PML Theory:

Sam Laing 6283670
Albert Catalan 6443478

a) we have

$$N(x;a,A)\,N(x;b,B)$$

$$= \frac{1}{(2\pi)^{D/2}|A|^{1/2}} \cdot \frac{1}{(2\pi)^{D/2}|B|^{1/2}} \exp\left[-\frac{1}{2}\left[(x-a)^T A^{-1}(x-a) + (x-b)^T B^{-1}(x-b)\right]\right]$$

$$\propto \exp\left[-\frac{1}{2}\left[x^T(A^{-1}+B^{-1})x - 2(A^{-1}a+B^{-1}b)^T x + a^T A^{-1}a + b^T B^{-1} b\right]\right]$$

$$= \exp\left[-\frac{1}{2}\left[(x-(A^{-1}+B^{-1})^{-1}(A^{-1}a+B^{-1}b))^T(A^{-1}+B^{-1})(x-(A^{-1}+B^{-1})^{-1}(A^{-1}a+B^{-1}b)) - (A^{-1}a+B^{-1}b)^T(A^{-1}+B^{-1})^{-1}(A^{-1}a+B^{-1}b) + a^T A^{-1} a + b^T B^{-1} b\right]\right]$$

completing the square

$$\propto \mathcal{N}(x;\underbrace{(A^{-1}+B^{-1})^{-1}(A^{-1}a+B^{-1}b)}_{=\mu_*},\underbrace{(A^{-1}+B^{-1})^{-1}}_{=\Sigma_*}) \cdot \exp\left[-\frac{1}{2}\left[a^T A^{-1} a + b^T B^{-1} b - (a^T A^{-1}+b^T B^{-1})(A^{-1}+B^{-1})^{-1}(A^{-1}a+B^{-1}b)\right]\right]$$

$$= \mathcal{N}(x;\mu_*,\Sigma_*)\exp\left(-\frac{1}{2}\left[a^T A^{-1} a + b^T B^{-1} b \cancel{+(a^T A^{-1}+b^T B^{-1})} - a^T A^{-1}(A^{-1}+B^{-1})^{-1}A^{-1}a - b^T B^{-1}(A^{-1}+B^{-1})^{-1}B^{-1}b - 2b^T B^{-1}(A^{-1}+B^{-1})^{-1}A^{-1}a\right]\right)$$

$$\propto \mathcal{N}(x;\mu_*,\Sigma_*)\exp\left(-\frac{1}{2}\left[a^T(A^{-1}-A^{-1}(A^{-1}+B^{-1})^{-1}A^{-1})a - b^T(B^{-1}-B^{-1}(A^{-1}+B^{-1})B^{-1})b\right]\right)$$

(where other terms are absorbed by proportionality)

$$\propto \mathcal{N}(x;\mu_*,\Sigma_*)\exp\left[-\frac{1}{2}(a-b)^T(A+B)^{-1}(a-b)\right]$$

woodbury identity

$$\propto N(x;(A^{-1}+B^{-1})^{-1}(A^{-1}a+B^{-1}b),(A^{-1}+B^{-1})^{-1}) \cdot N(a,b,A+B)$$

$\square$

b) for a matrix $\Sigma = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix}$, ~~where~~ we have

$$\Sigma^{-1} = \Sigma^{*} = \begin{pmatrix} \Sigma^{*}_{xx} & \Sigma^{*}_{xy} \\ \Sigma^{*}_{yx} & \Sigma^{*}_{yy} \end{pmatrix}$$

where $\Sigma^{*}_{xx} = A^{-1}$ $\quad \Sigma^{*}_{xy} = -A^{-1}\Sigma_{xy}\Sigma_{yy}^{-1}$

$\Sigma^{*}_{yx} = -\Sigma_{yy}^{-1}\Sigma_{yx}A^{-1}$ $\quad \Sigma^{*}_{yy} = \Sigma_{yy}^{-1} + \Sigma_{yy}^{-1}\Sigma_{yx}A^{-1}\Sigma_{xy}\Sigma_{yy}^{-1}$

with $A := \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}$

(by Schur's identity)

Now let $z^T := (x^T, y^T)$ & $\mu^T = (\mu_x^T, \mu_y^T)$

we have:

$$(z-\mu)^T \Sigma^{-1} (z-\mu) = (x^T-\mu_x^T, y^T-\mu_y^T) \begin{pmatrix} \Sigma^{*}_{xx} & \Sigma^{*}_{xy} \\ \Sigma^{*}_{yx} & \Sigma^{*}_{yy} \end{pmatrix} \begin{pmatrix} x-\mu_x \\ y-\mu_y \end{pmatrix}$$

$$= (x-\mu_x)^T \Sigma^{*}_{xx} (x-\mu_x) + (x-\mu_x)^T \Sigma^{*}_{xy} (y-\mu_y)$$

$$+ (y-\mu_y)^T \Sigma^{*}_{yx} (x-\mu_x) + (y-\mu_y)^T \Sigma^{*}_{yy} (y-\mu_y)$$

$$= (x-\mu_x)^T A^{-1} (x-\mu_x) - (x-\mu_x)^T (A^{-1}\Sigma_{xy}\Sigma_{yy}^{-1})(y-\mu_y)$$

$$- (y-\mu_y)^T \Sigma_{yy}^{-1}\Sigma_{yx}A^{-1}(x-\mu_x)$$

$$+ (y-\mu_y)^T \left[\Sigma_{yy}^{-1} + \Sigma_{yy}^{-1}\Sigma_{yx}A^{-1}\Sigma_{xy}\Sigma_{yy}^{-1}\right](y-\mu_y)$$

$$= (y-\mu_y)^T \Sigma_{yy}^{-1}(y-\mu_y)$$

$$+ \left(x-\left(\mu_x+\Sigma_{xy}\Sigma_{yy}^{-1}(y-\mu_y)\right)\right)^T A^{-1}\left(x-\left(\mu_x+\Sigma_{xy}\Sigma_{yy}^{-1}(y-\mu_y)\right)\right)$$

now we have

$$p(x|y) = \frac{p(x,y)}{p(y)} \qquad \| \; z^T = (x^T, y^T)$$

$$\propto \exp\left(-\frac{1}{2}\left[(z-\mu)^T\Sigma^{-1}(z-\mu) - (y-\mu_y)^T\Sigma_{yy}^{-1}(y-\mu_y)\right]\right)$$

$$= \exp\left[-\frac{1}{2}\left[\left(x-\left(\mu_x+\Sigma_{xy}\Sigma_{yy}^{-1}(y-\mu_y)\right)\right)^T A^{-1}\left(x-\left(\mu_x+\Sigma_{xy}\Sigma_{yy}^{-1}(y-\mu_y)\right)\right)\right]\right]$$

using derived expression

which we recognize as the functional form of a $\mathcal{N}\left(\mu_x+\Sigma_{xy}\Sigma_{yy}^{-1}(y-\mu_y),\ \Sigma_{xx}-\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}\right)$

~~c) Consider a multivariate normal~~ $\mathcal{N}(z; \mu_z, \Sigma_z)$ ~~where~~

$$\mu_z^T = (\mu_1, \mu_2, \mu^T) \qquad \Sigma_z = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} & \Sigma_{XY} \\ \Sigma_{21} & \Sigma_{22} & \\ \Sigma_{YX} & & \Sigma_{YY} \end{pmatrix}$$

~~where~~ $\mu \in \mathbb{R}^{n-2}$, $\Sigma_{YY} \in \mathbb{R}^{(n-2)\times(n-2)}$

$\mu_z \in \mathbb{R}^n$ $\quad \Sigma_z \in \mathbb{R}^{n\times n}$ $\quad$ ~~(since wlog,~~ $(i,j)=(1,2)$~~)~~

c) •) WLOG suppose $z = \begin{pmatrix} x_1 \\ x_2 \\ y \end{pmatrix}$ $\quad \Sigma = \left(\begin{array}{cc|c} \sigma_{11} & \sigma_{12} & \Sigma_{xy} \\ \sigma_{21} & \sigma_{22} & \\ \hline \Sigma_{yx} & & \Sigma_{yy} \end{array}\right)$

$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_y \end{pmatrix}$

then $p(x_1, x_2) = \mathcal{N}\left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} ; \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \cancel{\Sigma}, \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} \right)$ (*)

We have (using the result from (b))

$$p(x_1 | x_2) = \mathcal{N}\left(x_1 ; \mu_1 + \sigma_{12} \frac{1}{\sigma_{22}} (x_2 - \mu_2), \; \sigma_{11} - \cancel{\Sigma}\, \sigma_{12} \frac{1}{\sigma_{22}} \sigma_{21} \right)$$

$$= \mathcal{N}(x ; \mu_1, \sigma_{11}) = p(x_1)$$

if $\sigma_{12} = 0$

■

(*) The reason for this is that

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & 0 \cdots & 0 \end{pmatrix}}_{\text{“}A\text{''}} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

$\cancel{\Rightarrow p(x_1, x_2) = A\mu = \mu_1}$

$\Rightarrow A\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$ $\quad \cancel{A \Sigma A^T = \begin{pmatrix} 1 & 0 & 0 & 0 \cdots 0 \\ 0 & 1 & 0 & \cdots 0 \end{pmatrix} \begin{pmatrix} \sigma_{11} & & \\ & \ddots & \\ & & \sigma_{nn} \end{pmatrix}}$

& $A \Sigma A^{-1} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}$ (with A clearly linear)

$\Rightarrow (X_1, X_2) \sim \mathcal{N}\left((\mu_1, \mu_2), \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} \right)$

(linear / affine transformations of normal are normal)

•) As in part 1, we can linearly project to see that

$$p(x_1, x_2, x_3) = \mathcal{N}\left(\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}, \underbrace{\begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix}}_{\mu}, \underbrace{\left(\begin{array}{cc|c} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \hline \sigma_{31} & \sigma_{32} & \sigma_{33} \end{array}\right)}_{\Sigma}\right)$$

(where wlog $x_k = x_3$) (Using matrix $A = \begin{pmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \end{pmatrix}$)

Since Σ is symmetric, so is Σ^{-1}.
The ~~is~~ ij^{th} entry of Σ^{-1} is the ij^{th} cofactor divided by the determinant

therefore, if $(\Sigma^{-1})_{12} = 0$, $(\Sigma^{-1})_{21} = 0$

then $\begin{vmatrix} \sigma_{21} & \sigma_{23} \\ \sigma_{31} & \sigma_{33} \end{vmatrix} = \sigma_{21}\sigma_{33} - \sigma_{31}\sigma_{23} = 0$

$$\Rightarrow \sigma_{21} = \frac{\sigma_{31}\sigma_{23}}{\sigma_{33}} \quad (*)$$

and $\begin{vmatrix} \sigma_{12} & \sigma_{13} \\ \sigma_{31} & \sigma_{33} \end{vmatrix} = \sigma_{12}\sigma_{33} - \sigma_{13}\sigma_{31} = 0$

$$\Rightarrow \sigma_{12} = \frac{\sigma_{13}\sigma_{31}}{\sigma_{33}} \quad (*)$$

then, using part (b),

$$p(x_1, x_2 \mid x_3) = \mathcal{N}\left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}; \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \begin{pmatrix} \sigma_{13} \\ \sigma_{23} \end{pmatrix} \frac{1}{\sigma_{33}}(x_3 - \mu_3), \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} - \frac{1}{\sigma_{33}} \begin{pmatrix} \sigma_{13}\sigma_{31} & \sigma_{13}\sigma_{31} \\ \sigma_{23}\sigma_{31} & \sigma_{23}\sigma_{32} \end{pmatrix}\right)$$

$$\cancel{= \mathcal{N}\left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \begin{pmatrix} \mu_1 + \frac{\sigma_{13}}{\sigma_{33}}(x_3 - \mu_3) \\ \mu_2 + \frac{\sigma_{23}}{\sigma_{33}}(x_3 - \mu_3) \end{pmatrix}, \left(\sigma_{11} - \frac{\sigma_{13}\sigma_{31}}{\sigma_{33}}\right.\right.}$$

$$= N\left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}; \begin{pmatrix} \mu_1 + \frac{\sigma_{13}}{\sigma_{33}}(x_3-\mu_3) \\ \mu_2 + \frac{\sigma_{23}}{\sigma_{33}}(x_3-\mu_3) \end{pmatrix}, \begin{pmatrix} \sigma_{11} - \frac{\sigma_{13}\sigma_{31}}{\sigma_{33}} & \sigma_{12} - \frac{\sigma_{13}\sigma_{32}}{\sigma_{33}} \\ \sigma_{21} - \frac{\sigma_{23}\sigma_{31}}{\sigma_{33}} & \sigma_{22} - \frac{\sigma_{23}\sigma_{32}}{\sigma_{33}} \end{pmatrix}\right)$$

but using our identities $(*)$,
we see that the diagonals of the covariance matrix must then be 0

$$\Rightarrow p(x_1, x_2 | x_3) = N\left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}; \begin{pmatrix} \mu_1 + \frac{\sigma_{13}}{\sigma_{33}}(x_3-\mu_3) \\ \mu_2 + \frac{\sigma_{23}}{\sigma_{33}}(x_3-\mu_3) \end{pmatrix}, \begin{pmatrix} \sigma_{11} - \frac{\sigma_{13}\sigma_{31}}{\sigma_{33}} & 0 \\ 0 & \sigma_{11} - \frac{\sigma_{23}\sigma_{32}}{\sigma_{33}} \end{pmatrix}\right)$$

but then we see that (reading the matrix)

$$p(x_1 | x_3) = N\left(x_1 ; \mu_1 + \frac{\sigma_{13}}{\sigma_{33}}(x_3-\mu_3), \sigma_{11} - \frac{\sigma_{13}\sigma_{31}}{\sigma_{33}}\right)$$

$$p(x_2 | x_3) = N\left(x_2 ; \mu_2 + \frac{\sigma_{23}}{\sigma_{33}}(x_3-\mu_3), \sigma_{22} - \frac{\sigma_{23}\sigma_{32}}{\sigma_{33}}\right)$$

which are precisely the formulae ~~of~~ ~~the~~ ~~pr~~ from part (b) when applied to $\begin{pmatrix} x_1 \\ x_3 \end{pmatrix}$ & $\begin{pmatrix} x_2 \\ x_3 \end{pmatrix}$

~~for~~

We therefore conclude that x_1 & x_2 are conditionally independent given x_k for any $k \in \{3, \dots, n\}$ $\blacksquare$

Probabilistic Machine Learning

University of Tübingen, Summer Term 2023 © 2023 P. Hennig

Exercise Sheet No. 5 — Multivariate and Multi-Output Gaussian Processes

---

Submission by:

- Sam Laing, 6283670
- Albert Catalan Tatjer, 6443478

In [1]:
```
import functools

import jax
import jax.numpy as jnp
import numpy as np


jax.config.update("jax_enable_x64", True)
```

In [3]:
```
from matplotlib import pyplot as plt
from tueplots import bundles
from tueplots.constants.color import rgb


plt.rcParams.update(bundles.beamer_moml())
plt.rcParams.update({'figure.dpi': 200})
```

Exercise 5.2 (Coding Exercise)

In this exercise, we will use Gaussian processes to learn multivariate and vector-valued functions.

In [4]:
```
from collections.abc import Callable
import dataclasses


@dataclasses.dataclass
class Gaussian:
    # Gaussian distribution with mean mu and covariance Sigma
    mu: jnp.ndarray  # shape (D,)
    Sigma: jnp.ndarray  # shape (D,D)

    @functools.cached_property
    def L_factor(self):
        """Cholesky factorization of the covariance matrix
        (for use in jax.scipy.linalg.cho_solve)"""
        return jax.scipy.linalg.cho_factor(self.Sigma, lower=True)
```

```
    def __rmatmul__(self, A):
        """Linear maps of Gaussian RVs are Gaussian RVs
        A: linear map, shape (N,D)
        """
        return Gaussian(mu=A @ self.mu, Sigma=A @ self.Sigma @ A.T)

    @functools.singledispatchmethod
    def __add__(self, other):
        """Affine maps of Gaussian RVs are Gaussian RVs
        shift of a Gaussian RV by a constant.
        We implement this as a singledispatchmethod, because jnp.ndarrays
        and register the addition of two RVs below
        """
        other = jnp.asarray(other)
        return Gaussian(mu=self.mu + other, Sigma=self.Sigma)

    def condition(self, A, y, Lambda):
        """Linear conditionals of Gaussian RVs are Gaussian RVs
        Conditioning of a Gaussian RV on a linear observation
        A: observation matrix, shape (N,D)
        y: observation, shape (N,)
        Lambda: observation noise covariance, shape (N,N)
        """
        Gram = A @ self.Sigma @ A.T + Lambda
        L = jax.scipy.linalg.cho_factor(Gram, lower=True)
        mu = self.mu + self.Sigma @ A.T @ jax.scipy.linalg.cho_solve(L, y
        Sigma = self.Sigma - self.Sigma @ A.T @ jax.scipy.linalg.cho_solv
            L, A @ self.Sigma
        )
        return Gaussian(mu=mu, Sigma=Sigma)

    @functools.cached_property
    def std(self):
        # standard deviation
        return jnp.sqrt(jnp.diag(self.Sigma))

    def sample(self, key, num_samples=1):
        """
         sample from the Gaussian
        # alternative implementation: works because the @ operator contra
        # return (self.L @ jax.random.normal(key, shape=(num_samples, sel
        # or like this, more explicit, but not as easy to read
        # return jnp.einsum("ij,kj->ki", self.L, jax.random.normal(key, s
        # or the scipy version:
        """
        return jax.random.multivariate_normal(
            key, mean=self.mu, cov=self.Sigma, shape=(num_samples,), meth
        )


@Gaussian.__add__.register
def _add_gaussians(self, other: Gaussian):
    # sum of two Gaussian RVs
    return Gaussian(mu=self.mu + other.mu, Sigma=self.Sigma + other.Sigma


@dataclasses.dataclass
class GaussianProcess:
    # mean function
    m: Callable[[jnp.ndarray], jnp.ndarray]
```

```
    # covariance function
    k: Callable[[jnp.ndarray, jnp.ndarray], jnp.ndarray]

    def __call__(self, x):
        return Gaussian(mu=self.m(x), Sigma=self.k(x[:, None, :], x[None,

    def condition(self, y, X, sigma):
        return ConditionalGaussianProcess(
            self, y, X, Gaussian(mu=jnp.zeros_like(y), Sigma=sigma * jnp.
        )

    def plot(
        self,
        ax,
        x,
        color=rgb.tue_gray,
        mean_kwargs={},
        std_kwargs={},
        num_samples=0,
        rng_key=None,
    ):
        gp_x = self(x)
        ax.plot(x[:, 0], gp_x.mu, color=color, **mean_kwargs)
        ax.fill_between(
            x[:, 0],
            gp_x.mu - 2 * gp_x.std,
            gp_x.mu + 2 * gp_x.std,
            color=color,
            **std_kwargs
        )
        if num_samples > 0:
            ax.plot(
                x[:, 0],
                gp_x.sample(rng_key, num_samples=num_samples).T,
                color=color,
                alpha=0.2,
            )


class ConditionalGaussianProcess(GaussianProcess):
    """
    A Gaussian process conditioned on data.
    Implented as a proper python class, which allows inheritance from the
    A conditional Gaussian process contains a Gaussian process prior, pro
    """

    def __init__(self, prior, y, X, epsilon: Gaussian):
        self.prior = prior
        self.y = jnp.atleast_1d(y)  # shape: (n_samples,)
        self.X = jnp.atleast_2d(X)  # shape: (n_samples, n_inputs)
        self.epsilon = epsilon
        # initialize the super class
        super().__init__(self._mean, self._covariance)

    @functools.cached_property
    def predictive_covariance(self):
        return self.prior.k(self.X[:, None, :], self.X[None, :, :]) + sel

    @functools.cached_property
    def predictive_covariance_cho(self):
```

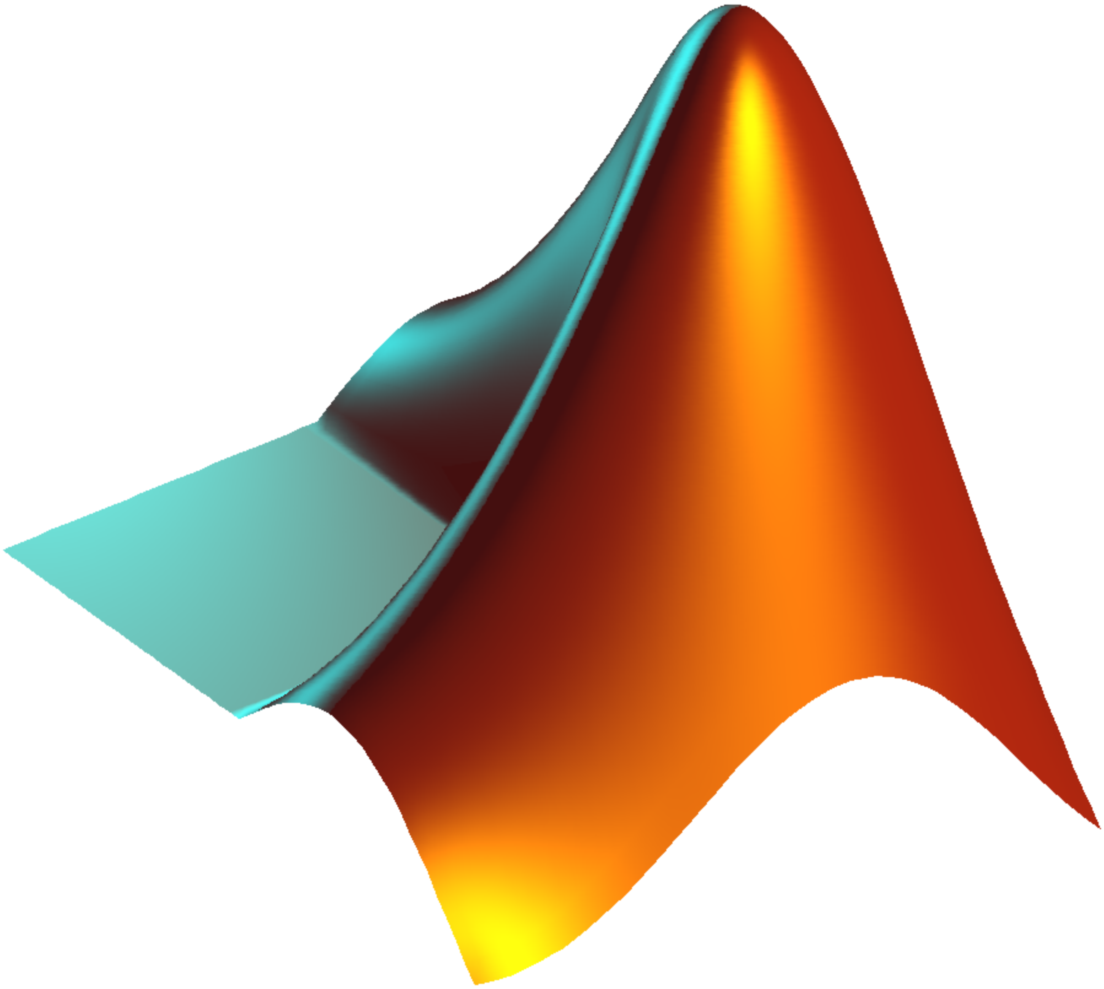
```
        return jax.scipy.linalg.cho_factor(self.predictive_covariance)

    @functools.cached_property
    def representer_weights(self):
        return jax.scipy.linalg.cho_solve(
            self.predictive_covariance_cho,
            self.y - self.prior(self.X).mu - self.epsilon.mu,
        )

    def _mean(self, x):
        x = jnp.asarray(x)
        return (
            self.prior(x).mu
            + self.prior.k(x[..., None, :], self.X[None, :, :])
            @ self.representer_weights
        )

    @functools.partial(jnp.vectorize, signature="(d),(d)->()", excluded={
    def _covariance(self, a, b):
        return self.prior.k(a, b) - self.prior.k(
            a, self.X
        ) @ jax.scipy.linalg.cho_solve(
            self.predictive_covariance_cho,
            self.prior.k(self.X, b),
        )
```

Task A: Multivariate Gausian Processes - The MathWorks / MATLAB Logo

The lecture mostly focused on Gaussian processes whose input is real-valued. However, Gaussian processes can be applied to arbitrary inputs, including graphs and images.

Here, we explore multivariate GPs, i.e. the input set $\mathbb{X}$ is a subset of $\mathbb{R}^d$, in particular $d = 2$.

We will use a Gaussian process to reconstruct the Mathworks / MATLAB logo



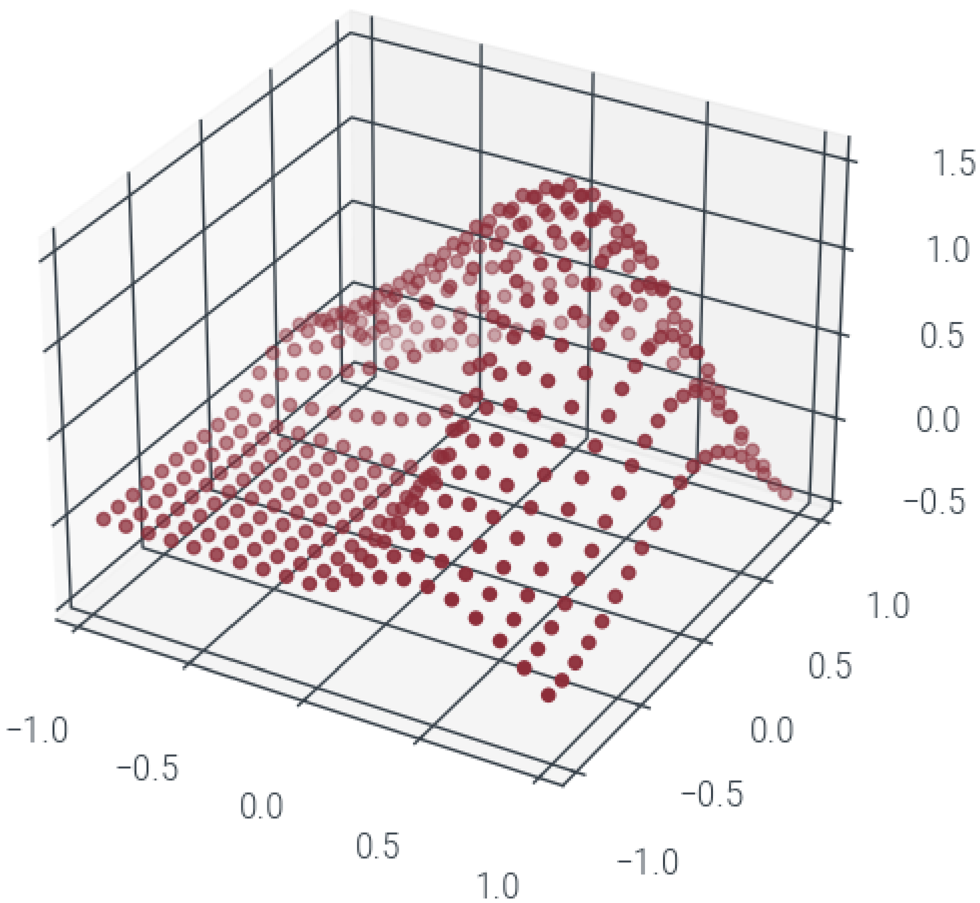
from a sparse set of measurements. The logo depicts an eigenfunction of the wave equation on an L-shaped domain.

In [5]:
```
# Load the measurements
data = np.load("matlab_logo.npz")

X = jnp.asarray(data["X"])
y = jnp.asarray(data["y"])
```

```
No GPU/TPU found, falling back to CPU. (Set TF_CPP_MIN_LOG_LEVEL=0 and rer
un for more info.)
```

In [6]:
```
# Visualize the data
fig, ax = plt.subplots(subplot_kw={"projection": "3d"})

ax.scatter(
    X[..., 0].ravel(),
    X[..., 1].ravel(),
    y.ravel(),
    marker=".",
    s=20,
)
```

Out[6]:
```
<mpl_toolkits.mplot3d.art3d.Path3DCollection at 0x7f2fd4383310>
```



We can reuse the `(Conditional)GaussianProcess` implementation from the lecture to learn multivariate functions.

However, we need to define a mean function m: $\R^2 \to \R$ and a positive definite kernel k: $\R^2 \times \R^2 \to \R$, which accept bivariate inputs.

For simplicity, we will choose the mean function to be zero everywhere, i.e. $m(x) = 0$.

In [7]:
```
def multivariate_zero_mean(x):
    return 0
```

There are many different choices for multivariate kernel functions. Here, we will focus on so-called **radial kernels**, which are of the form

$$k(\vec{x}_1, \vec{x}_2) = \tilde{k}\left(\frac{\|\vec{x}_1 - \vec{x}_2\|_2}{\ell}\right)$$

for some function $\tilde{k}$: $\mathbb{R}_{\geq 0} \to \mathbb{R}$.

Some of the kernels from the lecture, including the Gaussian, rational quadratic, and Matérn kernels, are radial. However, the lecture only introduced their 1D versions. For instance, the multivariate rational quadratic kernel is given by

$$k_{\alpha,l}(\vec{x}_1, \vec{x}_2) := \tilde{k}_\alpha\left(\frac{\|\vec{x}_1 - \vec{x}_2\|_2}{\ell}\right), \quad \text{where}$$
$$\tilde{k}_\alpha(r) = \left(1 + \frac{r^2}{2\alpha}\right)^{-\alpha}.$$

Implement the rational quadratic kernel with $\alpha = 1.0$, $\ell = 0.25$ and an output scale of $\sigma = 0.5$.

In [8]:
```
def multivariate_rational_quadratic(
    x0,
    x1,
    *,
    alpha=1.0,
    lengthscale=0.25,
    output_scale=0.5,
):
    x0 = jnp.asarray(x0)
    x1 = jnp.asarray(x1)

    def K_til(a, r):
        return (1 + r**2 / (2 * a)) ** (-a)

    dist = jnp.sqrt(jnp.sum((x0 - x1) ** 2, axis=-1))
    return output_scale * K_til(alpha, dist / lengthscale)
```

In [9]:
```
mv_prior = GaussianProcess(
    multivariate_zero_mean,
    multivariate_rational_quadratic,
)
```

Condition the Gaussian process prior on the training data. Here, we assume no observation noise.

Keep in mind that the inputs `y` and `X` to `GaussianProcess.condition` need to have shape `(n_samples,)` and `(n_samples, n_inputs)`, respectively.

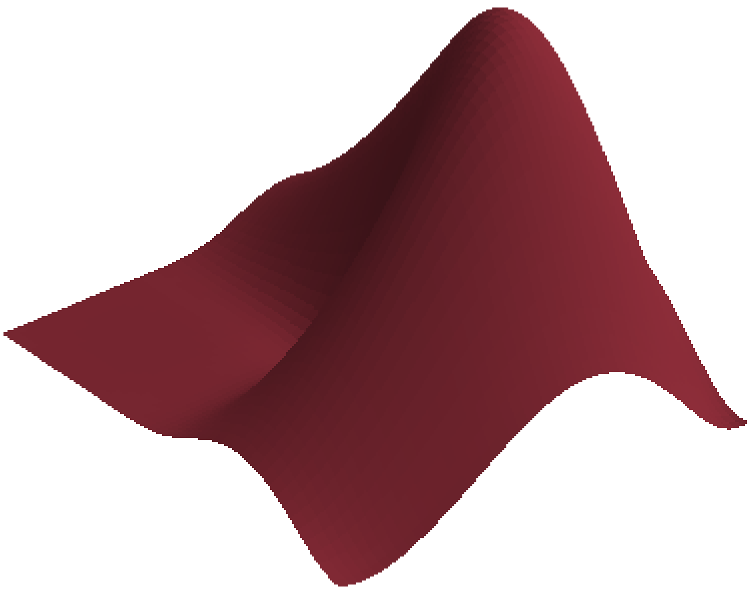
In [10]:
```
# reshape the data
X = X.reshape((400, 2))
y = y.ravel()

# inference
mv_posterior = mv_prior.condition(y, X, sigma=0.5)
```

Plot the posterior mean in a 3D surface plot.

In [11]:
```
plt_grid = np.stack(
    np.meshgrid(
        np.linspace(-1, 1, 200),
        np.linspace(-1, 1, 200),
    ),
    axis=-1,
)
```

In [12]:
```
import matplotlib.colors
```

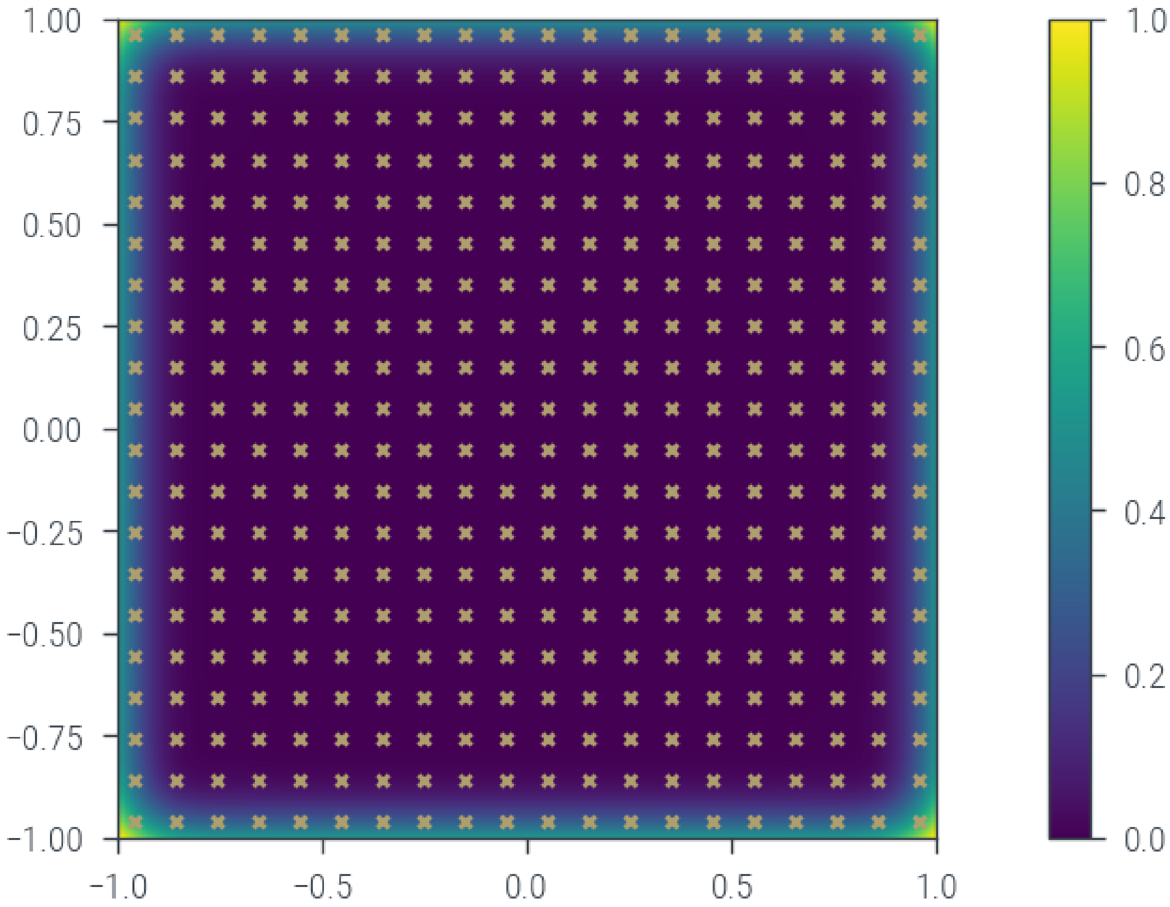
```
fig, ax = plt.subplots(subplot_kw={"projection": "3d"})

mX = mv_posterior.m(plt_grid)

ax.plot_surface(
    plt_grid[..., 0],
    plt_grid[..., 1],
    mX,
    color="C0",
    lightsource=matplotlib.colors.LightSource(30, 30),
    antialiased=False,
)

ax.set_aspect("equal")
ax.view_init(azim=-40)
ax.set_axis_off()
```



Plot the posterior standard deviation as a heatmap and superimpose the data points in a scatter plot:

In [13]:
```
stdX = np.sqrt(mv_posterior.k(plt_grid, plt_grid))

plt.imshow(
    stdX,
    extent=(-1, 1, -1, 1),
)
plt.scatter(
    X[..., 0],
    X[..., 1],
    c="C2",
```

```
    marker="x",
    s=5,
)
plt.colorbar()
plt.show()
```



Task B: Multi-output Gaussian Processes

In the second part of the exercise, we will construct a GP-based model for a vector-valued function - a so-called **multi-output Gaussian process**.

As a toy example, we will try to track a cargo ship on the ocean after losing communication for 15 minutes.

The trajectory of the ship can be modeled by a function $\vec{s}\colon \mathbb{R}_{\geq 0} \to \mathbb{R}^2$, which maps time to 2D coordinates on a map.

The ship started in a harbor at time $t = 0$ and coordinates $\vec{s}(0) = (0, 0)$. It receives GPS readings every $\frac{1}{12}$ hours (5 minutes), which have an accuracy of about 5 meters (in the sense of the 95% confidence interval). We assume the measurement noise to be Gaussian.

However, after 1 hour, we lose communication to the cargo ship. Your task is to predict the position of the ship after 1.25h using a multi-output Gaussian process.

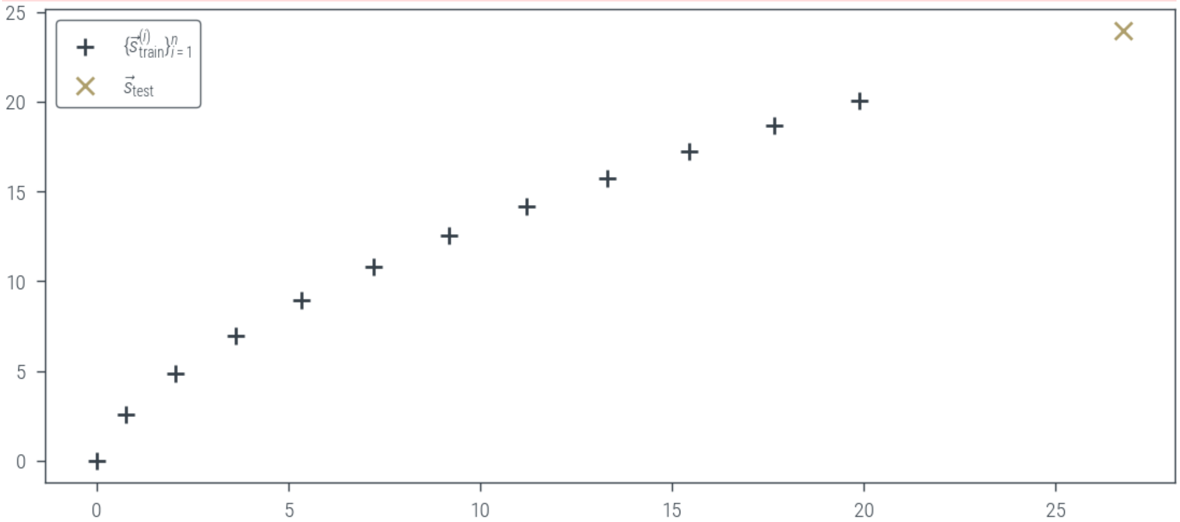
In [14]:
```
# Simulate some trajectory data
ts = np.linspace(
    0.0, 1.25, 15
```

```
)  # 1.25h of training data with training points every 5 minutes

ss = np.stack(
    (ts, np.sqrt(ts)),
    axis=-1,
)
# Average speed of cargo ship is 22-37 km/h
ss[1:, :] *= (
    30.0
    * (ts[1:] / np.cumsum(np.linalg.norm(ss[1:, :] - ss[:-1, :], ord=2, a
        :, None
    ]
)

sigma_train = (0.005 / 1.96) ** 2  # 5 m accuracy of GPS

# Train / test split
ts_train = ts[:12]
ss_train = ss[:12]

t_test = ts[-1]
s_test = ss[-1]
```

In [15]:
```
def plot_data(ax=plt.gca()):
    ax.plot(
        ss_train[:, 0],
        ss_train[:, 1],
        "+",
        c="C1",
        label=r"$\{ \vec{s}^{(i)}_\mathrm{train} \}_{i = 1}^n$",
    )

    ax.plot(*s_test, "x", c="C2", label=r"$\vec{s}_\mathrm{test}$")


plot_data()
plt.legend()
```

Out[15]: `<matplotlib.legend.Legend at 0x7f2f9c52acb0>`

```
findfont: Font family ['cursive'] not found. Falling back to DejaVu Sans.
findfont: Generic family 'cursive' not found because none of the following
families were found: Apple Chancery, Textile, Zapf Chancery, Sand, Script
MT, Felipa, Comic Neue, Comic Sans MS, cursive
```



We need to posit a Gaussian process prior over the unknown trajectory $s\colon \mathbb{R}_{\geq 0} \to \mathbb{R}^2$.

However, since kernel functions are defined to be scalar valued, this generalization is not as straightforward as the generalization to arbitrary inputs. Fortunately, we can use the fact that Gaussian processes can be defined on arbitrary input sets to "emulate" vector-valued functions.

To this end, note that a function $\vec{f} : \mathbb{X} \to \R^d$ is in some sense equivalent to the function

$$\tilde{f} : \{1, \dots, d\} \times \mathbb{X} \to \R, (i, x) \mapsto \vec{f}_i(x), \tag{1}$$

since

$$\vec{f}(x) = \begin{pmatrix} \tilde{f}(1, x) \\ \vdots \\ \tilde{f}(d, x) \end{pmatrix}. \tag{2}$$

We can use this equivalence to construct a vector-valued Gaussian process $f \sim \mathcal{GP}(m, k)$. In this case, the mean function is given by $m\colon \{1, \dots, d\} \times \mathbb{X} \to \R$ and the kernel function is given by $k\colon (\{1, \dots, d\} \times \mathbb{X}) \times (\{1, \dots, d\} \times \mathbb{X}) \to \R$, where $k((i, x_1), (j, x_2))$ computes the covariance between $f_i(x_1)$ and $f_j(x_2)$.

The specific GP prior for the ship tracking problem can be motivated as follows.

We assume that the velocity $\dot{\vec{s}}$ of the ship is well-modeled by two independent Wiener processes (one for each component) with output scales σ_1 and σ_2, respectively, and constant mean $\vec{v}_0$. Then, one can show that $\vec{s} = \int_0^t \dot{\vec{s}}(\tau) \mathrm{d}\tau \sim \mathcal{GP}(m, k)$ with

$$\begin{aligned} m(i, t) &= (\vec{v}_0)_i t, \qquad \text{and} \\ k((i, t_1), (j, t_2)) &= \begin{cases} \sigma_i^2 k_{\mathrm{IWP}}(t_1, t_2) & \text{if } i = j \\ 0 & \text{otherwise,} \end{cases} \end{aligned}$$

where k_{IWP} is the integrated Wiener process kernel (a.k.a. cubic spline kernel) from the lecture, i.e.

$$k_{\mathrm{IWP}}(t_1, t_2) = \frac{1}{3} \min^3(t_1, t_2) + \frac{1}{2} |t_1 - t_2| \min^2(t_1, t_2). \tag{3}$$

In [16]:

```
v0 = (ss_train[1, :] - ss_train[0, :]) / (ts_train[1, None] - ts_train[0,


def multi_output_mean(it):
    it = jnp.asarray(it)
    i, t = it[..., 0], it[..., 1]
    i = i.astype(jnp.int_)

    return v0[..., i] * t
```

In [17]:

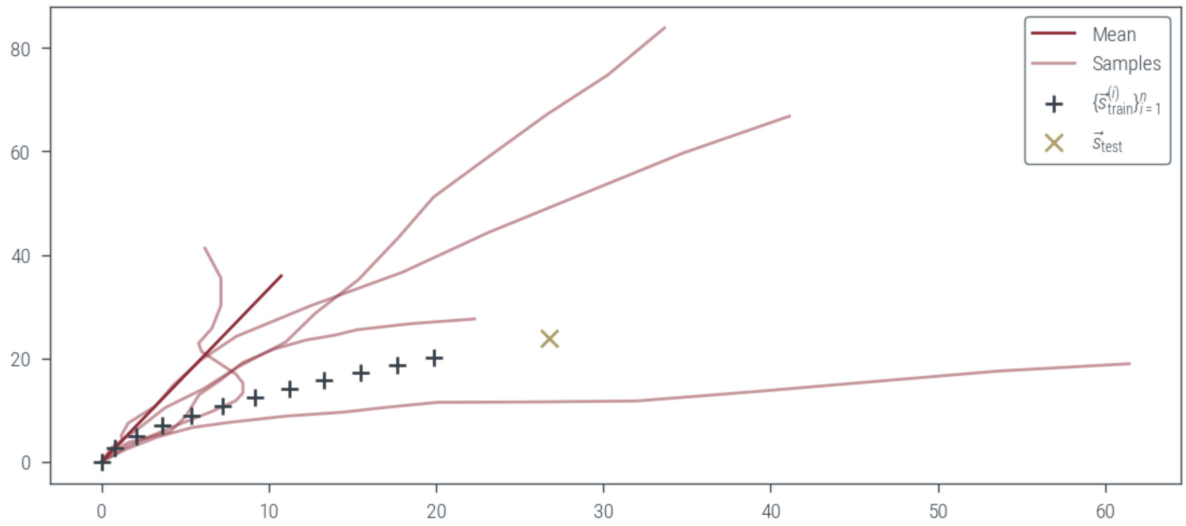
```
def multi_output_kernel(it0, it1, output_scales=(30.0**2, 30.0**2)):
    it0 = jnp.asarray(it0)
```

```
    i0, t0 = it0[..., 0], it0[..., 1]
    i0 = i0.astype(jnp.int_)

    it1 = jnp.asarray(it1)
    i1, t1 = it1[..., 0], it1[..., 1]
    i1 = i1.astype(jnp.int_)

    output_scales = jnp.asarray(output_scales)

    def K_iwp(T1, T2):
        return (1 / 3) * jnp.minimum(T1, T2) ** 3 + (1 / 2) * jnp.abs(
            T1 - T2
        ) * jnp.minimum(T1, T2) ** 2

    return output_scales[..., 0] * K_iwp(t0, t1) * (i0 == i1)
```

In [18]:
```
s_prior = GaussianProcess(multi_output_mean, multi_output_kernel)
```

In [19]:
```
ts_plot = np.linspace(0.0, 1.25, 200)
its_plot = np.stack(
    np.meshgrid(
        [0, 1],
        ts,
        indexing="ij",
    ),
    axis=-1,
)

# GP Mean
m_prior_its_plot = s_prior.m(its_plot)

plt.plot(
    m_prior_its_plot[0, :],
    m_prior_its_plot[1, :],
    c="C0",
    label="Mean",
)

# GP Samples
num_samples = 5
prior_samples = (
    s_prior(its_plot.reshape(-1, 2))
    .sample(jax.random.PRNGKey(3), num_samples=num_samples)
    .reshape(num_samples, 2, -1)
)

plt.plot(
    prior_samples[:, 0, :].T,
    prior_samples[:, 1, :].T,
    c="C0",
    alpha=0.5,
    label=["Samples"] + [None for _ in range(num_samples - 1)],
)

plot_data(plt.gca())

plt.legend()
```

Out[19]: <matplotlib.legend.Legend at 0x7f2fd43833a0>



Given a training dataset $\mathcal{D} = \{(t^{(i)}_{\text{train}}, \vec{s}^{(i)}_{\text{train}})\}_{i=1}^{n}$, we can use the `(Conditional)GaussianProcess` implementation from the lecture to condition our multi-output GP prior on the augmented dataset

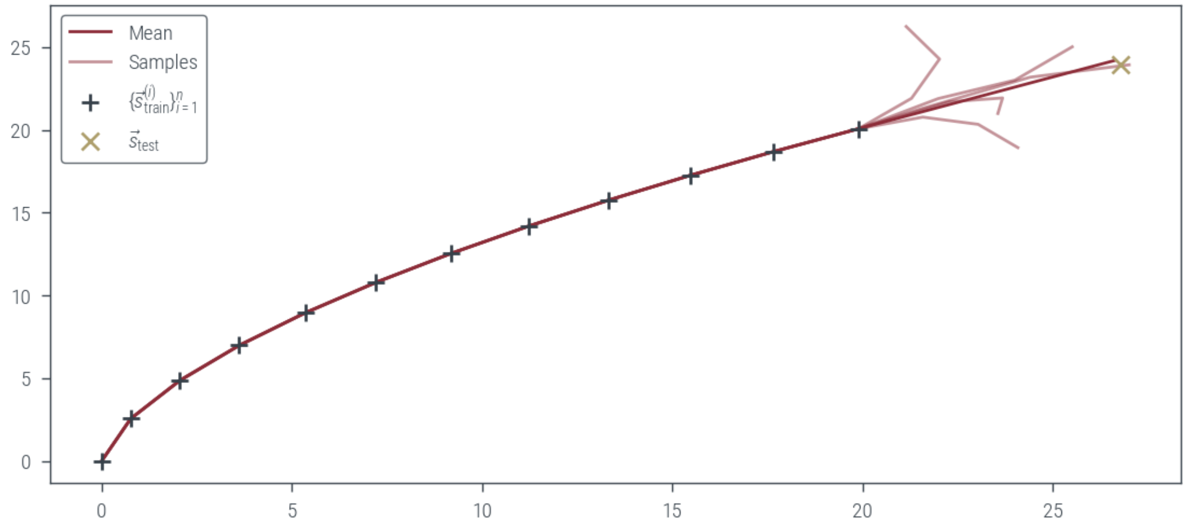
$$\tilde{\mathcal{D}} = \left\{ \dots, \left((1, t^{(i)}_{\text{train}}), (\vec{s}^{(i)}_{\text{train}})_1 \right), \left((2, t^{(i)}_{\text{train}}), (\vec{s}^{(i)}_{\text{train}})_2 \right), \dots \right\}. \tag{4}$$

In [20]:
```
ss_train_aug = ss_train.reshape(-1)

its_train_aug = jnp.asarray(
    [(i%2, t) for i, t in enumerate(np.repeat(ts_train, 2))]
)

s_posterior = s_prior.condition(
    ss_train_aug,
    its_train_aug,
    sigma_train,
)
```

In [21]:
```
# GP Mean
m_posterior_its_plot = s_posterior.m(its_plot)

plt.plot(
    m_posterior_its_plot[0, :],
    m_posterior_its_plot[1, :],
    c="C0",
    label="Mean",
)

# GP Samples
num_samples = 5
posterior_samples = s_posterior(its_plot.reshape(-1, 2)) \
                    .sample(jax.random.PRNGKey(3), num_samples=num_sample
                    .reshape(num_samples, 2, -1)

plt.plot(
    posterior_samples[:, 0, :].T,
    posterior_samples[:, 1, :].T,
    c="C0",
    alpha=0.5,
    label=["Samples"] + [None for _ in range(num_samples - 1)],
)
```

```
plot_data(plt.gca())

plt.legend()
```

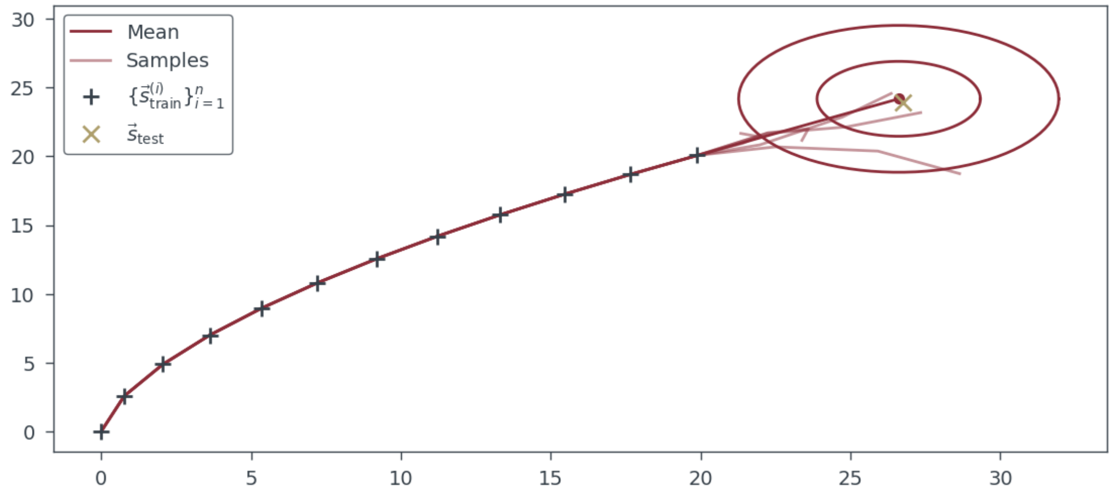
Out[21]: `<matplotlib.legend.Legend at 0x7f2f9c1d6f20>`



Write a function `plot_belief(ax, gp, t, ...)` that visualizes the mean and the 68% and 95% (1σ and 1.96σ) isoprobability contours of $\vec{s}(t)$.

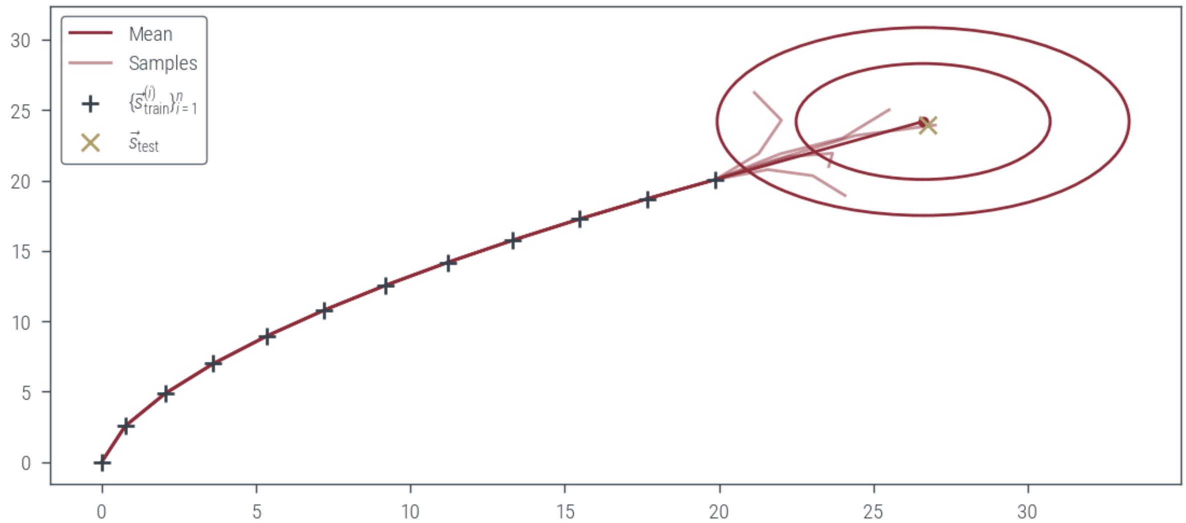
Hint: If $\vec{s} \sim \mathcal{GP}$, then $\vec{s}(t)$ follows a bivariate normal distribution, whose isoprobability contours are ellipses.

The final plot should look like this:



In [22]:
```
from matplotlib.patches import Ellipse


def plot_belief(ax, gp, t, perc):
    mean = gp(np.array([(0, t), (1, t)])).mu
    cov = gp(np.array([(0, t), (1, t)])).Sigma

    vals = jnp.asarray([cov[0][0], cov[1][1]])
    c = -2 * np.log(1 - perc)
    width, heigth = 2 * np.sqrt(c * vals)

    ax.add_patch(Ellipse(mean, width, heigth, fill=False, edgecolor="C0",
```

In [23]:
```
# GP Mean
m_posterior_its_plot = s_posterior.m(its_plot)

plt.plot(
    m_posterior_its_plot[0, :],
    m_posterior_its_plot[1, :],
    c="C0",
    label="Mean",
)

# GP Samples
num_samples = 5
posterior_samples = s_posterior(its_plot.reshape(-1, 2)) \
                    .sample(jax.random.PRNGKey(3), num_samples=num_sample
                    .reshape(num_samples, 2, -1)

plt.plot(
    posterior_samples[:, 0, :].T,
    posterior_samples[:, 1, :].T,
    c="C0",
    alpha=0.5,
    label=["Samples"] + [None for _ in range(num_samples - 1)],
)

# Predictive Belief
plot_belief(plt.gca(), s_posterior, t=1.25, perc=0.95)
plot_belief(plt.gca(), s_posterior, t=1.25, perc=0.68)

mean = s_posterior.m([(0, 1.25), (1, 1.25)])
plt.scatter(mean[0], mean[1], marker="o", color="C0", s=10)

plot_data(plt.gca())

plt.legend()
plt.show()
```



In [ ]: