y}$ and $\boldsymbol{f}(X_M^*)$ given as follows:

$$\begin{pmatrix}\boldsymbol{y}\\ \boldsymbol{f}_M\end{pmatrix}\sim N\left(\begin{pmatrix}\boldsymbol{\mu}_{0,N}\\ \boldsymbol{\mu}_{0,M}\end{pmatrix},\begin{pmatrix}H_{N,N}+\rho I_N & H_{N,M}\\ H_{M,N} & H_{M,M}\end{pmatrix}\right), \tag{18}$$

$$\boldsymbol{f}_M=\boldsymbol{f}(X_M^*), \tag{19}$$

$$\boldsymbol{\mu}_{0,N}=\boldsymbol{\mu}_0(X_N), \tag{20}$$

$$\boldsymbol{\mu}_{0,M}=\boldsymbol{\mu}_0(X_M^*), \tag{21}$$

$$H_{N,N}=H(X_N,X_N), \tag{22}$$

$$H_{M,M}=H(X_M^*,X_M^*), \tag{23}$$

$$H_{N,M}=H(X_N,X_M^*)=\begin{pmatrix}\tilde{G}(x_1,x_1^*) & \cdots & \tilde{G}(x_1,x_M^*)\\ \vdots & \ddots & \vdots\\ \tilde{G}(x_N,x_1^*) & \cdots & \tilde{G}(x_N,x_M^*)\end{pmatrix} \tag{24}$$

$$H_{M,N}=H(X_M^*,X_N)=\begin{pmatrix}\tilde{G}(x_1^*,x_1) & \cdots & \tilde{G}(x_1^*,x_N)\\ \vdots & \ddots & \vdots\\ \tilde{G}(x_M^*,x_1) & \cdots & \tilde{G}(x_M^*,x_N)\end{pmatrix}, \tag{25}$$

where Eqs. (19)-(25) is introduced for simplicity of notation. $H_{N,M}$ and $H_{M,N}$ are $N\times M$ and $M\times$N covariance matrices composed of the normalized Green function kernel, respectively.

Applying the results in terms of the partitioned covariance matrix[2)], we obtain the conditional distribution $p(\boldsymbol{f}_M|\boldsymbol{y})$, if $\mu(x)=0$ holds, given as

$$p(\boldsymbol{f}_M|\boldsymbol{y})=N\left(H_{M,N}\left(H_{N,N}+\rho I_N\right)^{-1}\boldsymbol{y},\ H_{M,M}-H_{M,N}\left(H_{N,N}+\rho I_N\right)^{-1}H_{N,M}\right). \tag{26}$$

Adding Gaussian noise ε to $\boldsymbol{f}_M$, the distribution of predictive value $\boldsymbol{y}^*$ is given as

$$p(\boldsymbol{y}^*|X_M^*,D)=p(\boldsymbol{y}^*|\boldsymbol{y})=N\left(\hat{\boldsymbol{\mu}},\hat{\Sigma}\right) \tag{27}$$

$$\hat{\boldsymbol{\mu}}=H_{M,N}\left(H_{N,N}+\rho I_N\right)^{-1}\boldsymbol{y} \tag{28}$$

$$\hat{\Sigma}=(H_{M,M}+\rho I_M)-H_{M,N}\left(H_{N,N}+\rho I_N\right)^{-1}H_{N,M}, \tag{29}$$

where $p(\boldsymbol{y}^*|X_M^*,D)$ represents the predictive distribution of $\boldsymbol{y}^*$ for new input variable X_M^* and the observed data set D, $\hat{\boldsymbol{\mu}}$ is an M dimensional mean vector, and $\hat{\Sigma}$ is an $M\times M$ covariance matrix. The second term of Eq. (29) represents a shift by existence of observation data D, and is regarded as a contribution from off-diagonal block matrices $H_{N,M}$ and $H_{M,N}$ of Eq. (18). The diagonal entries of covariance matrix $\hat{\Sigma}$ are equal to entries of a predictive variance vector $\boldsymbol{V}$=$(V_1,\ V_2,\cdots,V_M)^T$ with its i-th entry given by:

$$V_i=\hat{\Sigma}_{i,i}=\tilde{G}(x_i^*,x_i^*)+\rho-\boldsymbol{h}_{M,N,i}^T\left(H_{N,N}+\rho I_N\right)^{-1}\boldsymbol{h}_{N,M,i} \quad (i=1,2,\cdots,M), \tag{30}$$

$$\boldsymbol{h}_{M,N,i}^T=(\tilde{G}(x_i^*,x_1),\tilde{G}(x_i^*,x_2),\cdots,\tilde{G}(x_i^*,x_N)) \tag{31}$$

$$\boldsymbol{h}_{N,M,i}=(\tilde{G}(x_1,x_i^*),\tilde{G}(x_2,x_i^*),\cdots,\tilde{G}(x_N,x_i^*))^T \tag{32}$$

where $\boldsymbol{h}_{M,N,i}^T$ is an i-th row vector of the matrix $H_{M,N}$ and $\boldsymbol{h}_{N,M,i}$ is an i-th column vector of the matrix $H_{N,M}$. Note that each entry of the mean vector $\hat{\boldsymbol{\mu}}$ and variance vector $\boldsymbol{V}$ is a point-wise function of x_i^*. We also introduce a standard deviation vector $\boldsymbol{s}=(s_1,s_2,\cdots,\ s_M)^T$, where $s_i=\sqrt{V_i}$ $(i=1,2,\cdots,M)$.

2.4 Learning the hyperparameters of Green function kernel

In the case of the normalized Green function, a set of hyperparameters is $\boldsymbol{\theta}=(a,\rho)$, which determines the predictive distribution $p(\boldsymbol{y}^*|X_M^*,D)$ specified by mean and covariance matrix of Eqs. (28) and (29). In order to infer the predictive distribution $p(\boldsymbol{y}^*|X_M^*,D)$ we estimate the set of hyperparameters $\boldsymbol{\theta}=(a,\rho)$ by maximizing log likelihood, given as follows:

$$l=l(\boldsymbol{\theta})=\log p(\boldsymbol{y}^*|X_M^*,D,\boldsymbol{\theta}\,)$$

$$=-\frac{1}{2}\boldsymbol{y}^T(H_{NN}+\rho I_N)^{-1}\boldsymbol{y}-\frac{1}{2}\log\det(H_{NN}+\rho I_N)-\frac{N}{2}\log 2\pi \tag{33}$$

The parameter a is contained in the normalized Green function $\tilde{G}(x,x')=\tilde{G}(a\colon x,x')$ of Eq. (8) in the covariance matrix. A data set D is composed of input variable X_N and observation $\boldsymbol{y}$, and X_N are substituted to the covariance matrix $H_{N,N}$.

3. Numerical results

In this section, we present numerical results, which are performed by Python 3.7. We put a difference interval Δ= 0.01 for variable x^* in (0,1), and take an input variable $\mathrm{X}_M^*=\{x_1^*,x_2^*,\cdots,x_M^*\}$ as $x_i^*=0.01i$ $(i=1,2,\cdots,99)$ or equivalently $M=99$, throughout this section. We also give $N=15$ data.

3.1 Green function basis regression

We present numerical results concerning the application of Green function to a Gaussian process regression algorithm. Hereafter, we call this regression algorithm “the Green algorithm”, for short. The predictive distribution of $\boldsymbol{y}^*$ of Eqs. (27)-(29) by Green algorithm is obtained from observation data set D, kernel function $\tilde{G}(x,x')$ of Eq. (8) and the variance of Gaussian noise ρ. We consider two kinds of observed data sets D_1 and D_2 provided as $N=15$ observations.

$$D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_{15},y_{15})\}, \tag{34}$$

$$X_N=X_{15}=\{x_1,x_2,\cdots,x_{15}\},\ x_i=0.1+\tfrac{0.8}{14}(i-1) \tag{35}$$

$$D_1\colon y_i=F(x_i),F(x)=7\big(1-\exp(-5|x-0.5|)\big) \quad (i=1,2,\cdots,15) \tag{36}$$

$$D_2\colon y_i=F(x_i)+\text{randomness}\ (i=1,2,\cdots,15) \tag{37}$$

The above $F(x)$ possesses a cusp at $x=0.5$.

Concerning the kernel function and the noise, we search for a set of parameters $\boldsymbol{\theta}=(a,\rho)$, which attains maximum of log likelihood l of Eq. (33).

3.2 noise free $\rho=0$ fixed case

First we show the noise free $\rho=0$ fixed case. Figure 2 shows a dependence of a log likelihood l on the parameter a, and the predictive distribution $p(\boldsymbol{y}^*|X_M^*,D_1)$. Left figure shows that l attains its maximum at $a=2.88$. Table I shows l takes the maximum value $l=-24.79$ at $a=2.88$. Right figure of Fig. 2 illustrates the predictive distribution $p(\boldsymbol{y}^*|X_M^*,D_1)$, which means posterior distribution of $\boldsymbol{y}^*=(y_1^*,y_2^*,\cdots,y_{99}^*)^T$ for input variable $\mathrm{X}_M^*=\{x_1^*,x_2^*,\cdots,y_{99}^*\}$ if data set D_1 is observed. The shaded region spans between $[\mu-s,\mu+s]$ in the vertical direction. It is observed that the span of the shaded region depends on x^* and is the smallest in the neighborhood of the data points.

In Fig. 3, we put $a=2.88$. Figure 3 shows each term of the variance $V_i=V(x_i)$ of Eq. (30). The first term $\tilde{G}(x_i^*,x_i^*)$, which is shown as a black curve, is equivalent to the diagonal value of Fig. 1. The second term of Eq. (30), $\boldsymbol{h}_{M,N,i}^T{H_{N,N}}^{-1}\boldsymbol{h}_{N,M,i}$ in $\rho=0$ case, represents a shift by observation of data D_1, corresponding to the contribution of off-diagonal block matrix of covariance matrix of Eq. (18). Contribution of the second term is as large as the first term in the neighborhood of data points, but is smaller at a distance of data points. Hence, the variance V is smaller if x_i^* is closer to the data points and larger if x_i^* is further. This means that the precision increases near data points due to the observation of D_1. Concerning the value of V in Fig. 3 and that of $2s$, which is a span of the shaded region $[\mu - s, \mu + s]$ in Fig. 2, the equality $2\sqrt{V} = 2s$ holds.

Figure 4 shows a-dependence of the first and the second term of $V_i = V(x_i^*)$ in Eq. (30) at $x^* = 0.19$ and 0.5 in the $\rho = 0$ case. We select $x^* = 0.19$ as a middle position between data, and $x^* = 0.5$ as a data point. At $x^* = 0.19$ the first term $\tilde{G}(x_i^*, x_i^*)$ is monotone increasing with respect to a, and the second term (the blue curve) is convex upward. Whereas at $x^* = 0.5$ the first term and the second one (the red curve) are the same, and are monotone increasing with respect to a.

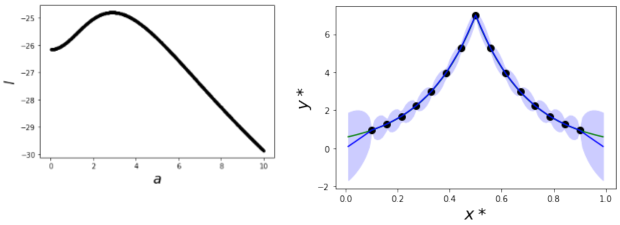

Figure 2　(left) a-dependence of l in the $\rho = 0$ fixed case by Green algorithm applied to data D_1. Maximum value is $l = -24.79$ at $a = 2.88$. (right) The illustration of $p(\boldsymbol{y}^*|X_M^*, D_1)$ with $(a, \rho) = (2.88, 0)$. In both $\rho = 0$ fixed case and changing ρ case, the same distribution is obtained. The blue curve: the predictive mean $\hat{\boldsymbol{\mu}}$. The shaded region: the correspondence to mean plus and minus s. Solid circles : data set D_1. The green curve: the function $F(x^*)$ on which observation data set D_1 exists.

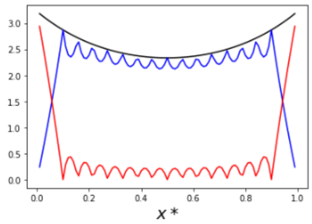

Figure 3　Each term of variance V of Eq. (30) as a point wise function of x^* in the $(a, \rho) = (2.88, 0)$ case by Green algorithm applied to data D_1. The black curve : the first term $\tilde{G}(x_i^*, x_i^*)$ of Eq. (30). The blue curve : the second term $\boldsymbol{h}_{M,N,i}^T {H_{N,N}}^{-1} \boldsymbol{h}_{N,M,i}$ of Eq. (30). The red curve: the variance $V_i = \tilde{G}(x_i^*, x_i^*) - \boldsymbol{h}_{M,N,i}^T {H_{N,N}}^{-1} \boldsymbol{h}_{N,M,i}$ as a point wise function of x_i^*.

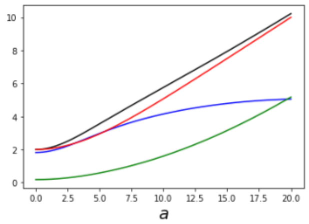

Figure 4　a-dependence of the first term $\tilde{G}(x_i^*, x_i^*)$ and second term $\boldsymbol{h}_{M,N,i}^T {H_{N,N}}^{-1} \boldsymbol{h}_{N,M,i}$ of variance $V(x_i^*)$ of Eq. (30) at $x^* = 0.19$ and 0.5 in $\rho = 0$ case by Green algorithm applied to data D_1. The black curve: the first term $\tilde{G}(0.19, 0.19)$. The blue curve: the second term $\boldsymbol{h}_{M,N,i}^T {H_{N,N}}^{-1} \boldsymbol{h}_{N,M,i}$ at $x^* = 0.19$. The green curve: variance $V(0.19)$. The red curve: the first term $\tilde{G}(0.5, 0.5)$, which is equal to the second term. Therefore variance $V(0.5)$, difference between the first and second term, is exactly equal to zero.

Table I　logarithm likelihood l with optimized hyperparameters

Data sets	Green algorithm	Gaussian algorithm
D_1 ($\rho = 0$ fixed)	$l = -24.79$ $a = 2.88$	$l = -26.35$ $\sigma = 0.07$
D_1	$l = -24.79$ $(a, \rho) = (2.88, 0.0)$	$l = -21.64$ $(\sigma, \rho) = (0.12, 0.1)$
D_2 ($\rho = 0$ fixed)	$l = -30.99$ $a = 5.61$	$l = -32.52$ $\sigma = 0.061$
D_2	$l = -30.00$ $(a, \rho) = (0.01, 0.8)$	$l = -30.29$ $(\sigma, \rho) = (0.20, 1.0)$

3.3　Comparison with Gaussian kernel regression in the noise free $\rho = 0$ fixed case

We here compare the results of Green algorithm obtained so far with the case of Gaussian kernel function, which is most widely used. We here call this algorithm “the Gaussian algorithm”, for short. As in the case of normalized Green function, we here adopt a normalized Gaussian kernel function,

$$v(x, x') = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x - x')^2}{2\sigma^2}\right). \qquad (38)$$

The parameter σ of Eq. (38) corresponds to parameters θ_1, θ_2 of a typical Gaussian kernel function $\theta_1 \exp\left(-\frac{(x - x')^2}{\theta_2}\right)$,[2), 5)] as $2\sigma^2 = \theta_2$ and $\frac{1}{\sqrt{2\pi}\sigma} = \theta_1$.

We search for a set of hyperparameters $\boldsymbol{\theta} = (\sigma, \rho)$, which attains maximum of log likelihood l[2), 5)].

Figure 5 shows σ-dependence of log likelihood l in the noise free $\rho = 0$ fixed case of Gaussian algorithm, and the predictive distribution $p(\boldsymbol{y}^*|X_M^*, D_1)$. Left figure shows the value of l increases if $0 < \sigma < 0.07$ and attains its maximum value $l = -26.35$ at $\sigma = 0.07$ and then decreases rapidly. The maximum value of l is also shown in Table I. Right figure shows a predictive distribution in the case of Gaussian kernel, which is smoother than the Green kernel.

Comparing Green and Gaussian algorithm in the case of $\rho = 0$ fixed for data D_1 of Table I, one finds that l is larger in the Green algorithm case. The a-dependence of l shows an obvious peak in Green algorithm case. On the other hand, l changes rapidly in the Gaussian case. From Figs. 2 and 5, we can observe that the blue curve of mean $\hat{\boldsymbol{\mu}}$ of Green algorithm coincides with the green curve of the function $F(x^*)$ better. The Green algorithm is considered to be suitable for predicting the function which possesses cusps. This fact is reflected by the fact that $G(x, y)$ has a pointed peak at $x = y$ whereas Gaussian function is a smooth function.

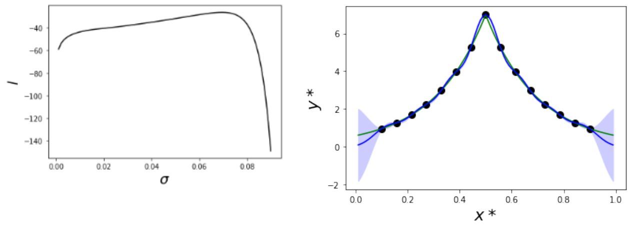

Figure 5 (left) σ-dependence of l in the $\rho = 0$ fixed case by Gaussian algorithm applied to data D_1. Maximum value is $l = -26.35$ at $\sigma = 0.07$. (right) The illustration of $p(\boldsymbol{y}^*|X_M^*, D_1)$ with $(\sigma, \rho) = (0.07, 0)$. The blue curve: the predictive mean $\hat{\boldsymbol{\mu}}$. The shaded region: the correspondence to mean plus and minus s. Solid circles: data set D_1. The green curve: the function $F(x^*)$.

3.4 The case of $\rho \geq 0$

Next, we take a Gaussian noise term ρ into account. In this case we change a pair of hyperparameters $\boldsymbol{\theta} = (a, \rho)$ in the Green algorithm, and $\boldsymbol{\theta} = (\sigma, \rho)$ in the Gaussian algorithm for data D_1.

In the Green algorithm, Fig. 6 shows a-dependence of l for $\rho = 0.0, 0.1, 0.2, \cdots, 1.0$. For every fixed ρ, l has a maximum with respect to a. The maximum value of l is monotone decreasing with respect to ρ. Hence l takes its maximum $l = -24.79$ at $(a, \rho) = (2.88, 0.0)$. Therefore predictive distribution $p(\boldsymbol{y}^*|X_M^*, D_1)$ by Green algorithm in the case of learning two parameters a and ρ is the same as that in the $\rho = 0$ fixed case of Fig. 2.

Figure 7 shows σ-dependence of l under the Gaussian algorithm for $\rho = 0.0, 0.1, 0.2, \cdots, 1.0$, and the predictive distribution $p(\boldsymbol{y}^*|X_M^*, D_1)$ with $(\sigma, \rho) = (0.12, 0.1)$. Left figure shows that l takes its maximum $l = -21.64$ at $(\sigma, \rho) = (0.12, 0.1)$, which is also shown in Table I. Right figure shows that mean $\hat{\boldsymbol{\mu}}$[2), 5)] gets smoother than in the case of $\rho = 0$ and shaded region is wider and smoother.

Comparing Green algorithm and Gauss algorithm, the maximum value of l of Green algorithm gets a little smaller than that of Gaussian algorithm, as is shown in Table I.

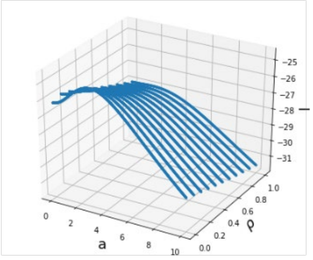

Figure 6 a and ρ-dependence of l by Green algorithm applied to data D_1. Maximum value is $l = -24.79$ at $(a, \rho) = (2.88, 0.0)$.

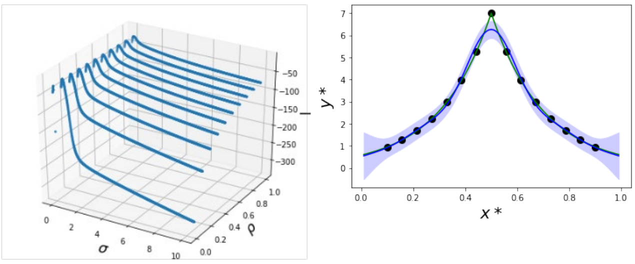

Figure 7 (left) σ and ρ-dependence of l by Gaussian algorithm applied to data D_1. Maximum value is $l = -21.64$ at $(\sigma, \rho) = (0.12, 0.1)$. (right) The illustration of $p(\boldsymbol{y}^*|X_M^*, D_1)$ with $(\sigma, \rho) = (0.12, 0.1)$. The blue curve: the predictive mean $\hat{\boldsymbol{\mu}}$. The shaded region: the correspondence to mean plus and minus s. Solid circles: data set D_1. The green curve: the function $F(x^*)$.

3.5 Introduction of randomness associated with $\boldsymbol{y}$

We finally introduce randomness in $\boldsymbol{y}$ of observation data D, which is written as D_2, following the section 6.4.2 in the reference[2)]. The dependences of l on the hyperparameters by Green and Gaussian algorithms are shown in Figs. 8 and 9.

We first investigate the Green algorithm. Left figure of Fig. 8 shows a-dependence of l for $\rho = 0.0, 0.1, \cdots, 1.0$. Table I shows that l in the $\rho = 0$ fixed case takes its maximum $l = -30.99$ at $a = 5.61$. If we change ρ, l takes its maximum $l = -30.00$ at

$(a,\rho)=(0.01,0.8)$. Right figure of Fig. 8 represents the predictive distribution $p(\boldsymbol{y}^*|X_M^*,D_2)$ in the $(a,\rho)=(0.01,0.8)$ case.

We next investigate the Gaussian algorithm. Left figure of Fig. 9 shows σ-dependence of l for $\rho=0.0,0.1,0.2,\cdots,2.0$. Table I shows that l, with $\rho=0$ fixed, takes its maximum $l=-32.52$ at $\sigma=0.061$. If we change ρ, l takes its maximum $l=-30.29$ at $(\sigma,\rho)=(0.20,1.0)$. Right figure of Fig. 9 represents the predictive distribution $p(\boldsymbol{y}^*|X_M^*,D_2)$ in the $(\sigma,\rho)=(0.20,1.0)$ case.

Comparing these two algorithms in Figs. 8 and 9, and Table I, we find that in both $\rho=0$ fixed case and the case of changing ρ, l takes a larger value in the Green algorithm case than in the Gaussian algorithm case.

In the cusp cases, Green algorithm shows almost the same or a little better performance, compared with Gaussian algorithm.

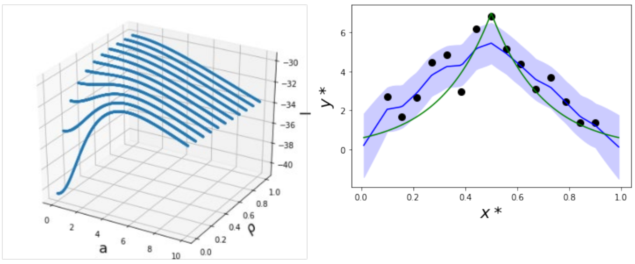

Figure 8. (left) a and ρ-dependence of l by Green algorithm applied to data D_2. Maximum value is $l=-30.00$ at $(a,\rho)=(0.01,0.8)$. (right) The illustration of $p(\boldsymbol{y}^*|X_M^*,D_2)$ with $(a,\rho)=(0.01,0.8)$. The blue curve: the predictive mean $\widehat{\boldsymbol{\mu}}$. The shaded region: the correspondence to mean plus and minus s. Solid circles : data set D_2. The green curve: the function $F(x^*)$.

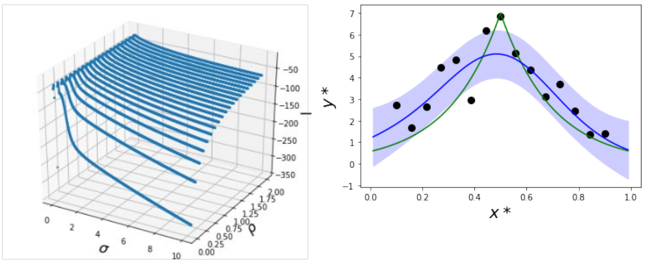

Figure 9. (left) σ and ρ-dependence of l by Gaussian algorithm applied to data D_2. Maximum value is $l=-30.29$ at $(\sigma,\rho)=(0.20,1.0)$. (right) The illustration of $p(\boldsymbol{y}^*|X_M^*,D_2)$ with $(\sigma,\rho)=(0.20,1.0)$. The blue curve: the predictive mean $\widehat{\boldsymbol{\mu}}$. The shaded region: the correspondence to mean plus and minus s. Solid circles: data set D_2. The green curve: the function $F(x^*)$.

4. Discussions

In order to take more information concerning covariance of y between x_i and x_j, we now investigate the covariance matrix $\widehat{\Sigma}$ of Eq. (29) hereafter. Generating a mesh of $M=99$, we introduce 99×99 matrix $\widehat{\Sigma}=\left(\widehat{\Sigma}_{i,j}\right)_{1\le i,j\le 99}$ whose (i,j)-th entry is also regarded as a function in $x_i^*=0.01i$ and $x_j^*=0.01j$ $(i,j=1,2,\cdots,99)$. That is to say, the covariance matrix of the predictive distribution is defined as a point-wise function, which is given as:

$$\begin{aligned}\widehat{\Sigma}_{i,j}&=\widehat{\Sigma}\left(x_i^*,x_j^*\right)=\tilde{G}\left(x_i^*,x_j^*\right)+\rho\delta_{i,j}\\&\qquad-\boldsymbol{h}_{M,N,i}^T\left(H_{N,N}+\rho I_N\right)^{-1}\boldsymbol{h}_{N,M,j}\\&=\tilde{G}(0.01i,0.01j)+\rho\delta_{i,j}-\boldsymbol{h}_{M,N,i}^T\left(H_{N,N}+\rho I_N\right)^{-1}\boldsymbol{h}_{N,M,j}\\&\qquad(i,j=1,2,\cdots,99).\end{aligned}\tag{39}$$

Figure 10 shows the covariance matrix $\widehat{\Sigma}_{i,j}$ of Eq. (39) corresponding to predictive distribution $p(\boldsymbol{y}^*|X_M^*,D_1)$ of Fig. 2, where x,y,z directions correspond to $x_i^*=0.01i$, $x_j^*=0.01j$, and $\widehat{\Sigma}_{i,j}$, respectively. Left and right figures are two-dimensional contour and three-dimensional plot of covariance at $(x_i^*,\ x_j^*)$, respectively. Due to the observation of data D_1, covariance between x_i^* and x_j^* localizes around $x_i^*=x_j^*$ line. The diagonal value at $x_i^*=x_j^*$ is equal to the variance, and its square root, which is a standard deviation, corresponds to the span of shaded region of Fig. 2 as a point-wise function of x_i^*. Figure 11 shows the covariance matrix for $p(\boldsymbol{y}^*|X_M^*,D_2)$ of Fig. 8. Since $\rho\neq 0$, the values on diagonal $x_i^*=x_j^*$ points contain the noise term ρ.

5. Conclusion

Let us summarize the obtained results. In this paper, we proposed and implemented a regression algorithm based on Green function theory. We propose a covariance matrix composed of the normalized Green function. By applying Bayesian approach, the covariance matrix gives a predictive distribution. The Green algorithm shows almost the same or slightly superior performance compared with Gaussian algorithm if $f(x)$ possesses cusps.

Acknowledgment

The author would like to thank S. Kamei and I. Kayo of Tokyo University of Technology, S. Tomizawa of Toyota Technological Institute, and T. Miura of National Institute of Advanced Industrial Science and Technology for useful comments.

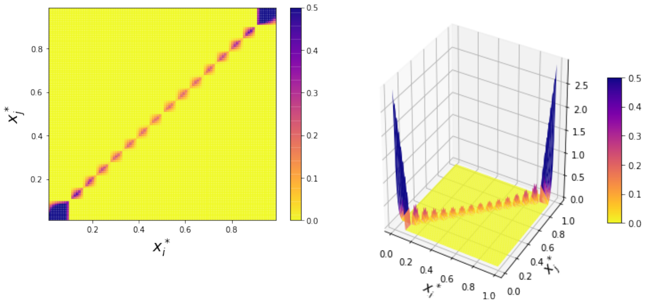

Figure 10. The plot of covariance matrix Eq. (39) of predictive distribution $p(\boldsymbol{y}^*|X_M^*, D_1)$ of Fig. 2 with $(a, \rho) = (2.88, 0)$ by Green algorithm. (left) 2-dimensional contour, (right) 3-dimensional plot.

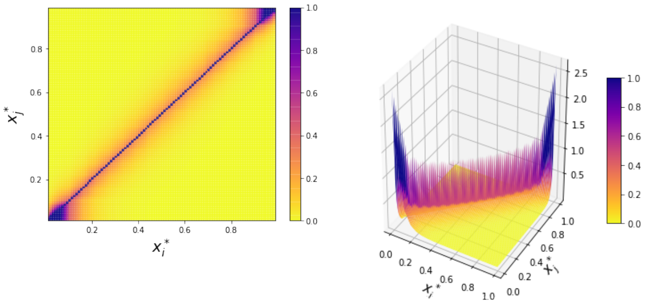

Figure 11. The plot of covariance matrix Eq. (39) of predictive distribution $p(\boldsymbol{y}^*|X_M^*, D_2)$ of Fig. 8 with $(a, \rho) = (0.01, 0.8)$ by Green algorithm. (left) 2-dimensional contour, (right) 3-dimensional plot.

References

1) B. Schölkopf and A. J. Smola, Learning with kernels, The MIT Press, Cambridge, 2002.

2) C. M. Bishop, Pattern Recognition and Machine Learning, Springer, Berlin, 2006.

3) C. E. Rasmussen and C. K. I. Williams, Gaussian Process for Machine Learning, The MIT Press, Cambridge, 2006.

4) K. Fukumizu, L. Song, and A. Gretton, Kernel Bayes' Rule: Bayesian Inference with Positive Definite Kernels, Journal of Machine Learning Research, vol.14, pp.3753-3783, 2013.

5) D. Mochihashi and S. Oba, Gaussian Process and Machine Learning, Kodansha, Tokyo, 2019.

6) M. Kanagawa, P. Hennig, D. Sejdinovic, and B. K Sriperumbudur, Gaussian Processes and Kernel Methods: A review on Connections and Equivalences, arXiv:1807.02582.

7) Y. Kametaka, Sugaku Seminar, vol.547, 2007 - vol.558, 2008 [in Japanese].

8) Y. Kametaka, K. Watanabe and A. Nagai, The best constant of Sobolev inequality in an n dimensional Euclidean space, Proc. Japan. Acad. Vol. 81, Ser. A, no. 3, pp.57-60, 2005.

9) M. Oikawa, A. Nagai and T. Yajima, Differential Equations 2nd edition, Saiensu-sha, Tokyo, 2019 [in Japanese].

10) S. Akaho, Kernel Tahenryo Kaiseki, Iwanami Shoten, Tokyo, 2008 [in Japanese].

11) K. Fukumizu, Kernel-ho Nyumon, Asakura Shoten, Tokyo, 2010 [in Japanese].