Exercises
Probabilistic Artificial Intelligence
Fall 2023

Institute for Machine Learning
Dept. of Computer Science, ETH Zürich
Prof. Dr. Andreas Krause

Homework #No. 2 (Gaussian Processes)

For questions, please refer to Moodle.
Released on **12/10/2023**

General Instructions

- Submission of solutions is not mandatory but solving the exercises are highly recommended. The master solution will be released after the exercise deadline.
- Part of the exercises are available on Moodle as a quiz. These problems are marked with [✓].

Exercise 1: Gaussian Process Kernels

Given a Gaussian process $GP(\mu, k)$ indexed by $\mathbb{R}$, which is the set of all real numbers, with mean function $\mu : \mathbb{R} \to \mathbb{R}$ and kernel function $k : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$. In the following, we assume that we have a RBF kernel function with lengthscale and variance of 1, i.e. $k(x, x') = e^{-\frac{(x-x')^2}{2}}$, and a linear mean function, i.e. $\mu(x) = x$. We want to model an unknown function $f : \mathcal{X} \to \mathbb{R}$.

(a) [✓]Given three data points $x_1 = 1, x_2 = 3, x_3 = 9$, what are the mean vector $\mathbf{m}$ and covariance matrix K of the marginal distribution $(f(x_1), f(x_2), f(x_3)) \sim \mathcal{N}(\mathbf{m}, K)$?

○ $\mathbf{m} = [1, 3, 9], K = \begin{bmatrix} 1 & e^{-2} & e^{-32} \\ e^{-2} & 1 & e^{-18} \\ e^{-32} & e^{-18} & 1 \end{bmatrix}$ ○ $\mathbf{m} = [0, 0, 0], K = \begin{bmatrix} 1 & e^{-2} & e^{-32} \\ e^{-2} & 1 & e^{-18} \\ e^{-32} & e^{-18} & 1 \end{bmatrix}$

○ $\mathbf{m} = [e^{-1}, e^{-3}, e^{-9}], K = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ ○ $\mathbf{m} = [e^{-1}, e^{-3}, e^{-9}], K = \begin{bmatrix} 0 & 2 & 8 \\ 2 & 0 & 6 \\ 8 & 6 & 0 \end{bmatrix}$

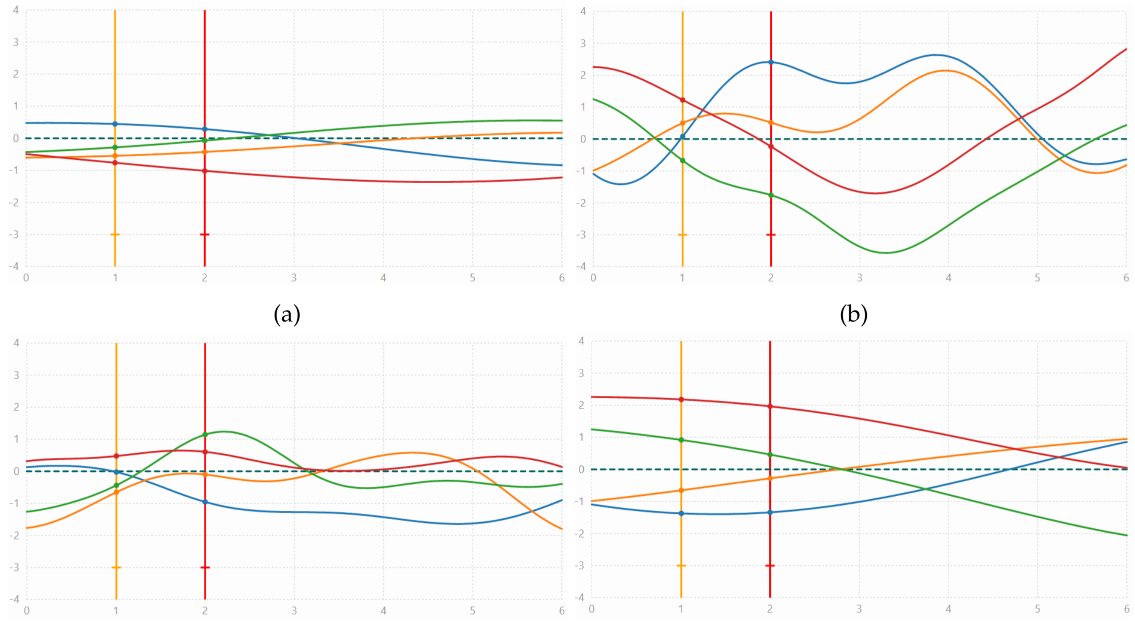
(b) [✓]We are now given noise-free observation for $f(x_1) = 3$ and $f(x_2) = 10$. What is the mean m_3^p and variance σ_3^p of of the posterior distribution $f(x_3)|f(x_1), f(x_2) \sim N(m_3^p, \sigma_3^2)$?

○ $m_3^p = 9 + \frac{2(e^{-32}-e^{-20})-7(e^{-34}-e^{-18})}{1-e^{-4}}, \sigma_3^2 = 1 - e^{-100}\frac{1-e^{-2}}{1-e^{-4}}$
○ $m_3^p = 9, \sigma_3^2 = 1 - e^{-100}\frac{1-e^{-2}}{1-e^{-4}}$
○ $m_3^p = 9, \sigma_3^2 = 1 - \frac{e^{-64}+e^{-36}-2e^{-52}}{1-e^{-4}}$
○ $m_3^p = 9 + \frac{2(e^{-32}-e^{-20})-7(e^{-34}-e^{-18})}{1-e^{-4}}, \sigma_3^p = 1 - \frac{e^{-64}+e^{-36}-2e^{-52}}{1-e^{-4}}$

(c) [✓]We now investigate the influence of kernel parameters on the prior distribution of functions. Imagine that you draw sample functions from Gaussian process with mean 0 and RBF kernel with four pairs of variance and length scale parameters.

$$k(t, t') = \sigma^2 \exp\left(-\frac{(t-t')^2}{2l^2}\right)$$

1. $\sigma^2 = 0.5, l = 1$.
2. $\sigma^2 = 0.5, l = 4$.
3. $\sigma^2 = 2, l = 1$.
4. $\sigma^2 = 2, l = 4$.



(a)

(b)

(c)

(d)

Figure 1: Samples drawn from Gaussian process with mean 0 and RBF kernel with four pairs of parameters. The x-axis is t and the y-axis is X_t. Each plot is generated using a specific pair of variance (σ^2) and length scale (l) parameters. Different colors represent different sample functions drawn from a Gaussian process. Function values at $t = 1$ (i.e., X_1) and $t = 2$ (i.e., X_2) are plotted as colored dots for simpler comparison.

Samples are plotted in Figure 1 in no particular order, each plot for samples drawn from **one** particular pair of parameters.

Applying the results from the previous question, which plot **most likely** corresponds to which pair of parameters?

1. $\sigma^2 = 0.5, l = 1$ ○ Fig.1a ○ Fig.1b ○ Fig.1c ○ Fig.1d
2. $\sigma^2 = 0.5, l = 4$ ○ Fig.1a ○ Fig.1b ○ Fig.1c ○ Fig.1d
3. $\sigma^2 = 2, l = 1$ ○ Fig.1a ○ Fig.1b ○ Fig.1c ○ Fig.1d
4. $\sigma^2 = 2, l = 4$ ○ Fig.1a ○ Fig.1b ○ Fig.1c ○ Fig.1d

Exercise 2: Gaussian Processes Regression With Linear Kernel

[☑]*Gaussian process (GP)*, denoted as $GP(\mu, k)$, is a stochastic process specified by some mean function $\mu : \mathcal{X} \to \mathbb{R}$ and some kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. In this exercise, you will show that Bayesian linear regression yields the same prediction as Gaussian process regression with the linear kernel $k(x, x') = \lambda x^T x'$.

Consider an unknown function $f : \mathcal{X} \to \mathbb{R}$ and a dataset $A = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ of noise-perturbed evaluations $y_i = f(x_i) + \varepsilon_i$, where $\varepsilon_i \sim \mathcal{N}(0, \sigma_n^2)$. Now, our task is to predict the distribution of f in a new point $x^* \in \mathcal{X}$.

GP regression: A priori, we assume $f \sim GP(0, k)$ with linear kernel $k(x, x') = \lambda x^T x'$ (since we want to emulate BLR). In this case, the posterior update for $\mu'(x)$ and $k'(x, x')$ based on the evaluations $y_A = [y_1, \ldots, y_m]^T$ can be computed as follows:

$$\mu'(x) = \mu(x) + k_{x,A}^T (K_{A,A} + \sigma_n^2 I)^{-1} y_A,$$
$$k'(x, x') = k(x, x') - k_{x,A}^T (K_{A,A} + \sigma_n^2 I)^{-1} k_{x,A},$$

where $K_{A,A} \in \mathbb{R}^{m \times m}$ is a matrix with elements $[K_{A,A}]_{i,j} = k(x_i, x_j)$, and $k_{x,A} \in \mathbb{R}^m$ is a vector with elements $[k_{x,A}]_i = k(x, x_i)$.

BLR: In the Bayesian Linear regression, we assume the liner model $f(x_i) = x_i^T w$ with the prior over weights $p(w) = \mathcal{N}(0, \sigma_p^2)$. In the class, we have shown that for the evaluations y_A (or, $y_{1:n}$ in the lecture notation):

$$p(w \mid y_A) = \mathcal{N}(\bar{\mu}, \bar{\Sigma}),$$
$$\bar{\mu} = \frac{1}{\sigma_n^2} \bar{\Sigma} X^T y_A,$$
$$\bar{\Sigma} = \left(\frac{1}{\sigma_n^2} X^T X + \frac{1}{\sigma_p^2} I \right)^{-1},$$

where $X \in \mathbb{R}^{m \times d}$ consists of rows $x_1^T, x_2^T, \ldots, x_m^T$.

> Given a new point x^*, find the predictive distribution both for BLR and GP regression. For which value λ do they coincide?
>
> *Hint:* One of the Woodbury identities might be useful.
>
> ○ $\lambda = \frac{1}{\sigma_n^2} + \frac{1}{\sigma_p^2}$ ○ $\lambda = \sigma_p^2$ ○ $\lambda = \sigma_n^2$ ○ $\lambda = \frac{1}{\sigma_p^2}$

Exercise 3: Kalman filters

Consider the following dynamical system describing a moving particle:

$$X_{t+1} = X_t + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0, \sigma_x^2),$$

where $X_t \in \mathbb{R}$ denotes a position of the particle at time $t > 0$ and ε_t is a time-independent and identically (i.i.d) distributed Gaussian noise. We would like to track the position of the particle over time, however, this is not easy. We cannot directly observe X_t but only see its measurement Y_t perturbed by i.i.d Gaussian noise η_t:

$$Y_t = X_t + \eta_t \qquad \eta_t \sim \mathcal{N}(0, \sigma_y^2).$$

To study this process, we divide the problem in prediction and conditioning.

(a) [✓]*Prediction*: assume the posterior distribution of X_t is a Gaussian with mean μ_t and variance σ_t^2 after having observed $y_{1:t} := \{Y_1 = y_1, \ldots, Y_t = y_t\}$:

$$p(X_t \mid y_{1:t}) = \mathcal{N}(\mu_t, \sigma_t^2).$$

What is the predictive distribution for the particle's position at the next time step $p(X_{t+1} \mid y_{1:t})$:

○ $\mathcal{N}(0, \sigma_t^2 + t\sigma_y^2)$ ○ $\mathcal{N}(\mu_t, \sigma_y^2 + \sigma_x^2)$ ○ $\mathcal{N}(\mu_t, \sigma_t^2 + \sigma_x^2)$ ○ $\mathcal{N}(\mu_t, \sigma_t^2 + \sigma_y^2)$

(b) [✓]*Conditioning*: once we have the prediction distribution $p(X_{t+1} \mid y_{1:t})$, we would like to understand how acquiring y_{t+1} would help. For what value k_{t+1}, called the Kalman gain, can we write $p(X_{t+1} \mid y_{1:t}, y_{t+1})$ as follows:

$$\begin{aligned} p(X_{t+1} \mid y_{1:t}, y_{t+1}) &= \mathcal{N}(\mu_{t+1}, \sigma_{t+1}^2), \\ \text{with} \qquad \mu_{t+1} &= \mu_t + k_{t+1}(y_{t+1} - \mu_t) \\ \text{and} \qquad \sigma_{t+1}^2 &= (1 - k_{t+1})(\sigma_x^2 + \sigma_t^2) \end{aligned}$$

○ $k_{t+1} = \frac{\sigma_x^2}{\sigma_x^2+\sigma_y^2}$ ○ $k_{t+1} = \frac{\sigma_x^2+\sigma_t^2}{\sigma_x^2+\sigma_t^2+\sigma_y^2}$ ○ $k_{t+1} = \frac{\sigma_x^2+\sigma_t^2+\sigma_y^2}{\sigma_x^2+\sigma_t^2}$ ○ $k_{t+1} = \frac{\sigma_x^2+\sigma_t^2}{\sigma_y^2}$

We now would like to discuss the importance of the Kalman gain k_t. This value plays an important role in the interpretation of the conditioning step.

(c) How the different parameters σ_x, σ_y and σ_t influence the value of the Kalman gain?

(d) How the value of the Kalman gain k_{t+1} gives more importance to y_{t+1} or, gives more importance to the previous knowledge μ_{t+1}? Can you guess why it's called "gain"?

(e) [✓]Now we will show that the Kalman filter can be seen as a GP. To this end, we define:

$$f : \mathbb{Z}_+ \to \mathbb{R} \quad \text{such that } f(t) = X_t.$$

Assuming that $X_0 \sim \mathcal{N}(0, \sigma_0^2)$ and $X_{t+1} = X_t + \varepsilon_t$, with $\varepsilon_t \sim \mathcal{N}(0, \sigma_x^2)$, we can show that $f \sim \mathcal{GP}(0, k_{KF})$ with which kernel function k_{KF}

○ $k_{KF}(t, t') = \sigma_x^2 + \sigma_0^2 \max\{t, t'\}$ ○ $k_{KF}(t, t') = \sigma_x^2 + \sigma_0^2 \min\{t, t'\}$
○ $k_{KF}(t, t') = \sigma_0^2 + \sigma_x^2 \min\{t, t'\}$ ○ $k_{KF}(t, t') = \sigma_0^2 + \sigma_x^2 \max\{t, t'\}$

Exercise 4: Hyperparameters Selection and Marginal Likelihood

Consider an unknown function $f : \mathcal{X} \to \mathbb{R}$ and a dataset $A = \{X, \mathbf{y}\} = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ of noise-perturbed evaluations $y_i = f(x_i) + \varepsilon_i$, where $\varepsilon_i \sim \mathcal{N}(0, \sigma_n^2)$, we make the hypothesis that $f \sim GP(0, k_\theta)$, with zero mean function and covariance function $k_\theta : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. We are interested into selecting hyperparameters θ by maximizing the marginal likelihood $p(\mathbf{y} \mid X, \theta)$.

(a) [✓]We define $\mathbf{K}_{\mathbf{y},\theta} = \mathbf{K}_{f,\theta} + \sigma_n^2 I$ as the covariance matrix of $\mathbf{y}$ for covariance function k_θ. We also define $\alpha = \mathbf{K}_{\mathbf{y},\theta}^{-1}\mathbf{y}$. What is the value of the the marginal likelihood gradient with respect to θ_j, $\frac{\partial}{\partial \theta_j} \log p(\mathbf{y} \mid X, \theta)$:

○ $\frac{1}{2} \operatorname{tr}\left(\left(\alpha\alpha^T - \mathbf{K}_{\mathbf{f},\theta}^{-1}\right) \frac{\partial \mathbf{K}_{\mathbf{f},\theta}}{\partial \theta_j}\right)$ ○ $\frac{1}{2} \operatorname{tr}\left(\left(\alpha\alpha^T - \mathbf{K}_{\mathbf{y},\theta}^{-1}\right) \frac{\partial \mathbf{K}_{\mathbf{y},\theta}}{\partial \theta_j}\right)$
○ $2 \operatorname{tr}\left(\left(\alpha\alpha^T - \mathbf{K}_{\mathbf{y},\theta}^{-1}\right) \frac{\partial \mathbf{K}_{\mathbf{y},\theta}}{\partial \theta_j}\right)$ ○ $2 \operatorname{tr}\left(\left(\alpha\alpha^T - \mathbf{K}_{\mathbf{f},\theta}^{-1}\right) \frac{\partial \mathbf{K}_{\mathbf{y},\theta}}{\partial \theta_j}\right)$

Hint: You can use the following indentites in your derivation:

- For any invertible matrix M, you have:

$$\frac{\partial}{\partial \theta_j} M^{-1} = -M^{-1} \frac{\partial M}{\partial \theta_j} M^{-1}$$

- For any symmetric positive definite matrix S, you have:

$$\frac{\partial}{\partial\theta_j}\log|S| = \operatorname{tr}\left(S^{-1}\frac{\partial S}{\partial\theta_j}\right)$$

(b) [✓]We now assume that covariance function for the noisy targets(i.e. including noise contribution) can be written $k_{\mathbf{y}}(x,x') = \theta_0\tilde{k}(x,x')$ with $\tilde{k}$ a valid kernel independent of θ_0. What is the closed-form solution θ_0^* to the equation $\frac{\partial}{\partial\theta_0}\log p(\mathbf{y}\mid X,\theta) = 0$:

○ $\frac{1}{m}\mathbf{y}^\top\mathbf{K}_{\mathbf{y}}^{-1}\mathbf{y}$ ○ $\frac{1}{m}\operatorname{tr}(\tilde{\mathbf{K}}^{-1})$ ○ $\frac{1}{m}\mathbf{y}^\top\tilde{\mathbf{K}}^{-1}\mathbf{y}$ ○ $\mathbf{y}^\top\mathbf{K}_{\mathbf{y}}^{-1}\mathbf{y}$

(c) [✓]In the noise-free case, given the covariance function defined as $k_{\theta_0}(x,x') = \theta_0\tilde{k}(x,x')$ with $\tilde{k}$ a valid kernel independent of θ_0 , how should we scale optimal parameter θ_0^* if we scale labels $\mathbf{y}$ by a scalar s:
○ s^2
○ θ_0^* stays identical because it is independent of s
○ s
○ It depends of the choice of kernel function $\tilde{k}$

Exercise 5: Equivalence between a subset of regressor and Gaussian process regression

Consider an unknown function $f : \mathcal{X} \to \mathbb{R}$. A priori we assume $f \sim GP(0,k)$. Using Gaussian process regression, after a set of $\mathbf{y}$ samples, the predictive mean $\mathbb{E}[f(x')]$ and variance $\mathbb{V}[f(x')]$ at x' are:

$$\mathbb{E}[f(x')] = \mathbf{k}(x')^\top\left(K+\sigma_n^2 I\right)^{-1}\mathbf{y} \tag{1}$$

$$\mathbb{V}[f(x')] = k(x',x') - \mathbf{k}(x')^\top(K+\sigma_n^2 I)^{-1}\mathbf{k}(x') \tag{2}$$

Unfortunately, the Gaussian process predictions scales typically $\mathcal{O}(n^3)$ due to matrix inversion. While using only $m < n$ subset of regressors helps to reduce the time complexity to $\mathcal{O}(m^2 n)$ with approximated predictive mean $\mathbb{E}[\tilde{f}(x')]$ and variance $\mathbb{V}[\tilde{f}(x')]$ at x' given by,

$$\mathbb{E}[\tilde{f}(x')] = \mathbf{k}_m(x')^\top\left(K_{mn}K_{nm}+\sigma_n^2 K_{mm}\right)^{-1}K_{mn}\mathbf{y} \tag{3}$$

$$\mathbb{V}[\tilde{f}(x')] = \sigma_n^2\mathbf{k}_m(x')^\top\left(K_{mn}K_{nm}+\sigma_n^2 K_{mm}\right)^{-1}\mathbf{k}_m(x') \tag{4}$$

Show that subset of regressors predictors for the mean and variance is equivalent to the full Gaussian process regression predictors while using Nyström approximate kernel function $\tilde{k}(x,x') = \mathbf{k}_m^\top(x)K_{mm}^{-1}\mathbf{k}_m(x')$.

Hints: In order to show equivalence, one can write predictive mean $\mathbb{E}[\tilde{f}(x')]$ and variance $\mathbb{V}[\tilde{f}(x')]$ for the subset of regressors as predictors of Gaussian process regression (1, 2) but with approximated kernel function $\tilde{k}(x,x') = \mathbf{k}_m^\top(x)K_{mm}^{-1}\mathbf{k}_m(x')$ instead of $k(x,x')$.

Using the approximated kernel function $\tilde{k}(x,x')$for each pair of data points in the training set, we get, $\tilde{\mathbf{k}}(x') = K_{nm}K_{mm}^{-1}\mathbf{k}_m(x')$ and $\tilde{K} = K_{nm}K_{mm}^{-1}K_{mn}$.

For a $n \times m$ matrix, Q, you can use matrix inversion lemma, $(\sigma^2 I_n + QQ^\top)^{-1} = \sigma^{-2}I_n - \sigma^{-2}Q(\sigma^2 I_m + Q^\top Q)^{-1}Q^\top$, to transform the inversion of an $n \times n$ to inversion of a $m \times m$ matrix.