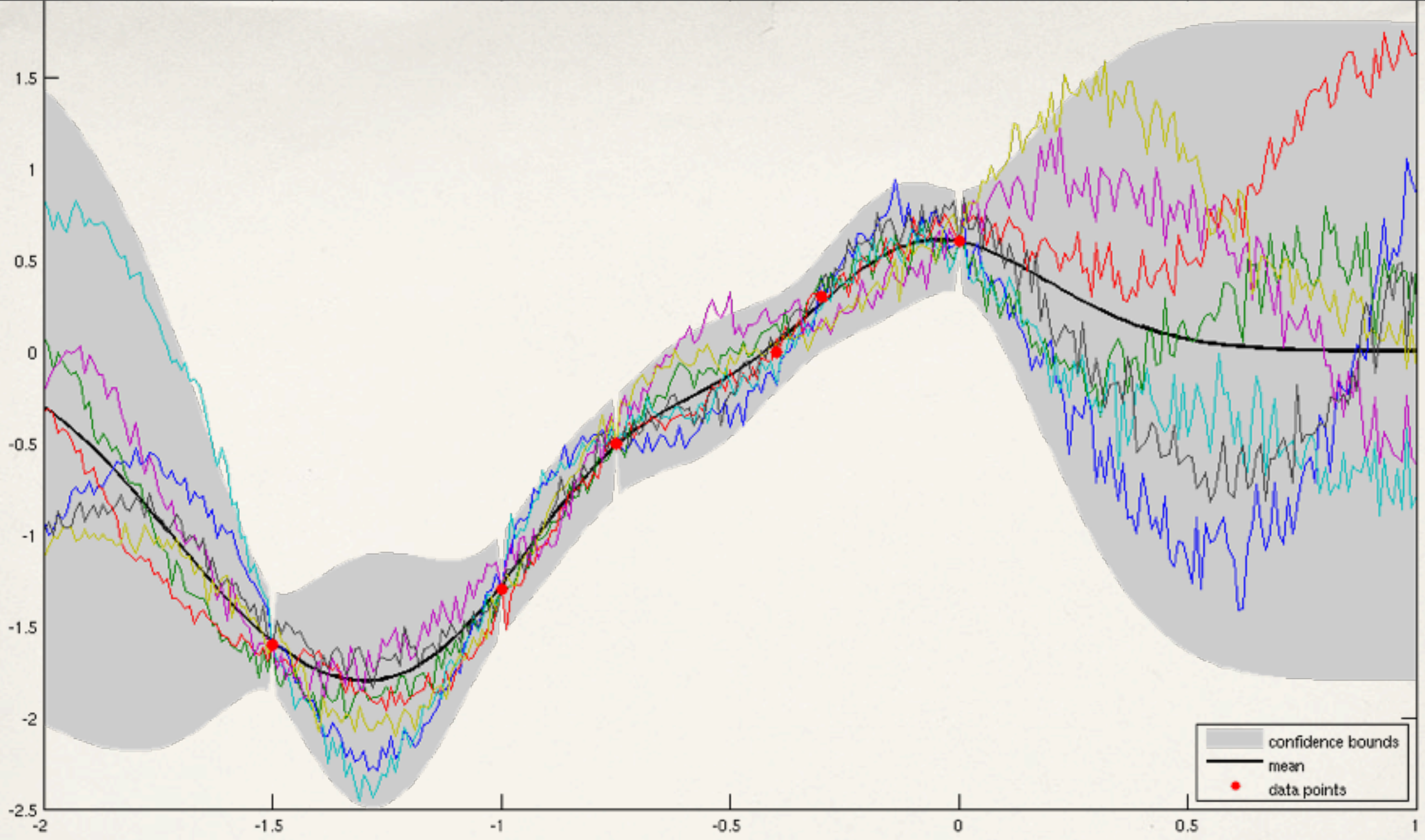


Gaussian Process - Introduction

Martin Kolář

Brno University of Technology - Faculty of Information Technology

Univariate Normal Sampling

$$\mathcal{N}(\mu, \sigma^2) \sim \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- given μ (the mean) and σ^2 (the variance), the distribution is set
- to sample, take a random number generated by the standard normal distribution:

$$\mathcal{N}(0,1)$$

- and apply

$$\mathcal{N}(\mu, \sigma^2) = \mathcal{N}(0,1) \cdot \sigma + \mu$$

[demo #1]

Multivariate Normal Sampling

$$\mathcal{N}(\mu, \Sigma) \sim |2\pi\Sigma|^{-\frac{1}{2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}$$

- for 2 dimensions, this is called the Bivariate Normal
- Sampling is similar, but we need the matrix A such that

$$A^T A = \Sigma$$

- then, given a sample from the standard Multivariate Normal

$$\mathcal{N}(\mu, \Sigma) = A\,\mathcal{N}(0, I) + \mu$$

[demo #2]

Simple Matrix Algebra

- Performing the eigendecomposition of Σ, we get two matrices such that:

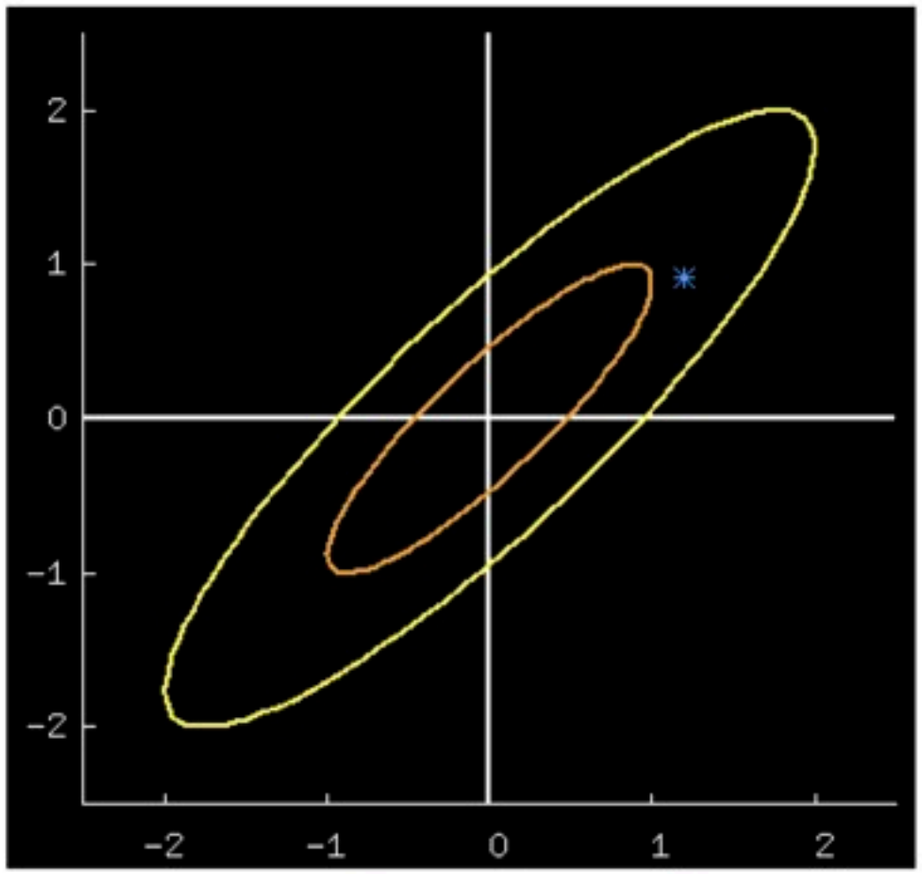
$$VDV^T = \Sigma$$

- Where D is a diagonal matrix. So, a matrix

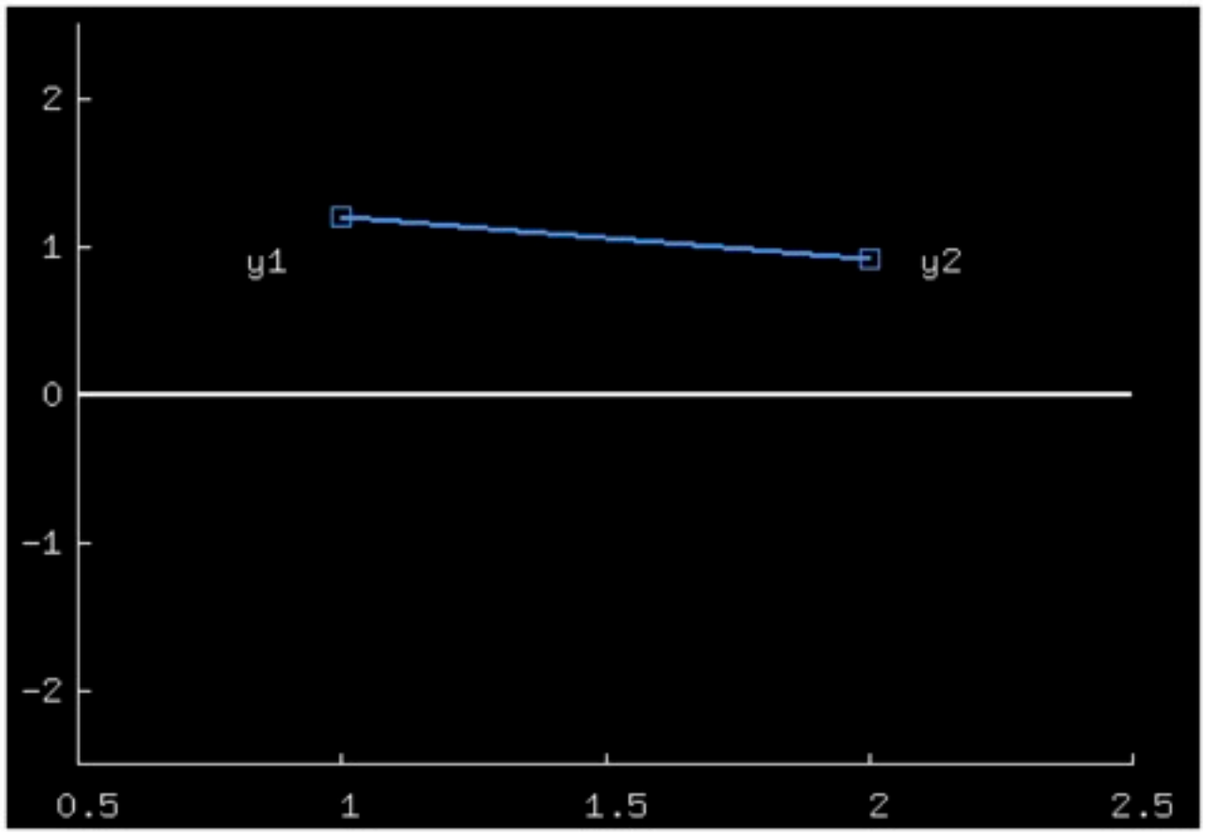
$$A = V\ diag(\sqrt{diag(D)})$$

- will satisfy

$$A^T A = \Sigma$$

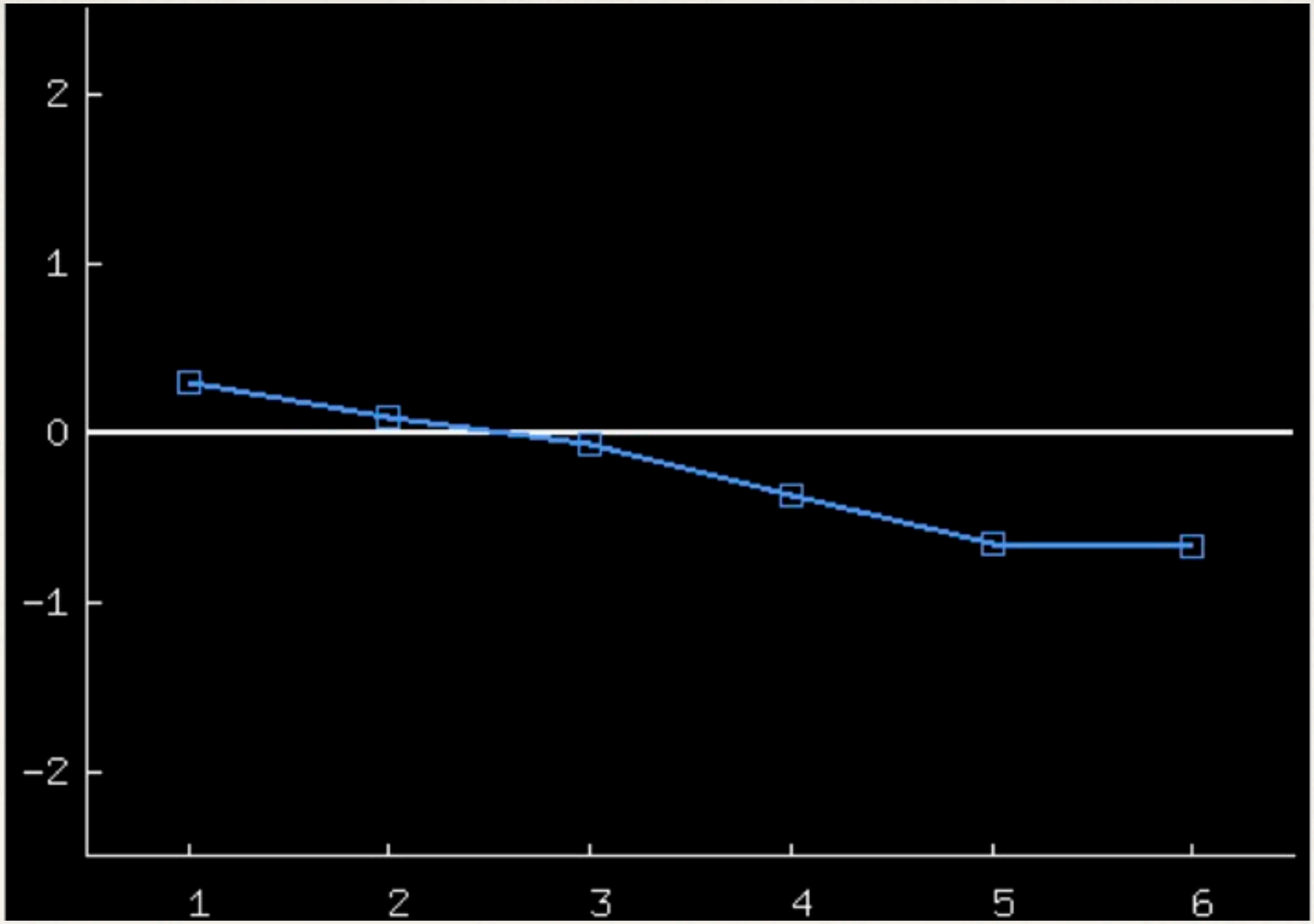


$$\mathbf{y} = \begin{bmatrix} 1.2 & 0.9 \end{bmatrix}$$



For a sample from a Multivariate Normal, any number of dimensions can be represented by a line.

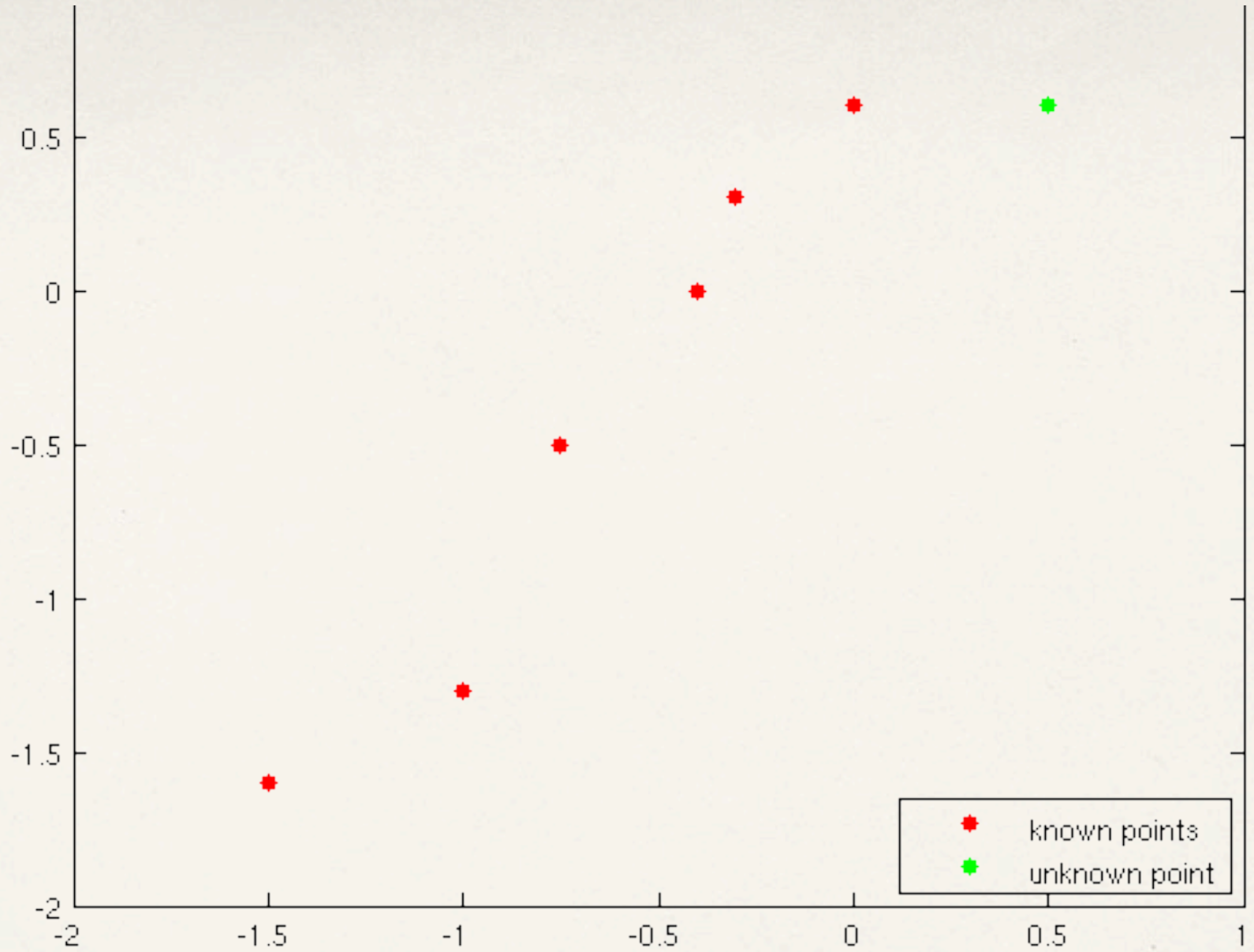
*MacKay [2007]



The covariance on adjacent dimensions can be set so that points n and n+1 are unlikely to differ, so that the line is smooth.

---

*MacKay [2007]



We would like to use a framework like this to construct a probability distribution over y at at any x, given other points and a type of function (oscillating/polynomial/smooth/...)

For that, we need a continuous enumeration of dimensions.

Covariance Kernel

- In order to have a continuous enumeration of dimensions, we need an infinite-dimensional multivariate normal.

- We define the infinite dimensional square matrix with a continuous function of two variables:

$$\Sigma_{ab} = k(x_a, x_b)$$

- We do not define the mean, but the covariance function implicitly creates restrictions on this.

- The trick is that we only ever construct the covariance matrix for the points we are interested in, so the problem is tangible

Gaussian Process

- There are many popular kernel functions, with nice properties.
- For example, this is the squared exponential kernel:

$$k(x_a, x_b) = \sigma_f^2 \, e^{-\frac{(x_a - x_b)^2}{2l^2}}$$

- Given no data, the mean and variance of a Gaussian Process is just:

$$\mu(x) = 0$$

$$\sigma(x) = k(x, x)$$

- (the mean and variance are continuous functions of x for all $\mathbb{R}$)

[demo #3]

Gaussian Process Sampling

- In order to draw a random sample, we need a covariance matrix and mean vector. If we have no observations, we construct those for the values of x we are interested in by:

$$x = \{-0.5, 0.99, 1\}$$

$$\mu = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, K = \begin{bmatrix} k(x_1,x_1) & k(x_1, x_2) & k(x_1, x_3) \\ k(x_2, x_1) & k(x_2, x_2) & k(x_2, x_3) \\ k(x_3, x_1) & k(x_3, x_2) & k(x_3, x_3) \end{bmatrix}$$

- And draw a sample using the same method as with the Gaussian

$$A^T A = K \qquad \mathcal{N}(\mu, \Sigma) = A\,\mathcal{N}(0, I) + \mu$$

[demo #4]

Gaussian Process Regression

- In order to incorporate date into the prior, we simply add it to the Kernel matrix:

$$x_{data} = \{0.5, 1.5\} \qquad y_{data} = \{-1, 1\} \qquad x_{prediction} = \{0.75\}$$

- However, we need to keep these in three separate sub-matrices

$$K_d = \begin{bmatrix} k(0.5,0.5) & k(0.5,1.5) \\ k(0.5,1.5) & k(1.5,1.5) \end{bmatrix}, K_{pd} = \begin{bmatrix} k(0.75,0.5) & k(0.75,1.5) \end{bmatrix}, K_p = \begin{bmatrix} k(0.75,0.75) \end{bmatrix}$$

$$K = \begin{bmatrix} K_d & K_{pd}^T \\ K_{pd} & K_d \end{bmatrix} = \begin{bmatrix} k(0.5,0.5) & k(0.5,1.5) & k(0.5,0.75) \\ k(1.5,0.5) & k(1.5,1.5) & k(1.5,0.75) \\ k(0.75,0.5) & k(0.75,1.5) & k(0.75,0.75) \end{bmatrix}$$

Complicated Matrix Algebra

- A result of matrix algebra is that

$$\begin{bmatrix} a \\ b \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} A & B^T \\ B & \mathrm{C} \end{bmatrix}\right)$$

- is equivalent to*

$$a \,|\, b \sim \mathcal{N}\left(BA^{-1}b, \mathrm{C} - BA^{-1}B^T\right)$$

*under specific conditions that are satisfied here

Gaussian Process Regression

- A result of matrix algebra is that

$$\begin{bmatrix} y_{data} \\ y_{prediction} \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K_d & K_{pd}^T \\ K_{pd} & K_p \end{bmatrix}\right)$$

- equals

$$y_{prediction} \mid y_{data} \sim \mathcal{N}\left(K_{pd}K_d^{-1}y_{data},\ K_d - K_{pd}K_d^{-1}K_{pd}^T\right)$$

- You can now calculate the mean, the variance, and take samples for any observed and predicted points. Congratulations!

[demo #5]

Properties of Kernels

- In order to find the desired Kernel parameters, we maximise the evidence:

$$\log p(y_{data} \mid x_{data}, \theta) = -\frac{1}{2} y^T K^{-1} y - \frac{1}{2} \log |K| - \frac{n}{2} \log(2\pi)$$

- This space is smooth, so we can use the Conjugate Gradient method, for example

[demo #6]

- Certain Kernels are harder to optimise, but we can integrate over all* values of θ in a Bayesian way to find the optimum (Rasmussen & Williams [2006] - Chapter 5)

Properties of Kernels

- Kernels in Gaussian Processes are analogous to SVM Kernels, and define a mapping in a different space.
  - This space can be infinite-dimensional
  - Making calculations in the Kernel space rather than the mapping space is called the kernel trick
  - That's why the Covariance matrix has a number of dimensions defined by the number of points, rather than the problem space
- Kernels can be added to produce a sum of the functions they represent

[demo #7]