Approximating Posterior Gaussian Processes

Carson James

April 25, 2022

Outline

Gaussian Processes

Note

Let T be a set, $f : T \to \mathbb{R}$, $\mu : T \to \mathbb{R}$, $c : T^2 \to \mathbb{R}$, $x = (x_j)_{j=1}^n \in T^n$ and $t \in T$. We will write

- $f(x) := (f(x_j))_{j=1}^n \in \mathbb{R}^n$
- $\mu(x) := (\mu(x_j))_{j=1}^n \in \mathbb{R}^n$
- $c(x, x) := (c(x_i, x_j))_{i,j} \in \mathbb{R}^{n \times n}$
- $c(x, t) := (c(x_j, t))_{i,j} \in \mathbb{R}^{n \times 1}$
- $c(t, x) := (c(t, x_j))_{i,j} \in \mathbb{R}^{1 \times n}$

Definition

Let T be a set and $c : T^2 \to \mathbb{R}$. Then c is said to be **positive definite** if for each $(x_j)_{j=1}^n \in T^n$, the matrix $c(x, x)$ is positive definite.

Gaussian Processes

Definition

Let T be a set, $(\Omega, \mathcal{F}, P)$ a probability space, $\mu : T \to \mathbb{R}$, $c : T^2 \to \mathbb{R}$ symmetric and positive definite and $f : T \to L^2(\Omega, \mathcal{F}, P)$ (i.e. f is a random function from T to $\mathbb{R}$). Then f is said to be a **Gaussian Process** with mean function μ and covariance function c, denoted $f \sim GP(\mu, c)$, if for each $x = (x_j)_{j=1}^n \in T^n$, $f(x) \sim N_n(\mu(x), c(x, x))$.

Gaussian Processes

Fact

Let T be a set, $c : T^2 \to \mathbb{R}$ positive definite, $x = (x_j)_{j=1}^n \in T^n$, $y = (y_j)_{j=1}^n \in \mathbb{R}^n$. Suppose we have the following model:

$$
\begin{aligned}
y_i &= f(x_i) + \epsilon_i \\
\epsilon_i &\sim N(0, \sigma^2) \\
f &\sim GP(0, c)
\end{aligned}
$$

Then

$$f|x, y \sim GP(\tilde{\mu}, \tilde{c})$$

where

$$\tilde{\mu}(t) = c(t, x)[c(x, x) + \sigma^2 I]^{-1} y$$

and

$$\tilde{c}(s, t) = c(s, t) - c(s, x)[c(x, x) + \sigma^2 I]^{-1} c(x, t)$$

Gaussian Processes

Question

Suppose that we want to evaluate $\tilde{\mu}(t) = c(t,x)[c(x,x) + \sigma^2 I]^{-1} y$ and $\tilde{c}(s,t) = c(s,t) - c(s,x)[c(x,x) + \sigma^2 I]^{-1} c(x,t)$ for many input values of (s,t).
What do we do when $[c(x,x) + \sigma^2 I]^{-1}$ is too expensive to compute, or perhaps computable after a fair bit of time, but finding the values $\tilde{c}(s,t)$ repeadtedly for many inputs of (s,t) is not feasible?

Gaussian Processes

Question

Suppose that we want to evaluate $\tilde{\mu}(t) = c(t, x)[c(x, x) + \sigma^2 I]^{-1} y$ and $\tilde{c}(s, t) = c(s, t) - c(s, x)[c(x, x) + \sigma^2 I]^{-1} c(x, t)$ for many input values of (s, t).
What do we do when $[c(x, x) + \sigma^2 I]^{-1}$ is too expensive to compute, or perhaps computable after a fair bit of time, but finding the values $\tilde{c}(s, t)$ repeadtedly for many inputs of (s, t) is not feasible?

Answer

One thing to try would be to approximate $\tilde{\mu}$ and $\tilde{c}$. In this case we discuss approximating $\tilde{\mu}$ and $\tilde{c}$ with neural networks. Then, after training, evaluation would be constant time (with respect to data size).

Gaussian Processes

Question

If we want to approximate $\tilde{\mu}$ and $\tilde{c}$ with neural networks, we need a loss function to train the networks. What should this loss function be?

Gaussian Processes

Question

If we want to approximate $\tilde{\mu}$ and $\tilde{c}$ with neural networks, we need a loss function to train the networks. What should this loss function be?

Answer

We propose using a loss function derived from an alternative formulation of the posterior mean and covariance. To make this less vague, we will discuss the necessary background, which consists of calculus on Banach spaces and convex analyis on Banach spaces and repoducing kernel Hilbert spaces .

Banach Spaces

Remark

When working with finite dimensional normed vector spaces, all linear operators are continuous. However, in general this is not true if we drop the assumption of finite dimensionality. We will introduce the concept of boundedness, which is equivalent to continuity in the context of linear operators.

Banach Spaces

Definition

Let X, Y be a normed vector spaces and $T : X \to Y$ a linear map. Then T is said to be **bounded** if there exists $C \geq 0$ such that for each $x \in X$,

$$\|Tx\| \leq C\|x\|$$

We define

$$L(X; Y) = \{T : X \to Y : T \text{ is linear and bounded}\}$$

Banach Spaces

Definition

Let X, Y be a normed vector spaces and $T : X \to Y$ a linear map. Then T is said to be **bounded** if there exists $C \geq 0$ such that for each $x \in X$,

$$\|Tx\| \leq C\|x\|$$

We define

$$L(X; Y) = \{T : X \to Y : T \text{ is linear and bounded}\}$$

Definition

Let X be a normed vector space over $\mathbb{R}$. We define the **dual space of** X, denoted X^*, by $X^* = L(X; \mathbb{R})$. Let $T : X \to \mathbb{R}$. Then T is said to be a **bounded linear functional on** X if $T \in X^*$.

Banach Spaces

Definition

Let $X_1, \ldots, X_n$ and Y be a normed vector spaces and $T : \prod_{j=1}^{n} X_j \to Y$ a multilinear linear map. Then T is said to be **bounded** if there exists $C \geq 0$ such that for each $(x_j)_{j=1}^{n} \in \prod_{j=1}^{n} X_j$,

$$\|T(x_1, \ldots, x_n)\| \leq C\|x_1\| \ldots \|x_n\|$$

We define

$$L^n(X_1, \ldots, X_n; Y) = \{T : X \to Y : T \text{ is multilinear and bounded}\}$$

If $X_1, \ldots, X_n = X$, we write $L^n(X; Y)$ in place of $L^n(X, \ldots, X; Y)$.

Banach Spaces

Definition

Let X and Y be normed vector spaces. We define the **operator norm**, denoted $\|\cdot\| : L^2(X;Y) \to [0,\infty)$ by

$$\|T\| = \inf\{C \geq 0 : \text{for each } (x_1, x_2) \in X^2,\ \|T(x_1, x_2)\| \leq C\|x_1\|\|x_2\|\}$$

Banach Spaces

Definition

Let X and Y be normed vector spaces. We define the **operator norm**, denoted $\|\cdot\| : L^2(X;Y) \to [0,\infty)$ by

$$\|T\| = \inf\{C \geq 0 : \text{for each } (x_1,x_2) \in X^2,\ \|T(x_1,x_2)\| \leq C\|x_1\|\|x_2\|\}$$

Fact

Let X and Y be normed vector spaces. Then $\|\cdot\| : L^2(X;Y) \to [0,\infty)$ is a norm.

Banach Spaces

Definition

Let X and Y be normed vector spaces. We define the **operator norm**, denoted $\|\cdot\| : L^2(X;Y) \to [0,\infty)$ by

$$\|T\| = \inf\{C \geq 0 : \text{for each } (x_1, x_2) \in X^2,\ \|T(x_1, x_2)\| \leq C\|x_1\|\|x_2\|\}$$

Fact

Let X and Y be normed vector spaces. Then $\|\cdot\| : L^2(X;Y) \to [0,\infty)$ is a norm.

Remark

Let X and Y be normed vector spaces. We may identify $L(X; L(X;Y))$ and $L^2(X;Y)$ via the isometric isomorphism $L(X; L(X;Y)) \to L^2(X;Y)$ given by $\phi \mapsto \psi_\phi$ where

$$\psi_\phi(x_1, x_2) = \phi(x_1)(x_2)$$

This immediately generalizes to higher dimensions.

Banach Spaces

Remark

To discuss differentiation on normed vector spaces, we need to be able to talk about limits. Assuming completeness lets us do just this.

Banach Spaces

Remark

To discuss differentiation on normed vector spaces, we need to be able to talk about limits. Assuming completeness lets us do just this.

Definition

Let X be a normed vector space. Then X is said to be a **Banach space** if X is complete.

Banach Spaces

Definition

Let X, Y be a banach spaces, $A \subset X$ open, $f : A \to Y$ and $x_0 \in A$. Then f is said to be **(1-st order) Frechet differentiable at x_0** if there exists $Df(x_0) \in L(X; Y)$ such that,

$$f(x_0 + h) = f(x_0) + Df(x_0)(h) + o(\|h\|) \qquad \text{as } h \to 0$$

If f is Frechet differentiable at x_0, we define the **Frechet derivative of f at x_0** to be $Df(x_0)$. We say that f is **(1-st order) Frechet differentiable** if for each $x_0 \in A$, f is Frechet differentiable at x_0.
If f is Frechet differentiable, we define the **(1-st order) Frechet derivative** of f, denoted $Df : A \to L(X; Y)$, by

$$x \mapsto Df(x)$$

Banach Spaces

Definition

Continuing inductively, if f is $(n-1)$-th order Frechet differentiable, f is said to be **n-th order Frechet differentiable at x_0** if $D^{n-1}f$ is Frechet differentiable at x_0. We define $D^n f(x_0) = D(D^{n-1}f)(x_0)$. If f is n-th order Frechet differentiable, we define the **(n-th) order Frechet derivative** of f, denoted $D^n f : A \to L^n(X; Y)$, by

$$x \mapsto D^n f(x)$$

Remark

Using the identification mentioned earlier, we may think of the n-th Frechet derivative as a bounded multilinear map: $D^n f(x_0) \in L^n(X; Y)$.

Banach Spaces

Remark

The tools used to obtain the following results:

- Frechet Derivative
- Bochner Integral
- Hahn-Banach Theorem

Fact

Let X, Y be Banach spaces and $f \in L(X; Y)$. Then f is Frechet differentiable and for each $x_0 \in X$, $Df(x_0) = f$.

Banach Spaces

Remark

The tools used to obtain the following results:

- Frechet Derivative
- Bochner Integral
- Hahn-Banach Theorem

Fact

Let X, Y be Banach spaces and $f \in L(X; Y)$. Then f is Frechet differentiable and for each $x_0 \in X$, $Df(x_0) = f$.

Fact

Let X, Y, Z be Banach spaces, $f : X \to Y$, $g : Y \to Z$ and $x_0 \in X$. If f is Frechet differentiable at x_0 and g is Frechet differentiable at $f(x_0)$, then $g \circ f$ is Frechet differentiable at x_0 and

$$D(g \circ f)(x_0) = Dg(f(x_0)) \circ Df(x_0)$$

Banach Spaces

Remark

The tools used to obtain the following results:

- Frechet Derivative
- Bochner Integral
- Hahn-Banach Theorem

Fact

Let X, Y be Banach spaces and $f \in L(X;Y)$. Then f is Frechet differentiable and for each $x_0 \in X$, $Df(x_0) = f$.

Fact

Let X, Y, Z be Banach spaces, $f : X \to Y$, $g : Y \to Z$ and $x_0 \in X$. If f is Frechet differentiable at x_0 and g is Frechet differentiable at $f(x_0)$, then $g \circ f$ is Frechet differentiable at x_0 and

$$D(g \circ f)(x_0) = Dg(f(x_0)) \circ Df(x_0)$$

Fact

Banach Spaces

Fact

Let X, Y be a Banach spaces, $A \subset X$ open and convex and $f : A \to Y$. Suppose that f is Frechet differentiable. If for each $x \in A$, $Df(x) = 0$, then f is constant.

Banach Spaces

Fact

Let X, Y be a Banach spaces, $A \subset X$ open and convex and $f : A \to Y$. Suppose that f is Frechet differentiable. If for each $x \in A$, $Df(x) = 0$, then f is constant.

Fact

Let X, Y be a Banach spaces, $A \subset X$ open and convex and $f, g : A \to Y$. Suppose that f and g are Frechet differentiable. If $Df = Dg$, then there exists $c \in Y$ such that $f = g + c$.

Banach Spaces

Fact

Let X, Y be a Banach spaces, $A \subset X$ open and convex and $f : A \to Y$. Suppose that f is Frechet differentiable. If for each $x \in A$, $Df(x) = 0$, then f is constant.

Fact

Let X, Y be a Banach spaces, $A \subset X$ open and convex and $f, g : A \to Y$. Suppose that f and g are Frechet differentiable. If $Df = Dg$, then there exists $c \in Y$ such that $f = g + c$.

Fact

Let X be a Banach spaces, $A \subset X$ open, $f : A \to \mathbb{R}$ and $x_0 \in A$. Suppose that f is Frechet differentiable at x_0. If f has a local minimum at x_0, then $Df(x_0) = 0$.

Banach Spaces

Fact

Let Y be a separable Banach space and $f \in C^1_Y(a,b)$. Then for each $x, x_0 \in (a,b)$, $x_0 < x$ implies that

1. f' is Bochner integrable on $(x_0, x]$
2. $$f(x) - f(x_0) = \int_{(x_0,x]} f' dm$$

Banach Spaces

Fact

Let Y be a separable Banach space and $f \in C^1_Y(a,b)$. Then for each $x, x_0 \in (a,b)$, $x_0 < x$ implies that

1. f' is Bochner integrable on $(x_0, x]$
2. $$f(x) - f(x_0) = \int_{(x_0,x]} f' dm$$

Fact

Let Y be a separable Banach space, $A \subset X$ open and convex, $f \in C^n_Y(A)$ and $x_0 \in A$. Then

$$f(x_0 + h) = \sum_{k=0}^{n} \frac{1}{k!} D^k f(x_0)(h, \ldots, h) + o(\|h\|^n) \qquad \text{as } h \to 0$$

Convex Analysis

Definition

Let X be a vector space, $A \subset X$ convex, $f : A \to \mathbb{R}$ convex and $x_0 \in A$. Then f is said to be **convex** if for each $t \in [0, 1]$ and $x, y \in A$,

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)$$

Convex Analysis

Definition

Let X be a vector space, $A \subset X$ convex, $f : A \to \mathbb{R}$ convex and $x_0 \in A$. Then f is said to be **convex** if for each $t \in [0, 1]$ and $x, y \in A$,

$$f(tx + (1 - t)y) \leq tf(x) + (1 - t)f(y)$$

Fact

Let X be a vector space, $A \subset X$ convex, $f : A \to \mathbb{R}$ convex and $x_0 \in A$. Then f has a local minimum at x_0 iff f has a global minimum at x_0.

Convex Analysis

Definition

Let X be a vector space, $A \subset X$ convex, $f : A \to \mathbb{R}$ convex and $x_0 \in A$. Then f is said to be **convex** if for each $t \in [0,1]$ and $x, y \in A$,

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)$$

Fact

Let X be a vector space, $A \subset X$ convex, $f : A \to \mathbb{R}$ convex and $x_0 \in A$. Then f has a local minimum at x_0 iff f has a global minimum at x_0.

Fact

Let X be a vector space, $A \subset X$ convex and $f : A \to \mathbb{R}$ strictly convex. If f has a local minimum, then there exists a unique $x_0 \in A$ such that $f(x_0) = \min\limits_{x \in A} f(x)$.

Convex Analysis

Fact

Let X be a Banach space, $A \subset X$ open and convex, $f : A \to \mathbb{R}$ convex, $x_0 \in A$. Suppose that f is 2nd order Frechet differentiable. Then f is convex iff for each $x_0 \in A$, $D^2 f(x_0)$ is positive semidefinite.

Remark

By positive semidefinite, we mean $D^2 f(x_0)(h, h) \geq 0$ for $h \neq 0$.

Convex Analysis

Fact

Let X be a Banach space, $A \subset X$ open and convex, $f : A \to \mathbb{R}$ convex, $x_0 \in A$. Suppose that f is 2nd order Frechet differentiable. Then f is convex iff for each $x_0 \in A$, $D^2 f(x_0)$ is positive semidefinite.

Remark

By positive semidefinite, we mean $D^2 f(x_0)(h, h) \geq 0$ for $h \neq 0$.

Fact

Let X be a Banach space, $A \subset X$ open and convex, $f : A \to \mathbb{R}$ convex, $x_0 \in A$. Suppose that f is 2nd order Frechet differentiable. If for each $x_0 \in A$, $D^2 f(x_0)$ is positive definite, then f is strictly convex.

Reproducing Kernel Hilbert Spaces

Definition

Let H be an inner product space. Then H is said to be a **Hilbert space** if H is complete with respect to the norm induced by the inner product.

Reproducing Kernel Hilbert Spaces

Definition

Let H be an inner product space. Then H is said to be a **Hilbert space** if H is complete with respect to the norm induced by the inner product.

Remark

We will be assuming the Hilbert space is real.

Reproducing Kernel Hilbert Spaces

Definition

Let H be an inner product space. Then H is said to be a **Hilbert space** if H is complete with respect to the norm induced by the inner product.

Remark

We will be assuming the Hilbert space is real.

Fact

Let H be an inner product space. Then for each $x, y \in H$, $|\langle x, y \rangle| \leq \|x\| \|y\|$ with equality iff $x \in \text{span}(y)$.

Reproducing Kernel Hilbert Spaces

Definition

Let H be a Hilbert space. Define $\phi : H \to H^*$ by $x \mapsto x^*$ where

$$x^* y = \langle x, y \rangle$$

Reproducing Kernel Hilbert Spaces

Definition

Let H be a Hilbert space. Define $\phi : H \to H^*$ by $x \mapsto x^*$ where

$$x^* y = \langle x, y \rangle$$

Fact

Let H be a Hilbert space. Then $\phi : H \to H^*$ defined above is an isometric isomorphism.

Reproducing Kernel Hilbert Spaces

Definition

Let H be a Hilbert space, $f : H \to \mathbb{R}$ and $x_0 \in H$. Suppose that f is Frechet differentiable at x_0 so that $Df(x_0) \in H^*$. We define the **gradient of** f **at** x_0, denoted $\nabla f(x_0) \in H$, by

$$\nabla f(x_0) = Df(x_0)^*$$

That is, $\nabla f(x_0)$ is the unique element of H such that for each $y \in H$,

$$\langle \nabla f(x_0), y \rangle = Df(x_0)(y)$$

Reproducing Kernel Hilbert Spaces

Definition

Let T be a set and $H \subset \mathbb{R}^T$ a hilbert space. For $t \in T$, we define the **evaluation functional at** t, denoted $l_t : H \to \mathbb{R}$, by

$$l_t(f) = f(t)$$

Reproducing Kernel Hilbert Spaces

Definition

Let T be a set and $H \subset \mathbb{R}^T$ a hilbert space. For $t \in T$, we define the **evaluation functional at** t, denoted $l_t : H \to \mathbb{R}$, by

$$l_t(f) = f(t)$$

The space H is said to be a **reproducing kernel Hilbert space (RKHS)** if for each $t \in T$, $l_t \in H^*$ (i.e. l_t is bounded).

Reproducing Kernel Hilbert Spaces

Definition

Let T be a set and $H \subset \mathbb{R}^T$ a hilbert space. For $t \in T$, we define the **evaluation functional at** t, denoted $l_t : H \to \mathbb{R}$, by

$$l_t(f) = f(t)$$

The space H is said to be a **reproducing kernel Hilbert space (RKHS)** if for each $t \in T$, $l_t \in H^*$ (i.e. l_t is bounded).
If H is an RKHS, the Riesz representation theorem implies that for each $t \in T$, there exists $c_t \in H$ such that for each $f \in H$, $\langle c_t, f \rangle = f(t)$.

Reproducing Kernel Hilbert Spaces

Definition

Let T be a set and $H \subset \mathbb{R}^T$ a hilbert space. For $t \in T$, we define the **evaluation functional at** t, denoted $l_t : H \to \mathbb{R}$, by

$$l_t(f) = f(t)$$

The space H is said to be a **reproducing kernel Hilbert space (RKHS)** if for each $t \in T$, $l_t \in H^*$ (i.e. l_t is bounded).
If H is an RKHS, the Riesz representation theorem implies that for each $t \in T$, there exists $c_t \in H$ such that for each $f \in H$, $\langle c_t, f \rangle = f(t)$.
If H is an RKHS, we define the **reproducing kernel** associated to H, denoted $c_H : T^2 \to \mathbb{R}$, by

$$c_H(s,t) = \langle c_s, c_t \rangle$$

Reproducing Kernel Hilbert Spaces

Fact

Let T be a set and $c : T^2 \to \mathbb{R}$. If c is symmetric and positive definite, then there exists a unique reproducing kernel Hilbert space $H \subset \mathbb{R}^T$ such that $c_H = c$.

Reproducing Kernel Hilbert Spaces

Fact

Let T be a set, $c : T^2 \to \mathbb{R}$ a symmetric, postivie definite kernel on T, $H \subset \mathbb{R}^T$ the corresponding RKHS, $x = (x_j)_{j=1}^n \in T^n$, $\lambda > 0$ and $y = (y_j)_{j=1}^n \in \mathbb{R}^n$.

Reproducing Kernel Hilbert Spaces

Fact

Let T be a set, $c : T^2 \to \mathbb{R}$ a symmetric, postivie definite kernel on T, $H \subset \mathbb{R}^T$ the corresponding RKHS, $x = (x_j)_{j=1}^n \in T^n$, $\lambda > 0$ and $y = (y_j)_{j=1}^n \in \mathbb{R}^n$.
Define $L_{\lambda,y} : H \to \mathbb{R}$ by

$$L_{\lambda,y}(f) = \sum_{j=1}^n (y_j - f(x_j))^2 + \lambda \|f\|_H^2$$

Reproducing Kernel Hilbert Spaces

Fact

Let T be a set, $c : T^2 \to \mathbb{R}$ a symmetric, postivie definite kernel on T, $H \subset \mathbb{R}^T$ the corresponding RKHS, $x = (x_j)_{j=1}^n \in T^n$, $\lambda > 0$ and $y = (y_j)_{j=1}^n \in \mathbb{R}^n$.
Define $L_{\lambda,y} : H \to \mathbb{R}$ by

$$L_{\lambda,y}(f) = \sum_{j=1}^{n} (y_j - f(x_j))^2 + \lambda \|f\|_H^2$$

Put $\hat{f} = \arg\min_{f \in H} L_{\lambda,y}(f)$.

Reproducing Kernel Hilbert Spaces

Fact

Let T be a set, $c : T^2 \to \mathbb{R}$ a symmetric, postivie definite kernel on T, $H \subset \mathbb{R}^T$ the corresponding RKHS, $x = (x_j)_{j=1}^n \in T^n$, $\lambda > 0$ and $y = (y_j)_{j=1}^n \in \mathbb{R}^n$.
Define $L_{\lambda,y} : H \to \mathbb{R}$ by

$$L_{\lambda,y}(f) = \sum_{j=1}^{n} (y_j - f(x_j))^2 + \lambda \|f\|_H^2$$

Put $\hat{f} = \underset{f \in H}{\arg\min}\, L_{\lambda,y}(f)$.
Then there exist $(\hat{\alpha}_j)_{j=1}^n \subset \mathbb{R}$ such that

$$\hat{f}(t) = \sum_{j=1}^{n} \hat{\alpha}_j c(t, x_j)$$

Reproducing Kernel Hilbert Spaces

Remark

Recall that we defined $c(x, x) \in \mathbb{R}^{n \times n}$ by $c(x, x)_{i,j} = c(x_i, x_j)$. Some regular calculus shows that $\hat{\alpha} = (c(x, x) + \lambda I)^{-1} y$. Another way to write this is

$$\hat{f}(t) = c(t, x)(c(x, x) + \lambda I)^{-1} y$$

Remark

Hopefully this looks familiar. Indeed, $\hat{f}$ is the posterior mean function of $f | x, y$ from our original model when $\lambda = \sigma^2$.

Reproducing Kernel Hilbert Spaces

Remark

Define $Q : H \to \mathbb{R}$ by

$$Q(f) = \sum_{j=1}^{n} (y_j - f(x_j))^2$$

Reproducing Kernel Hilbert Spaces

Remark

Define $Q : H \to \mathbb{R}$ by

$$Q(f) = \sum_{j=1}^{n} (y_j - f(x_j))^2$$

We can write rewrite $Q(f)$ as

$$Q(f) = \|I_x(f) - y\|_2^2$$

where $I_x \in L(H; \mathbb{R}^n)$ is given by

$$I_x(f) = (f(x_j))_{j=1}^{n}$$

Reproducing Kernel Hilbert Spaces

Remark

Writing this out, we see that

$$
\begin{aligned}
Q(f_0 + h) &= \|I_x(f_0) - y\|_2^2 + 2(I_x(f_0) - y)^T I_x(h) + \|I_x(h)\|_2^2 \\
&= Q(f_0) + [\text{lin funct of } h] + [\text{bilin funct of } (h, h)]
\end{aligned}
$$

Reproducing Kernel Hilbert Spaces

Remark

Writing this out, we see that

$$\begin{aligned} Q(f_0 + h) &= \|I_x(f_0) - y\|_2^2 + 2(I_x(f_0) - y)^T I_x(h) + \|I_x(h)\|_2^2 \\ &= Q(f_0) + [\text{lin funct of } h] + [\text{bilin funct of } (h, h)] \end{aligned}$$

Equating terms from Taylor's theorem, we see that $D^2 Q(f_0)(h, h) = 2\|I_x(h)\|_2^2$, which is p.s.d. So Q is convex. Since $\|f_0\|_H^2 = \langle f_0, f_0 \rangle$ and $\langle \cdot, \cdot \rangle$ is a positive definite bounded bilinear function on $H \times H$, $\|\cdot\|_H$ is strictly convex. Since $\lambda > 0$, L is strictly convex.

Reproducing Kernel Hilbert Spaces

Remark

Similar to before, writing out $L(f_0 + h)$, we get

$$L_{\lambda,y}(f_0 + h) = L_{\lambda,y}(f_0) + 2(I_x(f_0) - y)^T I_x(h) + 2\lambda\langle f_0, h\rangle + o(\|h\|^2)$$

Reproducing Kernel Hilbert Spaces

Remark

Similar to before, writing out $L(f_0 + h)$, we get

$$L_{\lambda,y}(f_0 + h) = L_{\lambda,y}(f_0) + 2(l_x(f_0) - y)^T l_x(h) + 2\lambda\langle f_0, h\rangle + o(\|h\|^2)$$

So

$$\begin{aligned} DL_{\lambda,y}(f_0)(h) &= 2(l_x(f_0) - y)^T l_x(h) + 2\lambda\langle f_0, h\rangle \\ &= 2\sum_{j=1}^{n}(f_0(x_j) - y_j)\langle c_{x_j}, h\rangle + 2\lambda\langle f_0, h\rangle \\ &= \left\langle 2\left[\sum_{j=1}^{n}(f_0(x_j) - y_j)c_{x_j} + \lambda f_0\right], h\right\rangle \end{aligned}$$

Remark

Similar to before, writing out $L(f_0 + h)$, we get

$$L_{\lambda,y}(f_0 + h) = L_{\lambda,y}(f_0) + 2(l_x(f_0) - y)^T l_x(h) + 2\lambda\langle f_0, h\rangle + o(\|h\|^2)$$

So

$$\begin{aligned} DL_{\lambda,y}(f_0)(h) &= 2(l_x(f_0) - y)^T l_x(h) + 2\lambda\langle f_0, h\rangle \\ &= 2\sum_{j=1}^{n}(f_0(x_j) - y_j)\langle c_{x_j}, h\rangle + 2\lambda\langle f_0, h\rangle \\ &= \left\langle 2\left[\sum_{j=1}^{n}(f_0(x_j) - y_j)c_{x_j} + \lambda f_0\right], h\right\rangle \end{aligned}$$

Hence

$$\nabla L_{\lambda,y}(f_0) = 2\left[\sum_{j=1}^{n}(f_0(x_j) - y_j)c_{x_j} + \lambda f_0\right]$$

Approximating Posterior Gaussian Processes

Remark

Recall our model: Let T be a set and $x = (x_j)_{j=1}^n \in T^n$, $y = (y_j)_{j=1}^n \in \mathbb{R}^n$. Recall that if

$$
\begin{aligned}
y_i &= f(x_i) + \epsilon_i \\
\epsilon_i &\sim N(0, \sigma^2) \\
f &\sim GP(0, c)
\end{aligned}
$$

Then

$$f|x, y \sim GP(\tilde{\mu}, \tilde{c})$$

where

$$\tilde{\mu}(t) = c(t, x)[c(x, x) + \sigma^2 I]^{-1} y$$

and

$$\tilde{c}(s, t) = c(s, t) - c(s, x)[c(x, x) + \sigma^2 I]^{-1} c(x, t)$$

Approximating Posterior Gaussian Processes

Remark

As pointed out previously,

$$\begin{aligned}\tilde{\mu} &= c(\cdot, x)[c(x, x) + \sigma^2 I]^{-1} y \\ &= \arg\min_{f \in H} L_{\sigma^2, y}(f)\end{aligned}$$

We may find $\tilde{c}$ similarly. For $t \in T$, we define $\bar{c}(\cdot, t)$ by

$$\begin{aligned}\bar{c}(\cdot, t) &= c(\cdot, x)[c(x, x) + \sigma^2 I]^{-1} c(x, t) \\ &= \arg\min_{f \in H} L_{\sigma^2, c(x,t)}(f)\end{aligned}$$

Approximating Posterior Gaussian Processes

Remark

As pointed out previously,

$$\begin{aligned}\tilde{\mu} &= c(\cdot, x)[c(x,x) + \sigma^2 I]^{-1} y \\ &= \arg\min_{f \in H} L_{\sigma^2, y}(f)\end{aligned}$$

We may find $\tilde{c}$ similarly. For $t \in T$, we define $\bar{c}(\cdot, t)$ by

$$\begin{aligned}\bar{c}(\cdot, t) &= c(\cdot, x)[c(x,x) + \sigma^2 I]^{-1} c(x, t) \\ &= \arg\min_{f \in H} L_{\sigma^2, c(x,t)}(f)\end{aligned}$$

Remark

Note that the posterior covariance function is given by $\tilde{c}(s,t) = c(s,t) - \bar{c}(s,t)$.

Approximating Posterior Gaussian Processes

Remark

Now, thanks to the background covered earlier, we know that

$$\nabla L_{\sigma^2,y}(\tilde{\mu}) = 2\left[\sum_{j=1}^{n}(\tilde{\mu}(x_j) - y_j)c_{x_j} + \sigma^2\tilde{\mu}\right]$$
$$= 0$$

and

$$\nabla L_{\sigma^2,c(x,t)}(\bar{c}(\cdot,t)) = 2\left[\sum_{j=1}^{n}(\bar{c}(x_j,t) - c(x_j,t))c_{x_j} + \sigma^2\bar{c}(\cdot,t)\right]$$
$$= 0$$

Approximating Posterior Gaussian Processes

Remark

This gives us the following two restrictions:

- for each $s \in T$,

$$\sum_{j=1}^{n} (\tilde{\mu}(x_j) - y_j) c(s, x_j) + \sigma^2 \tilde{\mu}(s) = 0$$

- for each $s, t \in T$,

$$\sum_{j=1}^{n} (\bar{c}(x_j, t) - c(x_j, t)) c(s, x_j) + \sigma^2 \bar{c}(s, t) = 0$$

Approximating Posterior Gaussian Processes

Remark

Now if we approximate $\tilde{\mu} : T \to \mathbb{R}$ by a neural network $g_\theta : T \to \mathbb{R}$ and $\bar{c} : T^2 \to \mathbb{R}$ by a neural network $h_\eta : T^2 \to \mathbb{R}$, substitution yields the following restrictions:

- for each $s \in T$,

$$\sum_{j=1}^{n} (g_\theta(x_j) - y_j) c(s, x_j) + \sigma^2 g_\theta(s) = 0$$

- for each $s, t \in T$,

$$\sum_{j=1}^{n} (h_\eta(x_j, t) - c(x_j, t)) c(s, x_j) + \sigma^2 h_\eta(s, t) = 0$$

Approximating Posterior Gaussian Processes

Remark

Focusing on g_θ, let $(s_k)_{k=1}^a$ be a grid of T. Using the triangle inequality and Jensen's inequality we obtain l^1 and l^2 loss functions given by

- $$l_1(\theta) = \frac{1}{a}\sum_{k=1}^{a}\left[\sum_{j=1}^{n}|g_\theta(x_j) - y_j|c(s_k, x_j) + \sigma^2|g_\theta(s_k)|\right]$$

- $$l_2(\theta) = \frac{1}{a}\sum_{k=1}^{a}\left[\left(\sum_{j=1}^{n}(g_\theta(x_j) - y_j)c(s_k, x_j)\right)^2 + \sigma^4 g_\theta(s_k)^2\right]$$

Approximating Posterior Gaussian Processes

Remark

We could also add an MSE penalty term,

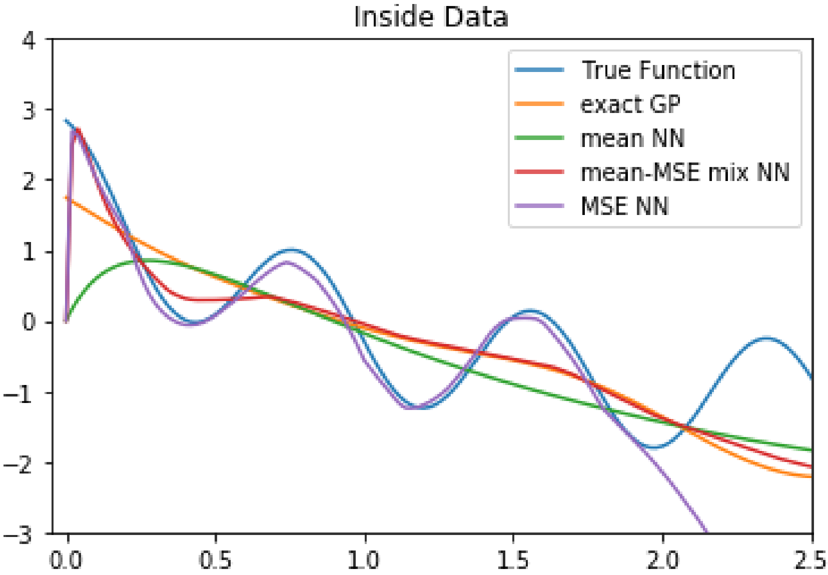
$$MSE(\theta) = \frac{1}{n}\sum_{j=1}^{n}(y_j - g_\theta(x_j))^2$$

to $l(\theta)$. The motivation for doing this is that it might yield a better approximation inside the data range while still retaining desirable properties of the posterior mean outside the data range. In the following slides we make a comparison. The model is

$$y_i \sim N(f(x_i, 0.1^2))$$

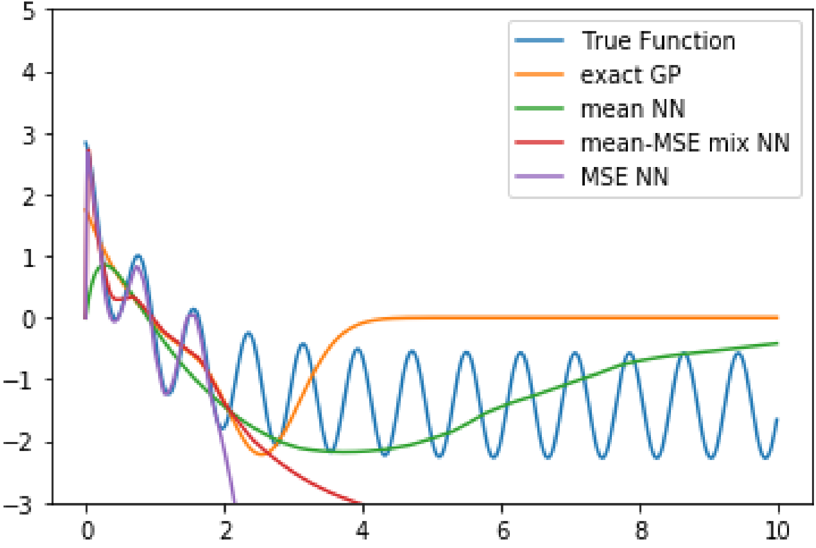
with the true data generating function given by

$$f_0(x) = 4 + 2e^{-x_i} + 0.5\cos(8 \cdot x_i)$$



Inside Data
True Function
exact GP
mean NN
mean-MSE mix NN
MSE NN
4
3
2
1
0
−1
−2
−3
0.0
0.5
1.0
1.5
2.0
2.5

Outside Data



Approximating Posterior Gaussian Processes

Remark

Focusing on h_η, let $(s_k, t_l)_{k,l}^{a,b}$ be a grid of T^2. Using the triangle inequality like before, except now adding a symmetric penalty term, we obtain an l^1 loss function given by

$$l_1(\eta) = \frac{1}{ab}\sum_{k=1}^{a}\sum_{l=1}^{b}\Bigg[\sum_{j=1}^{n}|h_\eta(x_j, t_l) - c(x_j, t_l)|c(s_k, x_j) + \sigma^2|h_\eta(s_k, t_l)|\Bigg] + \frac{1}{ab}\sum_{k=1}^{a}\sum_{l=1}^{b}|c(s_k, t_l) - c(t_l, s_k)|$$

Approximating Posterior Gaussian Processes

Remark

Using the Jensen's inequality like before, except now adding a symmetric penalty term, we obtain an l^2 loss function given by

$$
\begin{aligned}
l_2(\eta) = \frac{1}{ab}\sum_{k=1}^{a}\sum_{l=1}^{b}\Bigg[\bigg(\sum_{j=1}^{n}(h_\eta(x_j,t_l) - c(x_j,t_l))c(s_k,x_j)\bigg)^2 \\
+ \sigma^2 h_\eta(s_k,t_l)^2\Bigg] \\
+ \frac{1}{ab}\sum_{k=1}^{a}\sum_{l=1}^{b}(c(s_k,t_l) - c(t_l,s_k))^2
\end{aligned}
$$

References

- analysis notes
- integration notes
- RKHS's
- Representer Theorem