STATS205: Introduction to Nonparametric Statistics

Lecturer: Tengyu Ma | Lecture # 9
Scribe: Yibing Du and Scott Maran | May 28th, 2021

9.1 Review and overview

In the last lecture, we introduced the parametric Bayesian linear regression model with a theorem on the distributions of $\theta|S$ and then $y^*|x^*, S$. With the setup of $\theta \sim N(0, \tau^2 I), y^{(i)} = {x^{(i)}}^\top \theta + \epsilon^{(i)}$ where $\epsilon^{(i)} \sim N(0, \sigma^2)$, our goal is to predict $y^*|x^*, S$ where $S = \{(x^{(1)}, y^{(1)}), ..., (x^{(n)}, y^{(n)})\}$. S is the set of all data points, where all the $x^{(i)}$'s are deterministic whereas the $y^{(i)}$s are random. We're always conditioning on x^*, even if $y^*|x^*, S$ is sometimes written as $y^*|S$. In particular, we claimed

Theorem 9.1. *It follows that*

$$\theta|S \sim N(\frac{1}{\sigma^2} A^{-1} x^\top \overrightarrow{y}, A^{-1}), \tag{9.1}$$

$$y^*|x^*, S \sim N(\frac{1}{\sigma^2} {x^*}^\top A^{-1} x^\top \overrightarrow{y}, {x^*}^\top A^{-1} x^* + \sigma^2), \tag{9.2}$$

where the collection of all data points $X = \begin{bmatrix} {x^{(1)}}^\top \\ ... \\ {x^{(n)}}^\top \end{bmatrix} \in \mathbb{R}^{n \times d}$*, the collection of all labels* $\overrightarrow{y} = \begin{bmatrix} y^{(1)} \\ ... \\ y^{(n)} \end{bmatrix} \in \mathbb{R}^n$*, the covariance matrix* $A = \frac{1}{\sigma^2} x^\top x + \frac{1}{\tau^2} I$*, and the noise* $\overrightarrow{\epsilon} = \begin{bmatrix} \epsilon^{(1)} \\ ... \\ \epsilon^{(n)} \end{bmatrix}$.

In this class, we will prove the second claim and discuss nonparametric Bayesian methods, i.e. Gaussian process.

9.2 Proof of the Eq. (9.2) in Theorem 9.1

Extending from our discussion about interpretations of the theorem last time, we will prove the theorem using an approach that's also useful in the nonparametric scenario—we are motivated to prove the parametric case as it will be insightful to the nonparametric case. We note that we can directly prove the second claim without having to prove the first claim, which is helpful as claim one is harder to prove in the nonparametric case.

The general idea is that the posterior, conditional distributions and the prior are all Gaussian here. With a group of jointly Gaussian-distributed random variables, the following lemma is useful for looking for the posterior of one conditional on the other. Suppose we have a joint Gaussian distribution, then we can actually compute via an analytical formula. We will prove it later.

Lemma 9.2. *Suppose*

$$\begin{bmatrix} x_A \\ x_B \end{bmatrix} \sim N(\begin{bmatrix} \mu_A \\ \mu_B \end{bmatrix}, \begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix})$$

where Σ_{AA} is the covariance matrix of x_A and Σ_{AB} is the correlation between x_A and x_B. Then,

$$x_B|x_A \sim N(\mu_B + \Sigma_{BA}\Sigma_{AA}^{-1}(x_A - \mu_A), \Sigma_{BB} - \Sigma_{BA}\Sigma_{AA}^{-1}\Sigma_{AB}) \tag{9.3}$$

We see that $x_B|x_A$ is a function of x_A and thus the mean and variance depend on x_A. By symmetry, we know that

$$x_A|x_B \sim N(\mu_A + \Sigma_{AB}\Sigma_{BB}^{-1}(x_B - \mu_B), \Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA}) \tag{9.4}$$

9.2.1 Proof using Lemma 9.2

First, we prove the second claim (9.2) assuming the Lemma holds. Fixing $X_1, ..., X_n$, we drop them to directly deal with the outcome $y^*|x^*, S$. Since x^* is fixed globally and S contains x and y, this is equivalent to $y^*|y^{(1)}, ..., y^{(n)}$ ($y^{(i)} \in \mathbb{R}$). Each of them is a Gaussian random variable and their joint distribution is also a Gaussian random variable, which gives us the condition to apply the lemma.

The first Gaussian random variable is $x_A = \overrightarrow{y} = \begin{bmatrix} y^{(1)} \\ ... \\ y^{(n)} \end{bmatrix} \in \mathbb{R}^n$, and the second is a scalar $x_B = y^* \in \mathbb{R}$. It's easier to compute the joint distribution compared to conditional distribution. We first compute the joint distribution

$$\begin{bmatrix} \overrightarrow{y} \\ y^* \end{bmatrix} \sim N(\begin{bmatrix} \mu_A \\ \mu_B \end{bmatrix}, \begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix}).$$

Using the definition of mean and covariance and the fact that θ is the random, we have

$$\mu_A = \mathbb{E}[\overrightarrow{y}] = \mathbb{E}[X\theta + \overrightarrow{\epsilon}] = X\,\mathbb{E}[\theta] + \mathbb{E}[\overrightarrow{\epsilon}] = 0 + 0 = 0.$$

Similarly, we can obtain $\mu_B = \mathbb{E}[y^*] = \mathbb{E}[x^{*T}\theta + \epsilon^*] = 0$. Since $\theta \sim N(0, \tau^2 I)$, the covariance of $\overrightarrow{y}$ is

$$\begin{aligned}
\Sigma_{AA} &= \mathbb{E}[(\overrightarrow{y} - \mu_A)(\overrightarrow{y} - \mu_A)^\top] \\
&= \mathbb{E}[\overrightarrow{y}\,\overrightarrow{y}^\top] \\
&= \mathbb{E}[(X\theta + \overrightarrow{\epsilon})(X\theta + \overrightarrow{\epsilon})^\top] \\
&= \mathbb{E}[X\theta\theta^\top X^\top + X\theta\overrightarrow{\epsilon}^\top + \overrightarrow{\epsilon}\theta^\top Z^\top + \overrightarrow{\epsilon}\,\overrightarrow{\epsilon}^\top] \\
&= X\,\mathbb{E}[\theta\theta^\top]Z^\top + X\,\mathbb{E}[\theta\overrightarrow{\epsilon}^\top] + \mathbb{E}[\overrightarrow{\epsilon}\theta^\top]X^\top + \mathbb{E}[\overrightarrow{\epsilon}\,\overrightarrow{\epsilon}^\top] \\
&= \tau^2 XX^\top + 0 + 0 + \sigma^2 I \\
&= \tau^2 XX^\top + \sigma^2 I.
\end{aligned}$$

Note that $x^{*\top}\theta$ is a scalar and thus it equals to its transpose, we have

$$\begin{aligned}
\Sigma_{AB} &= \mathbb{E}[(\overrightarrow{y} - \mu_A)(y^* - \mu_B)] \\
&= \mathbb{E}[\overrightarrow{y}y^*] \\
&= \mathbb{E}[X\theta(x^{*\top}\theta)] \\
&= \mathbb{E}[X\theta\theta^\top x^*] \\
&= X\,\mathbb{E}[\theta\theta^\top]x^* \\
&= \tau^2 Xx^*,
\end{aligned}$$

and thus

$$\Sigma_{BA} = \Sigma_{AB}^\top = \tau^2 x^* X^\top.$$

Finally we can get

$$\begin{aligned}
\Sigma_{BB} &= \mathbb{E}[(y^* - \mu_B)^2] \\
&= \mathbb{E}[(y^*)^2] \\
&= \mathbb{E}[(x^{*\top}\theta + \Sigma^*)(x^{*\top}\theta + \Sigma^*)] \\
&= \mathbb{E}[x^{*\top}\theta(x^{*\top}\theta) + (\Sigma^*)^2] \\
&= \mathbb{E}[x^{*\top}\theta(x^{*\top}\theta) + (\Sigma^*)^2] \\
&= \mathbb{E}[x^{*\top)\theta\theta^\top x^* + (\Sigma^*)^2}] \\
&= \tau^2 x^{*\top} I x^* + \sigma^2 \\
&= \tau^2 \|x^*\|_2^2 + \sigma^2.
\end{aligned}$$

Putting together and invoking Lemma 9.2,

$$x_B | x_A \sim N(\mu_B + \Sigma_{BA}\Sigma_{AA}^{-1}(x_A - \mu_A), \Sigma_{BB} - \Sigma_{BA}\Sigma_{AA}^{-1}\Sigma_{AB}),$$

it yields that

$$\mu_B + \Sigma_{BA}\Sigma_{AA}^{-1}(x_B - \mu_A) = \tau^2 x^* X^\top (\tau^2 X X^\top + \sigma^2 I)^{-1} x_B.$$

and our goal is then to show

$$\tau^2 x^* X^\top (\tau^2 X X^\top + \sigma^2 I)^{-1} \overrightarrow{y} = \frac{1}{\sigma^2} x^* A^{-1} X^\top \overrightarrow{y}. \tag{9.5}$$

9.2.2 Proof of Eq. (9.5) using singular value decomposition

With $A = \frac{1}{\sigma^2} Z^\top X + \frac{I}{\tau^2}$, we will show that

$$x^* X^\top (\tau^2 X X^\top + \sigma^2)^{-1} = \frac{1}{\tau^2} x^* A^{-1} X^\top,$$

using singular value decomposition (SVD). First, we will have a brief summary of why SVD is useful in our case.

Consider $A \in \mathbb{R}^{n \times m}$ where $n \geq m$. It can be decomposed as $A = U \Sigma V^\top$ where $V \in \mathbb{R}^{m \times m}, U \in \mathbb{R}^{n \times m}, \Sigma \in \mathbb{R}^{m \times m}$. Note that:

- Σ is diagonal.
- U is column-wise orthogonal, meaning that every column of U has norm 1 and orthogonal of each other. This means that $U^\top U = [U^\top][U] = I_{m \times m}$ where every entry ij is the inner product of i-th and j-th column of U.
- V has orthogonal columns. Similar to U, $V^\top V = I$.
- The columns of U are basis of the column span of A; similarly, the rows of $V^\top$ are basis of the row span of A.

What we want to show is that

$$\begin{aligned}&\tau^2 x^* X^\top(\tau^2 XX^\top + \sigma^2 I)^{-1}\overrightarrow{y}\\ =&x^\top X^\top(XX^\top + \frac{\sigma^2}{\tau^2}I)^{-1}\overrightarrow{y}\\ =&\frac{1}{\sigma^2}x^* X^\top(\frac{XX^\top}{\sigma^2} + \frac{1}{\tau^2}I)^{-1}\overrightarrow{y}.\end{aligned}$$

Thus, we'd like to prove

$$X^\top(\frac{ZZ^\top}{\sigma^2} + \frac{I}{\tau^2})^{-1} = (\frac{X^\top X}{\sigma^2} + \frac{I}{\tau^2})^{-1}X^\top.$$

which is not obviously true with X being a matrix. Since SVD allows us to extend U and V to a bigger orthonormal matrices, the claim can be proved as the following. Suppose

$$X = U\Sigma V^\top = U\begin{bmatrix} r_1 & & & & & \\ & \ddots & & & & \\ & & r_d & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0\end{bmatrix}V^\top,$$

where $X \in \mathbb{R}^{n\times d}, n \geq d$. Given that U is orthogonal, we have

$$\begin{aligned}\frac{XX^\top}{\sigma^2} + \frac{I}{\tau^2} &= \frac{U\Sigma V^\top V\Sigma U^\top}{\sigma^2} + \frac{I}{\tau^2} = \frac{U\Sigma^2 U^\top}{\sigma^2} + \frac{I}{\tau^2}\\ &= \frac{1}{\sigma^2}U\begin{bmatrix} r_1^2 & & & & & \\ & \ddots & & & & \\ & & r_d^2 & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0\end{bmatrix}U^\top + \frac{I}{\tau^2}\\ &= U\begin{bmatrix} \frac{r_1^2}{\sigma^2} & & & & & \\ & \ddots & & & & \\ & & \frac{r_d^2}{\sigma^2} & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0\end{bmatrix}U^\top + \frac{UU^\top}{\tau^2}\\ &= U\begin{bmatrix} \frac{r_1^2}{\sigma^2}+\frac{1}{\tau^2} & & & & & \\ & \ddots & & & & \\ & & \frac{r_d^2}{\sigma^2}+\frac{1}{\tau^2} & & & \\ & & & \frac{1}{\tau^2} & & \\ & & & & \ddots & \\ & & & & & \frac{1}{\tau^2}\end{bmatrix}U^\top,\end{aligned}$$

and therefore

$$(\frac{XX^\top}{\sigma^2}+\frac{I}{\tau^2})^{-1}=U\begin{bmatrix} (\frac{r_1^2}{\sigma^2}+\frac{1}{\tau^2})^{-1} & & & & & \\ & \ddots & & & & \\ & & (\frac{r_d^2}{\sigma^2}+\frac{1}{\tau^2})^{-1} & & & \\ & & & \tau^2 & & \\ & & & & \ddots & \\ & & & & & \tau^2 \end{bmatrix}U^\top.$$

Since we have

$$(U\Sigma U^\top)^{-1}=U\Sigma^{-1}U^\top,$$

and

$$U\Sigma U^\top U\Sigma^{-1}U^\top=U\Sigma\Sigma^{-1}U^\top=UU^\top=I, \tag{9.6}$$

we can further show that

$$\begin{aligned} & X^\top(\frac{XX^\top}{\sigma^2}+\frac{I}{\tau^2})^{-1} \\ =& V\Sigma U^\top U\begin{bmatrix} (\frac{r_1^2}{\sigma^2}+\frac{1}{\tau^2})^{-1} & & & & & \\ & \ldots & & & & \\ & & (\frac{r_d^2}{\sigma^2}+\frac{1}{\tau^2})^{-1} & & & \\ & & & \tau^2 & & \\ & & & & \ldots & \\ & & & & & \tau^2 \end{bmatrix}U^\top \\ =& V\Sigma\begin{bmatrix} \frac{r_1}{\frac{r_1^2}{\sigma^2}+\frac{1}{\tau^2}} & & & & & \\ & \ldots & & & & \\ & & \frac{r_d}{\frac{r_d^2}{\sigma^2}+\frac{1}{\tau^2}} & & & \\ & & & \tau^2 & & \\ & & & & \ldots & \\ & & & & & \tau^2 \end{bmatrix}U^\top \\ =& V\begin{bmatrix} \frac{r_1}{\frac{r_1^2}{\sigma^2}+\frac{1}{\tau^2}} & & & & & \\ & \ldots & & & & \\ & & \frac{r_d}{\frac{r_d^2}{\sigma^2}+\frac{1}{\tau^2}} & & & \\ & & & \tau^2 & & \\ & & & & \ldots & \\ & & & & & \tau^2 \end{bmatrix}U^\top. \end{aligned}$$

We expand the RHS $(\frac{X^\top X}{\sigma^2}+\frac{I}{\tau^2})^{-1}X^\top$ in the same way and thus arrive at the same quantity. As we've shown, SVD reduces everything to diagonal matrices. Thus, it's a very useful trick in these problem settings.

Back to our original proof, we can first show that

$$\begin{aligned}\mu_B + \Sigma_{BA}\Sigma_{AA}^{-1}(x_A - \mu_A) &= \frac{1}{\sigma^2} z^* Z^\top (\frac{XX^\top}{\sigma^2} + \frac{I}{\tau^2})^{-1}\overrightarrow{y} \\ &= \frac{1}{\sigma^2} x^* (\frac{XX^\top}{\sigma^2} + \frac{I}{\tau^2})^{-1} X^\top \overrightarrow{y} \\ &= \frac{1}{\sigma^2} x^* A^{-1} X^\top \overrightarrow{y}.\end{aligned} \tag{9.7}$$

Then, for the covariance:

$$\begin{aligned}&\Sigma_{BB} - \Sigma_{BA}\Sigma_{AA}^{-1}\Sigma_{AB} \\ &= \tau^2\|x^*\|_2^2 + \sigma^2 - X^{4\top} X^\top \tau^2 (\tau^2 XX^\top + \sigma^2 I)^{-1} \tau^2 X x^* \\ &= \tau^2\|x^*\|_2^2 + \sigma^2 - \frac{\tau^2}{\sigma^2} x^{*\top} X^\top (\frac{XX^\top}{\sigma^2} + \frac{I}{\tau^2})^{-1} X x^* \\ &= \tau^2\|x^*\|_2^2 + \sigma^2 - \frac{\tau^2}{\sigma^2} x^{*\top} (\frac{X^\top X}{\sigma^2} + \frac{I}{\tau^2})^{-1} X^\top X x^* \\ &= \tau^2\|x^*\|_2^2 + \sigma^2 - \tau^2 x^{*\top} (\frac{X^\top y}{\sigma^2} + \frac{I}{\tau^2})^{-1} (\frac{X^\top X}{\sigma^2} + \frac{I}{\tau^2}) x^* + \tau^2 x^{*\top} (\frac{X^\top X}{\sigma^2} + \frac{I}{\tau^2})^{-1} \frac{I}{\tau^2} x^* \\ &= \tau^2\|x^*\|_2^2 + \sigma^2 - \tau^2 x^{*\top} x^* + \tau^2 x^{*\top} A^{-1} \frac{I}{\tau^2} x^* \\ &= \sigma^2 + x^{*\top} A^{-1} x^*.\end{aligned} \tag{9.8}$$

9.3 Nonparametric Bayesian regression

The setup of nonparametric Bayesian regression is the following. Suppose that we have $S = \{(x^{(i)}, y^{(i)})\}_{i=1}^n$, where high-dimensional data is allowed. We assume that $y^{(i)} = f(x^{(i)}) + \epsilon^{(i)}$ where f can be non-linear and the noise $\epsilon^{(i)} \sim N(0, \sigma^2), \forall i$.

First, we revisit the frequentist approach, which is the kernel method. Assumed that $f(x^{(i)}) = \theta^\top \phi(x^{(i)})$ where $\phi : \mathbb{R}^d \to \mathbb{R}^m$ is a fixed feature map (m could be possibly infinite). As we've discussed in the previous lectures, the computational efficiency here depends on the inner product of the features.

Secondly, if we extend the above to the Bayesian approach, we have a prior on f. Assume that $f \sim P(f)$, a distribution of non-linear functions. Given test example x^*, compute $P(y^*|x^*, S) = \int P(y^*|f, x^*, S)P(f|S)df$ where $P(f|S)$ is the posterior of f given S.

Here, the challenge is to define the prior. There are two approaches. While we'll emphasize the second one, which is more convenient, the two methods are actually equivalent.

9.3.1 Approach 1: frequentist approach

The first approach defines the prior as the following:

$$f \sim P(f) \iff f(x) = \theta^\top \phi(x) = \sum_{i=1}^n \theta_i \phi_i(x),$$

$$\theta_i \overset{\text{iid}}{\sim} N(0, \tau^2 I), \forall i = 1, ..., n.$$

This reduces to a Bayesian linear regression in the feature space. Our input is now $\phi(x)$ rather than x. Moreover, we need to know that $\phi(x)$ needs to be powerful enough such that $\theta^\top \phi(x)$ can cover a large family of functions f (or even all functions).

Then, with $\theta(.) \in \mathbb{R}^\infty$, we need the kernel trick for computational efficiency. Choose ϕ such that $< \phi(x), \phi(z) >= K(x, z)$ is computationally efficient.

The Bayesian linear regression algorithm should be rewritten as only using calls of the kernel function $K(x, z)$. Thus, this method is somewhat complicated.

9.3.2 Approach 2: Gaussian process

The second approach, the Gaussian process, takes a cleaner and more fundamental viewpoint, though conceptually, it requires more work.

As a warm-up, assume that the input space is finite and our goal is to design a prior over functions with finite input space. After this, it takes a slight leap of faith to extend it to the infinite case.

Consider $\mathcal{F} = \{\text{all functions that maps } X \to \mathbb{R}\}$ where $X = \{t_1, ..., t_m\}$. We want to design a prior over $\mathcal{F}$.

To describe the function $f \in$, we only to specify specify the values of the function on a finite number of inputs:

$$\begin{bmatrix} f(t_1) \\ ... \\ f(t_m) \end{bmatrix}.$$

In other word, we can represent f by a vector

$$\overrightarrow{f} = \begin{bmatrix} f(t_1) \\ ... \\ f(t_m) \end{bmatrix} \in \mathbb{R}^m.$$

In this case, designing a prior over the function space is the same as designing a prior over the m-dimensional vector; the latter is much easier.

Consider a Gaussian prior on f (or $\overrightarrow{f}$).

$$\overrightarrow{f} \sim N(\mu, \Sigma), \mu \in \mathbb{R}^m, \Sigma \in \mathbb{R}^{m \times m}.$$

The prior, or the density of the function, is

$$P(f) = P(\overrightarrow{f}) = \frac{1}{(\sqrt{2\pi})^d \det(\Sigma)} y_2 \exp(-\frac{1}{2}(\overrightarrow{f} - \mu)^\top \Sigma^{-1} (\overrightarrow{f} - \mu)).$$

The key takeaway is that the distribution of f is equivalent to the distribution of vector $\overrightarrow{f}$. Then, what if X is infinite ($m = \infty$)? A straightforward extension is the following:

$$\begin{aligned} \mu \in \mathbb{R}^m &\to \mu(\cdot) \\ \Sigma \in \mathbb{R}^{m \times m} &\to k(\cdot, \cdot), \end{aligned}$$

where $\mu(\cdot)$ is a function over $\mathcal{X}$ and $k(\cdot, \cdot)$ is a function over $\mathcal{X} \times \mathcal{X}$.

Now, we're tempted to say that $f \sim N(\mu, k)$. To formalize the method, we define a stochastic process as a collection of random variables $\{f(x) : x \in \mathcal{X}\}$ indexed by elements in $\mathcal{X}$. These random variable have correlation with each other, just like different entries of a vector.

A Gaussian process is a stochastic process such that for any finite number of variables $t_1, ..., t_m \in \mathcal{X}$, $(f(t_1), ..., f(t_m))$ has Gaussian distribution. We will design a prior that has this property. Returning to definition, suppose $\{f(x); x \in \mathcal{X}\}$ is a Gaussian process, then let

$$\mu(x) = \mathbb{E}[f(x)]$$

be the mean function and

$$k(x, z) = \mathbb{E}[(f(x) - \mu(x))(f(z) - \mu(z))]$$

be the covariance function. Formally,

$$f \sim GP(\mu(\cdot), k(\cdot, \cdot)). \tag{9.9}$$

Note that the Gaussian process has the following interesting properties.

1. μ, k uniquely describe a Gaussian process $f \sim GP(\mu, k)$.

   For a Gaussian random vector $\mathcal{W} = \begin{bmatrix} f(x_1) \\ ... \\ f(x_n) \end{bmatrix}$, there is

   $$\mathcal{W} \sim N(\begin{bmatrix} \mu(x_1) \\ ... \\ \mu(x_n) \end{bmatrix}, \big[k(x_i, x_j)\big]_{i,j})$$

2. If a Gaussian process has mean μ and covariance function $k(\cdot, \cdot)$, then $k(\cdot, \cdot)$ is a valid kernel function. In other words, there exists ϕ such that $k(x, z) = \phi(x)^\top \phi(z)$.

   Proof. Suppose that in a Gaussian process, $\forall x_1, ..., x_n, K = \big[k(x_i, x_j)\big]_{i,j=1,...,n})$ is the covariance of $\begin{bmatrix} f(x_1) \\ ... \\ f(x_n) \end{bmatrix}$, $\forall x_1, ..., x_n, K \geq 0$. By Mercer's Theorem, $k(\cdot, \cdot)$ is a valid kernel function. □

3. Vice versa, if $k(\cdot, \cdot)$ is a valid kernel function, then there exists ϕ such that $k(x, z) = \phi(x)^\top \phi(z)$.

For simplicity, assume $\phi(x) \in \mathbb{R}^m$. Let $f(x) = \theta^\top \phi(x)$ where $\theta_i \sim N(0,1)$, then

$$\begin{aligned}
f &\sim GP(0, K), \\
cov(\begin{bmatrix} f(x_1) \\ ... \\ f(x_n) \end{bmatrix})_{ij} &= \mathbb{E}[f(x_i)f(x_j)] \\
&= \mathbb{E}[\theta^\top \phi(x_i)\theta^\top \phi(x_j)] \\
&= \mathbb{E}[\phi(x_i)^\top \theta\theta^\top \phi(x_j)] \\
&= \phi(x_i)^\top \mathbb{E}[\theta\theta^\top]\phi(x_j) \\
&= \phi(x_i)^\top I \phi(x_j) \\
&= \phi(x_i)^\top \phi(x_j) \\
&= k(x_i, x_j).
\end{aligned}$$

Thus, GP is a properly-defined Gaussian process if and only if $K(\cdot,\cdot)$ is a valid kernel function.

What we've been doing is to go from any choice of kernel function $K(\cdot,\cdot)$ to defining $GP(\mu, k)$ to getting the prior $f \sim GP(\mu, k)$. Typically, $\mu(\cdot)$ is chosen to be zero function. $k(\cdot,\cdot)$ can be any common kernel function, where the most popular is the squared exponential kernel or Gaussian kernel.

$$k_{SE}(x, z) = \exp(-\frac{1}{2\tau^2}\|x - z\|_2^2).$$

Qualitatively, what does $f \sim GP(\mu, k)$ look like?

- $f(x), f(z)$ have high correlation is x is close to z because $\exp(-\frac{1}{2\tau^2}\|x - z\|_2^2) \approx \exp(0) \approx 1$.
- $f(x), f(z)$ have low correlation if they are far away because with big $\|x-z\|$, $\exp(-\frac{\|x-z\|_2^2}{2\tau^2}) \approx 0$.
- The parameter τ controls smoothing. Thus, if τ is very big, then even faraway points have strong correlations, meaning that there is strong smoothing. If τ is very small, there is weak smoothing.

To summarize, $GP(\mu(\cdot), k(\cdot,\cdot))$ is the distribution of functions that satisfies the following properties:

1. $f(x)$ is Gaussian, $\forall x$;
2. $(f(x_1), ..., f(x_n))$ is Gaussian;
3. The correlation between $f(x), f(z)$ is $k(x, z)$.

9.3.3 Bayesian prediction

Our next challenge is to compute $y^*|x^*, S$. Following the parametric case, $P(y^*|x^*, S)$ should be defined as $\int P(y^*|x^*, f)P(f|S)df$, which is not really defined when f is a function instead of vectors.

Our plan is to directly compute $P(y^*|x^*, S)$. We will practically test on multiple test points $x^{*(1)}, ..., x^{*(m)}$. We will compute posteriors $y^{*(1)}, ..., y^{*(m)}|x^{*(1)}, ..., x^{*(m)}, S$; more symmetry makes math cleaner.

We will reuse the lemma about the conditional distribution of Gaussian. Recall the lemma: Suppose

$$\begin{bmatrix} x_A \\ x_B \end{bmatrix} \sim N(\begin{bmatrix} \mu_A \\ \mu_B \end{bmatrix}, \begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix}),$$

then

$$x_B|x_A \sim N(\mu_B + \Sigma_{BA}\Sigma_{AA}^{-1}(x_A - \mu_A), \Sigma_{BB} - \Sigma_{BA}\Sigma_{AA}^{-1}\Sigma_{AB}).$$

Consider $x_A \to y^{(1)}, ...y^{(n)}$ to be training observation and $x_B \to y^{*(1)}, ...y^{*(m)}$ to be labels to predict. Then $x_B|x_A = y^{*(1)}, ...y^{*(m)}|y^{(1)}, ...y^{(n)}$ is our target. To apply the lemma, we need the joint distribution of

$$\begin{aligned}
\overrightarrow{f} &= \begin{bmatrix} f(x^{(1)}) \\ ... \\ f(x^{(n)}) \end{bmatrix} \in \mathbb{R}^n, \\
\overrightarrow{f}^* &= \begin{bmatrix} f(x^{*(1)}) \\ ... \\ f(x^{*(m)}) \end{bmatrix} \in \mathbb{R}^m, \\
\overrightarrow{y} &= \begin{bmatrix} f(y^{(1)}) \\ ... \\ f(y^{(n)}) \end{bmatrix} = \overrightarrow{f} + \begin{bmatrix} f(\epsilon^{(1)}) \\ ... \\ f(\epsilon^{(n)}) \end{bmatrix} = \overrightarrow{f} + \overrightarrow{\epsilon} \in \mathbb{R}^n, \\
\overrightarrow{y}^* &= \begin{bmatrix} f(y^{*(1)}) \\ ... \\ f(y^{*(m)}) \end{bmatrix} = \overrightarrow{f}^* + \begin{bmatrix} f(\epsilon^{*(1)}) \\ ... \\ f(\epsilon^{*(m)}) \end{bmatrix} = \overrightarrow{f}^* + \overrightarrow{\epsilon}^* \in \mathbb{R}^m; \\
f &\sim GP, \\
\overrightarrow{f} &\sim N(0, K(X, X)),
\end{aligned}$$

where $K(X, X) = [K(x^{(i)}, x^{(j)})]_{i,j=1,...,n}$. By defining

$$\begin{aligned}
K(X, X^*) &:= [K(X^{(i)}, X^{*(j)})]_{i=1,...,n;j=1,...,m}, \\
K(X^*, X) &:= [K(X^{*(i)}, X^{(j)})]_{i=1,...,m;j=1,...,n}, \\
K(X^*, X^*) &:= [K(X^{*(i)}, X^{*(j)})]_{i,j=1,...,m}; \\
\overrightarrow{f} &\sim N(0, K(X, X)), \\
\overrightarrow{f}^* &\sim N(0, K(X^*, X^*)), \\
\begin{bmatrix} \overrightarrow{f} \\ \overrightarrow{f}^* \end{bmatrix} &\sim N(0, \begin{bmatrix} K(X, X) & K(X, X^*) \\ K(X^*, X) & K(X^*, X^*) \end{bmatrix}); \\
\overrightarrow{y} &= \overrightarrow{f} + \overrightarrow{\Sigma} \sim N(0, K(X, X) + \sigma^2 I), \\
\overrightarrow{y}^* &= \overrightarrow{f}^* + \overrightarrow{\Sigma}^* \sim N(0, K(X^*, X^*) + \sigma^2 I),
\end{aligned}$$

we can have

$$\begin{aligned}
\mathbb{E}[\overrightarrow{y}\,\overrightarrow{y}^{*\top}] &= \mathbb{E}[(\overrightarrow{f} + \overrightarrow{\Sigma})(\overrightarrow{f}^* + \overrightarrow{\Sigma}^*)^\top] \\
&= \mathbb{E}[\overrightarrow{f}\,\overrightarrow{f}^{*\top}] + \mathbb{E}[\overrightarrow{\Sigma}\,\overrightarrow{f}^{*\top}] + \mathbb{E}[\overrightarrow{f}\,\overrightarrow{\Sigma}^{*\top}] + \mathbb{E}[\overrightarrow{\Sigma}\,\overrightarrow{\Sigma}^{*\top}] \\
&= K(X, X^*).
\end{aligned}$$

In summary, we have

$$\begin{bmatrix} \overrightarrow{y} \\ \overrightarrow{y}^* \end{bmatrix} \sim N(0, \begin{bmatrix} K(X,X)+\sigma^2 I & K(X,X^*) \\ K(X^*,X) & K(X^*,X^*)+\sigma^2 I \end{bmatrix}) = N(0, \begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix}).$$

Applying the conditional Gaussian distribution lemma,

$$\begin{aligned} \mu_B &= \Sigma_{BA}\Sigma_{AA}^{-1}(x_A - \mu_A) - \Sigma_{BA}\Sigma_{AA}^{-1}x_A = K(X^*,X)(K(X,X)+\sigma^2 I)^{-1}\overrightarrow{y}, \\ \Sigma_{BB} - \Sigma_{BA}\Sigma_{AA}^{-1}\Sigma_{AB} &= K(X^*,X^*)+\sigma^2 I - K(X^*,X)(K(X,X)+\sigma^2 I)^{-1}K(X,X^*), \end{aligned}$$

and thus

$$\begin{aligned} \overrightarrow{y}^*|\overrightarrow{y} &\sim N(\mu^*,\Sigma^*), \\ \mu^* &= K(X^*,X)(K(X,X)+\sigma^2 I)^{-1}\overrightarrow{y}, \\ \Sigma^* &= K(X^*,X^*)+\sigma^2 I - K(X^*,X)(K(X,X)+\sigma^2 I)^{-1}K(X,X^*). \end{aligned} \tag{9.10}$$

Interestingly, not only do we have the prediction μ^*, but we also have Σ^*, which gives uncertainty quantification. When we have few data points, the confidence interval is large, whereas the confidence interval is smaller when there are multiple data points.