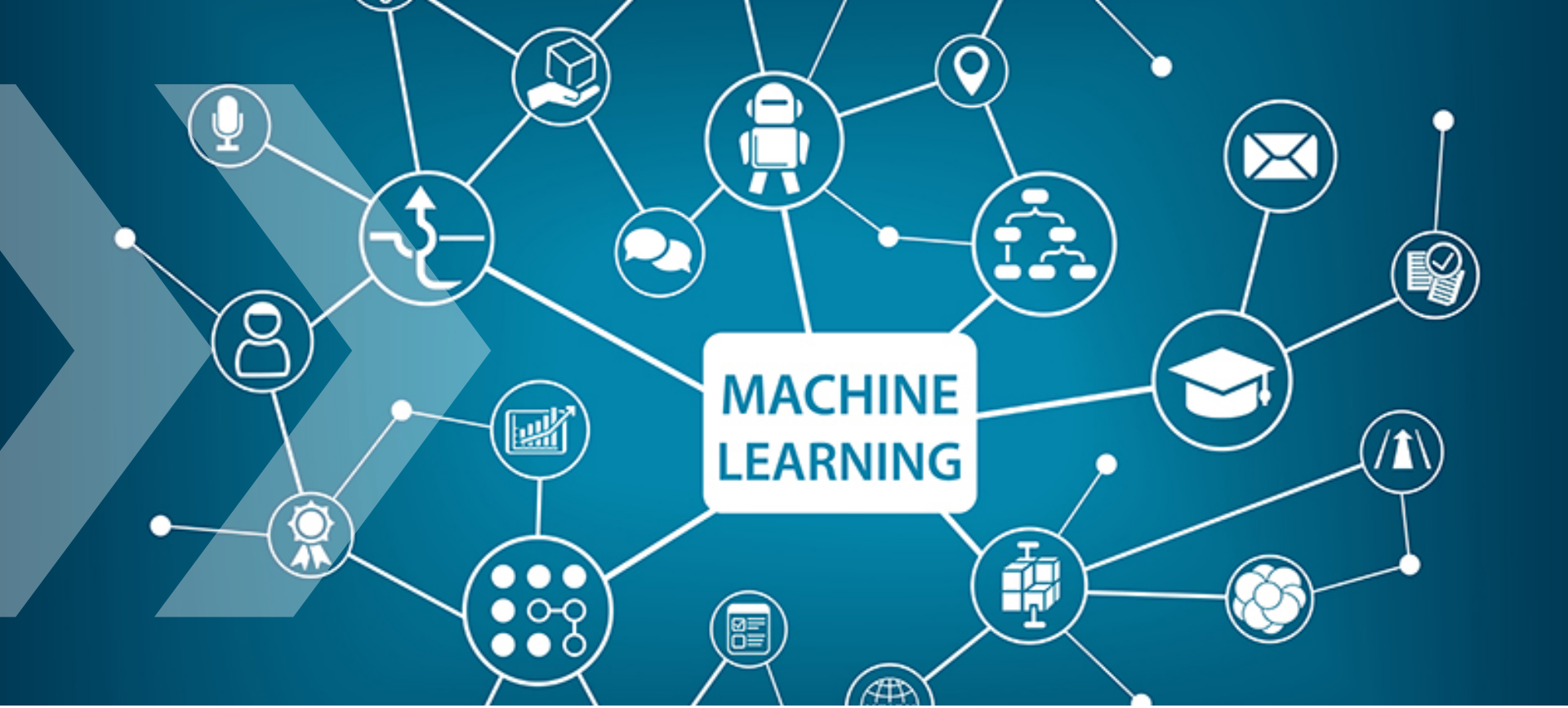


CS489/698: Intro to ML

Lecture 09: Gaussian Processes

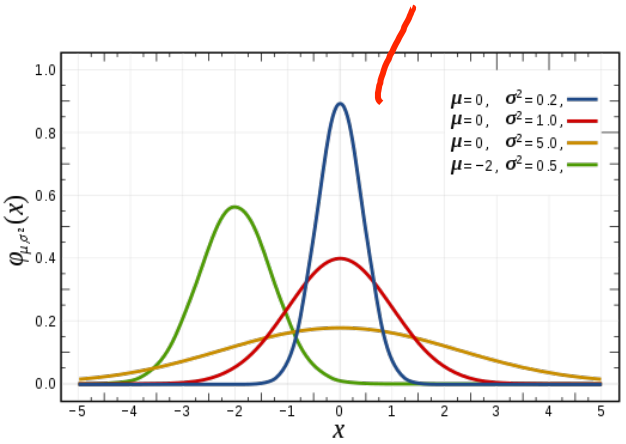
UNIVERSITY OF WATERLOO

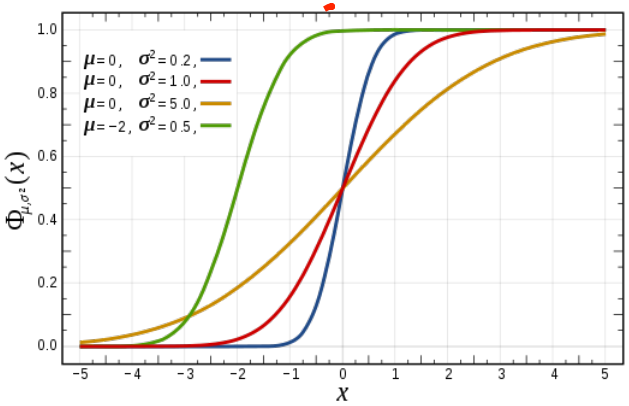
Outline

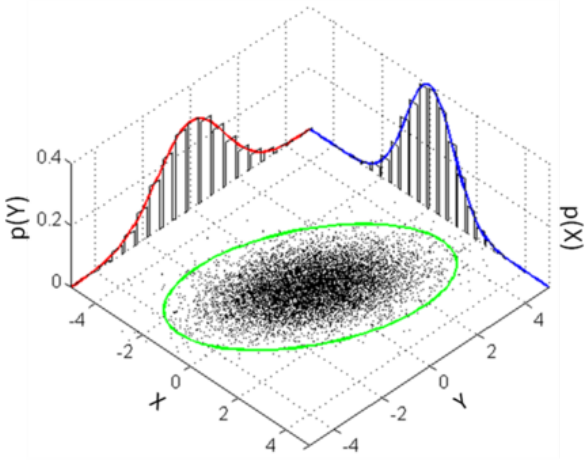
- Gaussian Distribution
- Gaussian Process
- Gaussian Linear Regression
- Advanced

Gaussian distribution

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[- \frac{(x-\mu)^2}{2\sigma^2} \right]$$







$$p(\mathbf{x}) = (2\pi)^{-d/2} |\Sigma|^{-1/2} \exp\left[- \tfrac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right]$$

$$X \sim \mathcal{N}_d(\boldsymbol{\mu}, \Sigma)$$

dimension — mean — covariance, PSD

UNIVERSITY OF WATERLOO

Carl Friedrich Gauss (1777 – 1855)

UNIVERSITY OF WATERLOO

Important facts

$$X \sim \mathcal{N}_d(\boldsymbol{\mu}, \Sigma), \quad A \in \mathbb{R}^{p \times d} \Rightarrow AX \sim \mathcal{N}_p(A\boldsymbol{\mu}, A\Sigma A^\top)$$

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right)$$

joint = marginal ✖ conditional

$$X_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \Sigma_{11})$$

$$X_2 | X_1 \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$$

$$\boldsymbol{\mu}_2 + \Sigma_{21}\Sigma_{11}^{-1}(X_1 - \boldsymbol{\mu}_1)$$

$$\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$$

Derivation

$$\begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}^{-1} = \begin{bmatrix} \Sigma_{11}^{-1} + \Sigma_{11}^{-1}\Sigma_{12}\Sigma^{-1}\Sigma_{21}\Sigma_{11}^{-1} & -\Sigma_{11}^{-1}\Sigma_{12}\Sigma^{-1} \\ -\Sigma^{-1}\Sigma_{21}\Sigma_{11}^{-1} & \Sigma^{-1} \end{bmatrix}$$

$$\Sigma = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$$

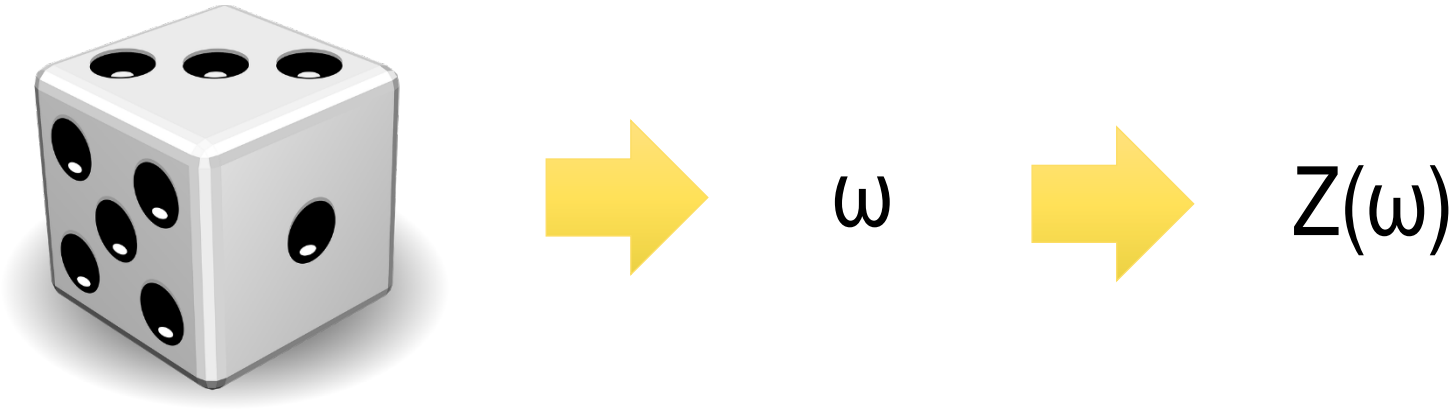
Schur Complement

Outline

- Gaussian Distributions
- Gaussian Processes
- Gaussian Linear Regression
- Advanced

What is a random variable?

- A random variable is a function $Z(\omega)$



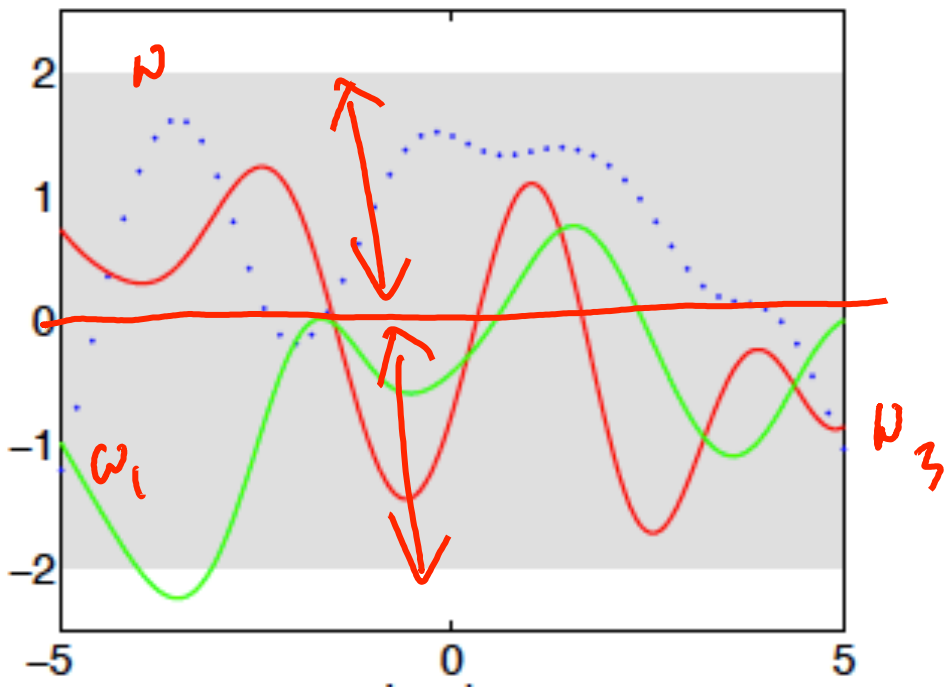
Gaussian process

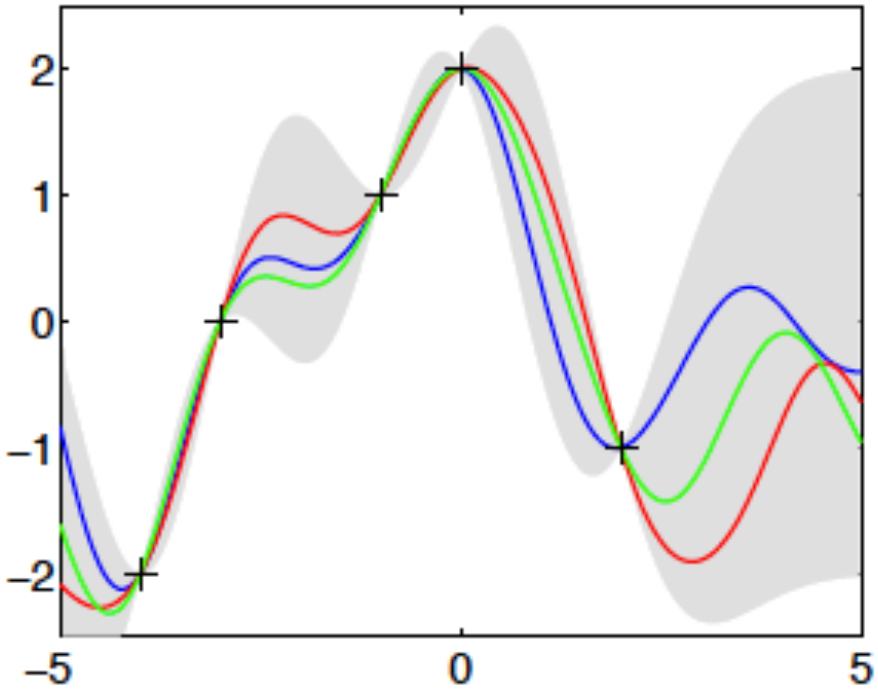
- A collection of Gaussian random variables $\{Z_t : t \text{ in } T\}$ such that for any finite N, $\{Z_t : t \text{ in } N\}$ is jointly Gaussian

- A Gaussian process is a function of two variables $Z(t,\omega)$
  - For any finite N, $\{Z_t := Z(t,\omega) \mid t \text{ in } N\}$ is a Gaussian random vector
  - For any ω, $Z_\omega := Z(t, \omega) : T \to \mathbf{R}$ is a function of one variable t (sample path)

- Does Gaussian process exist?

Example





Mean and covariance function

- For each t, Z(t,ω) is a Gaussian random variable hence has mean $m(t) := \mathbf{E}[Z_t]$

- For each s and t, the covariance between Z_s and Z_t:

$$\kappa(s,t) := \mathbf{E}[(Z_s - m(s))(Z_t - m(t))]$$

- Say $Z \sim \mathcal{GP}(m, \kappa)$

mean function

covariance function

Recap: verifying a kernel

For any *n*, for any $x_1, x_2, \ldots, x_n$, the kernel matrix *K* with

$$K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$$

is symmetric and positive semidefinite ($K \in \mathbb{S}_+^d$)

- Symmetric: $K_{ij} = K_{ji}$
- Positive semidefinite (PSD): for all $\boldsymbol{\alpha} \in \mathbf{R}^n$

$$\boldsymbol{\alpha}^\top K \boldsymbol{\alpha} = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j K_{ij} \geq 0$$

What is a covariance function?

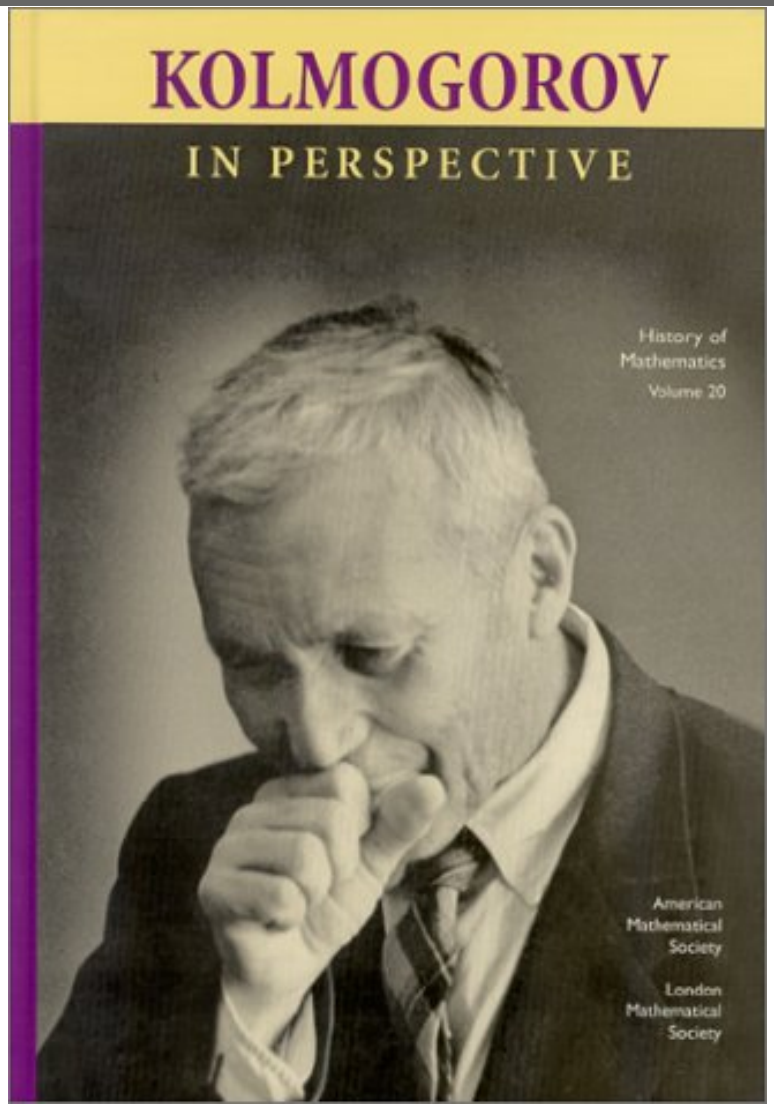
- $\kappa : T \times T \to \mathbb{R}$
- For $t_1, t_2, \ldots, t_n$, $K_{ij} = \kappa(t_i, t_j)$ by definition is the covariance between Z_{t_i} and Z_{t_j}
- K is symmetric and PSD
- Thus, the covariance function κ is a kernel !

Conversely

Theorem. Given any function $m(t)$ and kernel function $\kappa(s,t)$, exist $Z \sim \mathcal{GP}(m, \kappa)$

This may not hold for other distributions !!!

Andrey Kolmogorov (1903 - 1987)



ERGEBNISSE DER MATHEMATIK
UND IHRER GRENZGEBIETE
HERAUSGEGEBEN VON DER SCHRIFTLEITUNG
DES
„ZENTRALBLATT FÜR MATHEMATIK"
ZWEITER BAND
3

GRUNDBEGRIFFE DER
WAHRSCHEINLICHKEITS-
RECHNUNG

VON

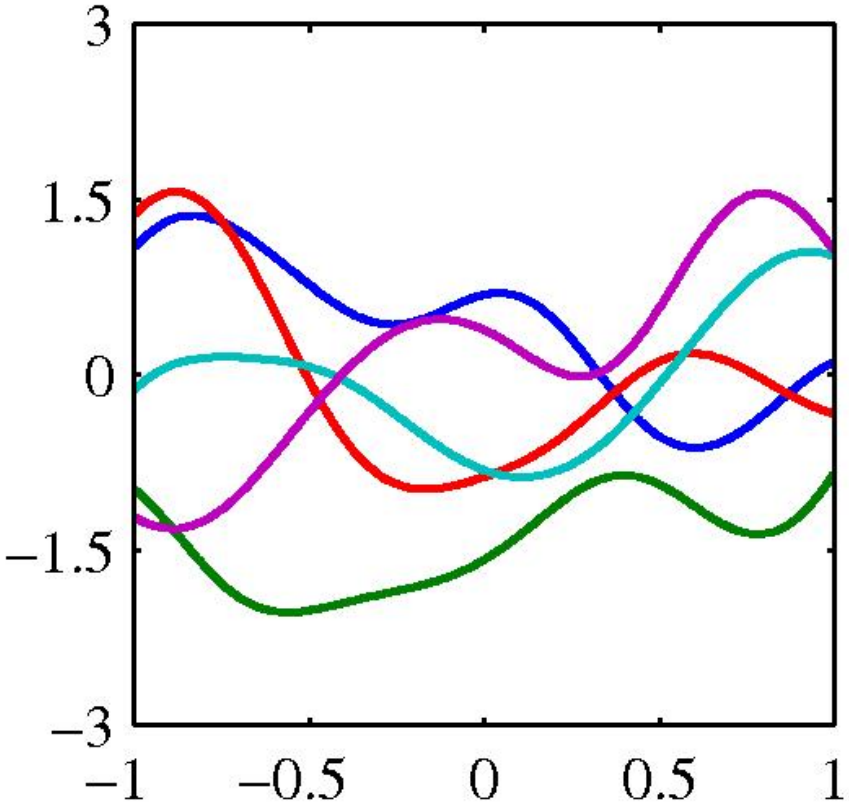
A. KOLMOGOROFF

BERLIN
VERLAG VON JULIUS SPRINGER
1933

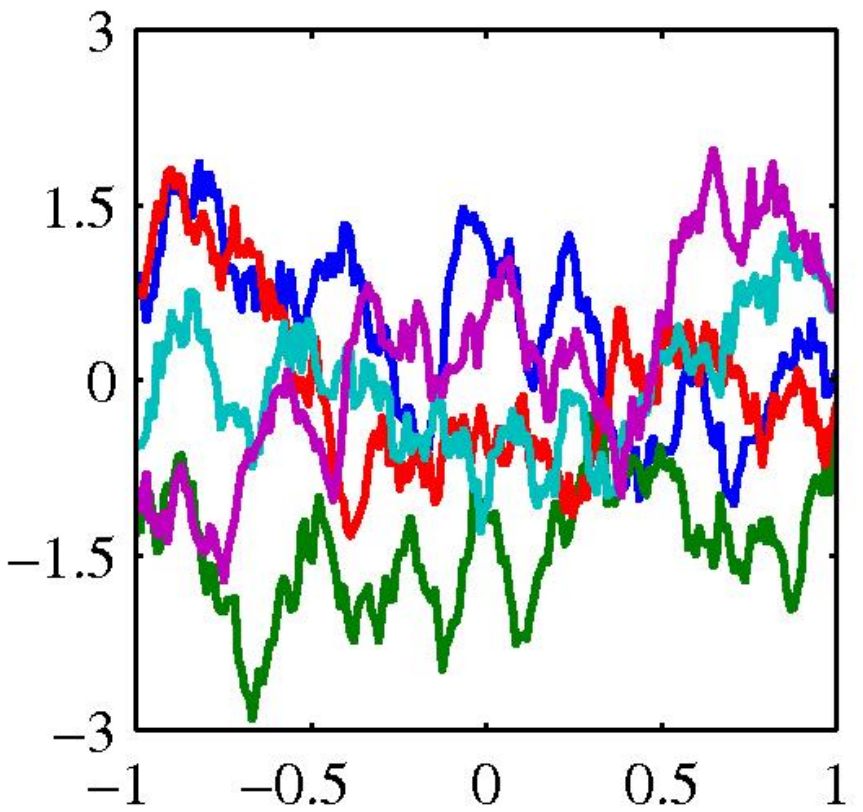
Foundations of the theory of probability

UNIVERSITY OF WATERLOO

Effect of kernel



$$\exp(-\|\mathbf{x}-\mathbf{x}'\|_2^2/\sigma)$$



$$\exp(-\|\mathbf{x}-\mathbf{x}'\|_2/\sigma)$$

Example: Linear Regression

unknown but deterministic

independent of each other for different x

$$y = \mathbf{x}^\top \mathbf{w} + \epsilon$$

z t $\mathcal{N}(0, \sigma^2)$

$x_1, x_2, \cdots, x_n$

$y_1, y_2, \cdots, y_n$

$y_i = x_i^\top w + \epsilon_i$

- Y is a Gaussian process

$$Y \sim \mathcal{GP}(m(\mathbf{x}), \kappa(\mathbf{x}, \mathbf{x}'))$$

$\mathbf{w}^\top \mathbf{x}$

$\sigma^2 \mathbf{1}_{\mathbf{x}=\mathbf{x}'}$

$\Sigma_y = \sigma^2 I$

$\mu_y = Xw$

Maximum Likelihood

- Having observed $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)$
- Need to estimate **w**
- Choose **w** that explains the observations best

$$\max_{\mathbf{w}} \Pr[(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)] \overset{\text{i.i.d.}}{=} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[- \frac{(y_i - \mathbf{w}^\top \mathbf{x}_i)^2}{2\sigma^2} \right]$$

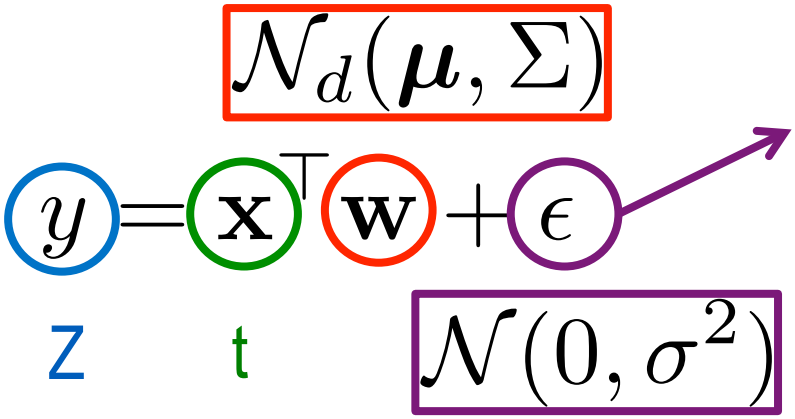
likelihood of data

$\log$

$$= \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi}\sigma} + - \frac{(y_i - w^\top x_i)^2}{2\sigma^2}$$

$$\Leftrightarrow \min_w \sum_i (y_i - w^\top x_i)^2 = \|Xw - y\|_2^2$$

Example: Bayesian Linear Regression



- Y is a Gaussian process

$$Y \sim \mathcal{GP}(m(\mathbf{x}), \kappa(\mathbf{x}, \mathbf{x}'))$$

$$\boldsymbol{\mu}^\top \mathbf{x}$$

$$\mathbf{x}^\top \Sigma \mathbf{x}' + \sigma^2 \mathbf{1}_{\mathbf{x}=\mathbf{x}'}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ y_n \end{pmatrix} = \begin{pmatrix} x_1^\top w + \epsilon_1 \\ x_2^\top w + \epsilon_2 \\ x_n^\top w + \epsilon_n \end{pmatrix} = \begin{pmatrix} x_1^\top \\ x_2^\top \\ x_n^\top \end{pmatrix} w + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_n \end{pmatrix}$$

Maximum A Posteriori

likelihood of data

prior

$$\max_{\mathbf{w}} \Pr[\mathbf{w}|(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)] = \frac{\Pr[(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)|\mathbf{w}] \cdot \Pr[\mathbf{w}]}{\Pr[(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)]}$$

posterior

normalization

- Different prior on **w** leads to different regularization
  - **w** ~ N(0, I) is equivalent as ridge regression
  - **w** ~ Lap(0, I) is equivalent as Lasso

Outline

- Gaussian Distributions
- Gaussian Process
- Gaussian Linear Regression
- Advanced

Abstract View

- Let $Z \sim \mathcal{GP}(m, \kappa)$ be a Gaussian process
- Having observed $(t_1, Z_{t_1}), (t_2, Z_{t_2}), \ldots, (t_n, Z_{t_n})$
- Need to predict Z_t

Familiar view

Feature map of k

$$Z = \varphi(t)^\top \mathbf{w} + m(t)$$

$$\mathcal{N}(\mathbf{0}, \mathbb{I})$$

- Equivalent in finite dimensions
- Incorrect but “intuitive” in infinite dimensions

Back to abstract

- $Z_{t_1}, Z_{t_2}, \ldots, Z_{t_n}, Z_t$ is jointly Gaussian $\mathcal{N}(\boldsymbol{\mu}, K)$

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} \qquad K = \begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix}$$

- What is the distribution of $Z_t | Z_{t_1}, Z_{t_2}, \ldots, Z_{t_n}$?

Recap: Important facts

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right)$$

joint = marginal ✖ conditional

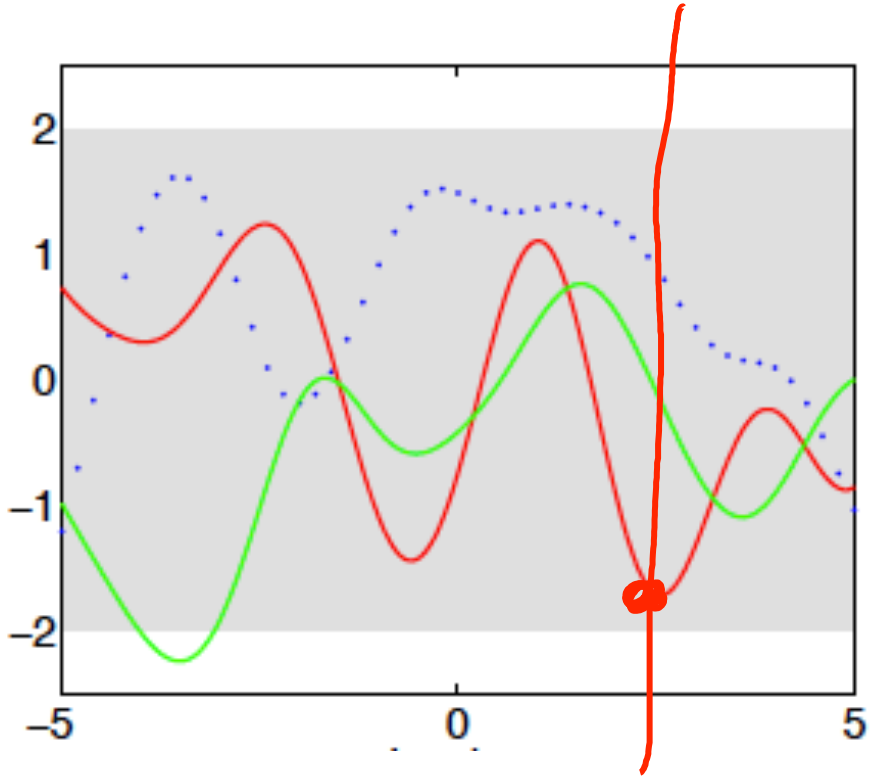
$$X_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \Sigma_{11})$$

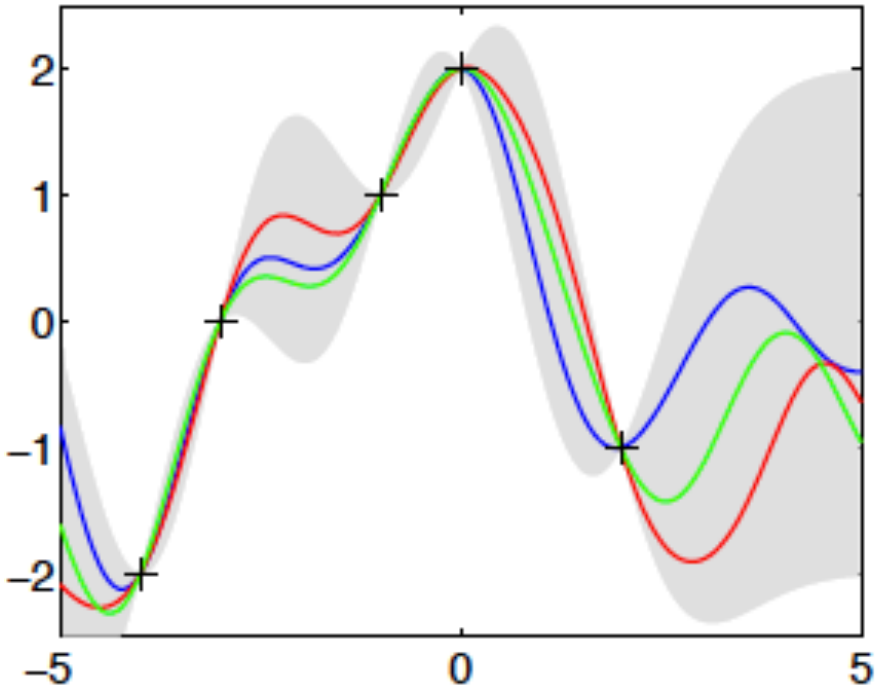
$$\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$$

$$X_2 | X_1 \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$$

$$\boldsymbol{\mu}_2 + \Sigma_{21}\Sigma_{11}^{-1}(X_1 - \boldsymbol{\mu}_1)$$

Recap: Example





Outline

- Gaussian Distributions
- Gaussian Process
- Gaussian Linear Regression
- Advanced

Gaussian process classification

- Binary label Y generated by

$$\Pr(Y_t = 1) = \frac{1}{1 + \exp(Z_t)}$$

- Z_t is a Gaussian process (real-valued)

- Given observations $(t_1, Y_{t_1}), \ldots, (t_n, Y_{t_n})$

- Need to predict

integrate out

latent variable

$$\Pr(Y_t = 1 | t, t_i, Y_{t_i}) = \int \frac{1}{1+\exp(z_t)} p(z_t | t, t_i, Y_{t_i}) \mathrm{d}z_t$$