Gaussian Processes – Literature Study 2019

Ivan De Boi

Op3Mech
University of Antwerp

Outline

Gaussian Processes

Bayesian Inference

Kernels

Classification

Sparse Approximations

Bayesian Optimization – Active Learning

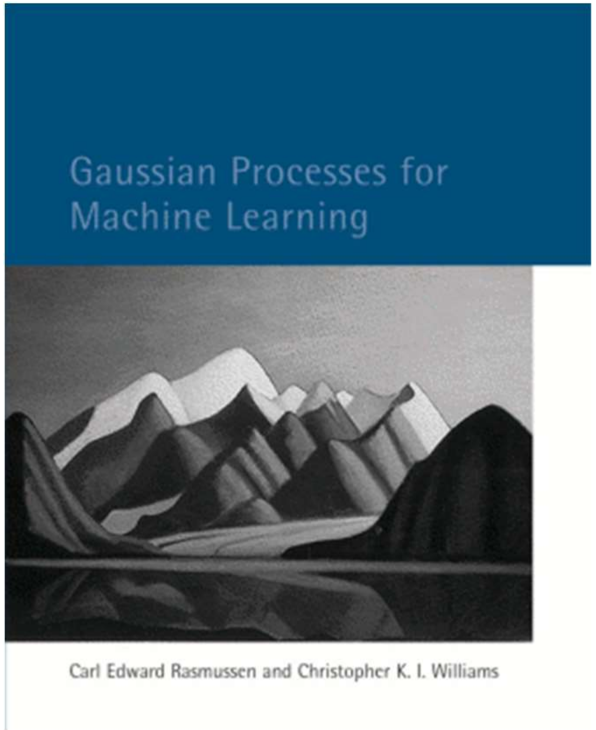
Bayesian Quadrature

Deep Gaussian Processes

Stochastic Differential Equations

Software

What are Gaussian Processes?



A collection of random variables,
any finite subset of which
is a multivariate Gaussian distribution.

http://www.gaussianprocess.org/

What are Gaussian Processes?

Gaussian Process Summer Schools

iudicium posterium discipulus est prioris

University of Sheffield since 2013

http://gpss.cc/

Notation: Wiki & Math vs. Machine Learning & Statistics

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

PDF: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$

$\mathrm{E}[X] = \int x\, f(x) dx$, $\mathrm{E}[X] = \sum_{i=0}^{\infty} x_i p_i$

$X \sim \mathrm{Ber}(p)$
PMF: $Pr(X = k|p) = p^k (1-p)^{1-k}$

$X = (X_1, \dots, Xn)$

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

PDF: $p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$

$\mathrm{E}[X] = \int x\, p(x) dx$, $\mathrm{E}[X] = \sum_{i=0}^{\infty} x_i p_i$

$X \sim \mathrm{Ber}(\theta)$
PMF: $Pr(X = k|\theta) = \theta^k (1-\theta)^{1-k}$

$X = (\mathrm{x}_1, \dots, \mathrm{x}_n)$

Notation: Wiki & Math vs. Machine Learning & Statistics

X matrix of inputs $\in \mathbb{R}^{n \times D}$

X=$\{X_1, \dots, Xn\}$

X=$\{\mathrm{x}_1, \dots, \mathrm{x}_n\}$

x=$\{\mathrm{x}_1, \dots, \mathrm{x}_n\}$

x_i vector $\in \mathbb{R}^D$
x_i element i of all inputs X

x_i element i of input x

x_i^j = element i of input j

$\mathcal{D} = \{(\mathrm{x}_i, y_i) | i = 1, \dots, n\} = (X, \mathrm{y})$

$y_i \in \mathbb{R}$, observation i of input x_i
$\mathrm{y} \in \mathbb{R}^n$, vector of observations

X, y training points (data)

x_*, y_* test points (predictions)
X_*, y_* test points (predictions)

μ vector, μ scalar

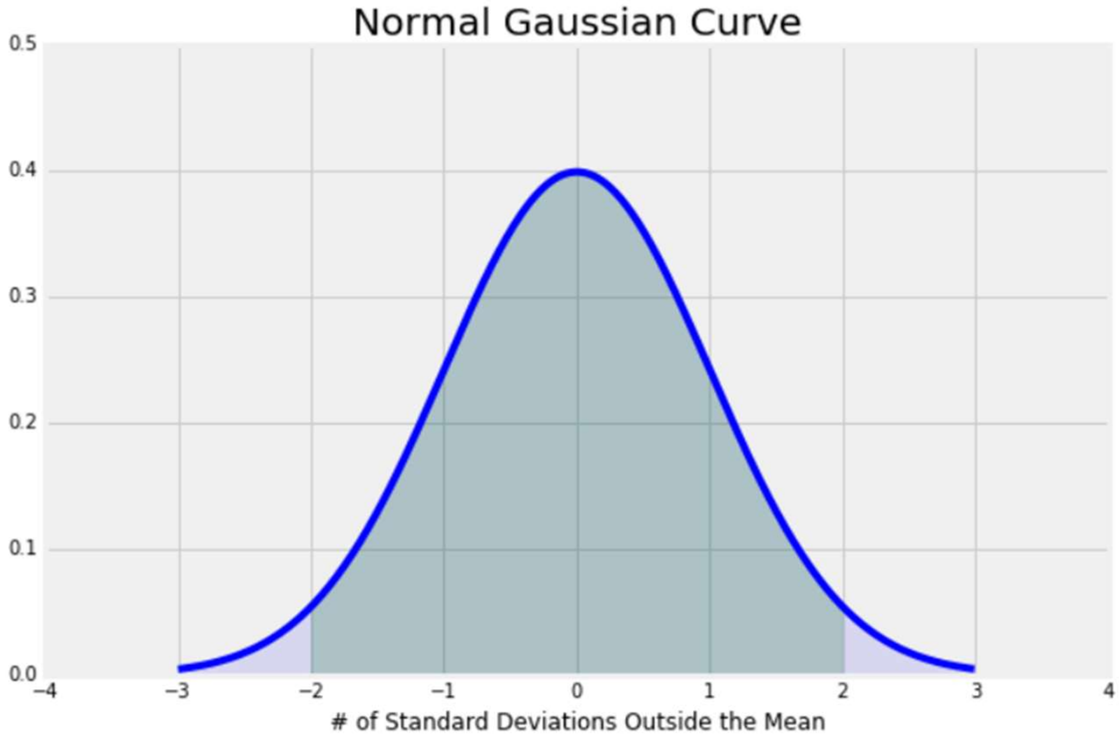
Gaussian aka Normal Distribution

Central Limit Theorem

$$Y = aX + b$$
$$Y \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

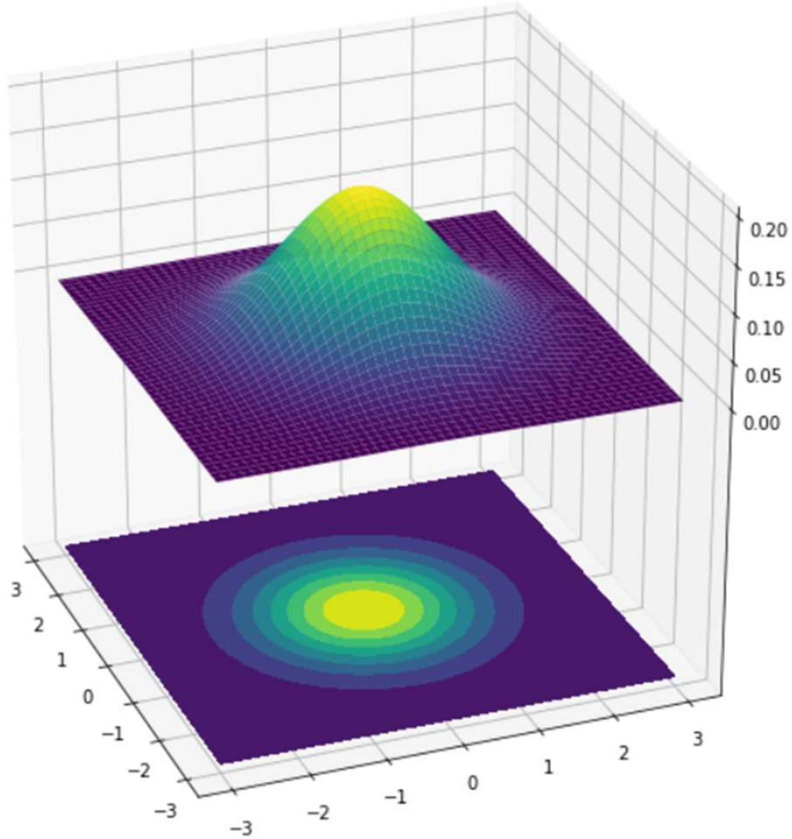
$$X \sim \mathcal{N}(\mu = 0, \sigma^2 = 1)$$



Bivariate or joint Gaussian distribution

$$p(X) = p(\mathrm{x}_1, \mathrm{x}_2) = \frac{1}{\sqrt{(2\pi)^2|\Sigma|}} e^{-\frac{1}{2}(X-\mu)^{\mathrm{T}}\Sigma^{-1}(X-\mu)}$$

$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



$$\mathrm{E}[X] = \int_{-\infty}^{+\infty} x\, p(x) dx\, , \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = \begin{bmatrix} \mathrm{E}[\mathrm{x}_1] \\ \mathrm{E}[\mathrm{x}_2] \end{bmatrix}$$

$\mathrm{Var}[X] = \mathrm{E}[(X - \mathrm{E}[X])^2] = \sigma^2$ for univariate

$\mathrm{Cov}[\mathrm{x}_1, \mathrm{x}_2] = \mathrm{E}[(\mathrm{x}_1 - \mu_1)(\mathrm{x}_2 - \mu_2)]$

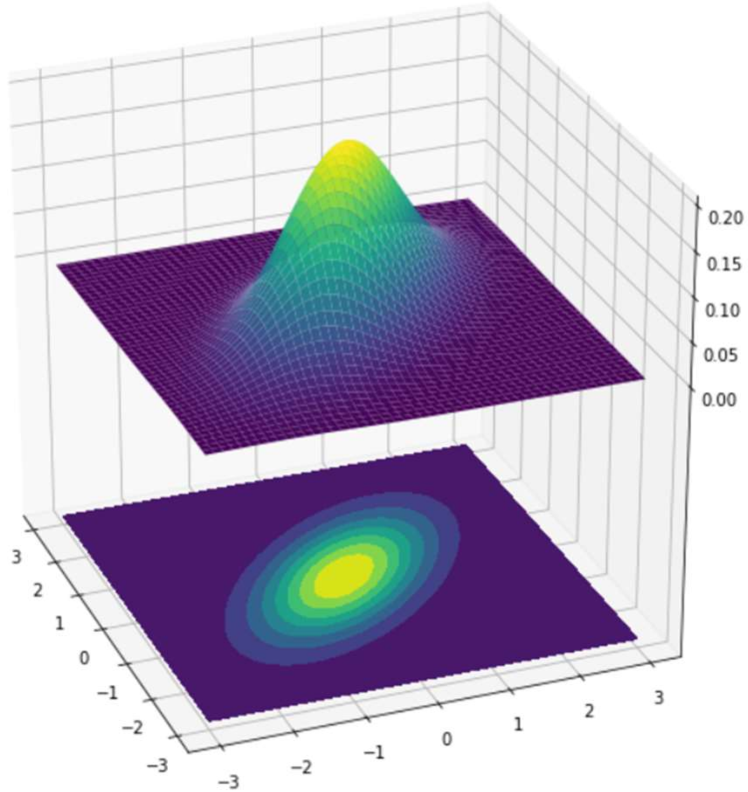
$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} = \begin{bmatrix} \mathrm{Cov}[\mathrm{x}_1, \mathrm{x}_1] & \mathrm{Cov}[\mathrm{x}_1, \mathrm{x}_2] \\ \mathrm{Cov}[\mathrm{x}_2, \mathrm{x}_1] & \mathrm{Cov}[\mathrm{x}_2, \mathrm{x}_2] \end{bmatrix}$$

$$= \begin{bmatrix} {\sigma_1}^2 & \mathrm{Cov}[\mathrm{x}_1, \mathrm{x}_2] \\ \mathrm{Cov}[\mathrm{x}_2, \mathrm{x}_1] & {\sigma_2}^2 \end{bmatrix} = \begin{bmatrix} {\sigma_1}^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & {\sigma_2}^2 \end{bmatrix}$$

Bivariate or joint Gaussian distribution

$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} 1 & 0.6 \\ 0.6 & 1 \end{bmatrix}$$

$$p(X) = p(\mathrm{x}_1, \mathrm{x}_2) = \frac{1}{\sqrt{(2\pi)^2|\Sigma|}} e^{-\frac{1}{2}(X-\mu)^{\mathrm{T}}\Sigma^{-1}(X-\mu)}$$



$$\mathrm{E}[X] = \int_{-\infty}^{+\infty} x\, p(x) dx, \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = \begin{bmatrix} \mathrm{E}[\mathrm{x}_1] \\ \mathrm{E}[\mathrm{x}_2] \end{bmatrix}$$

$\mathrm{Var}[X] = \mathrm{E}[(X - \mathrm{E}[X])^2] = \sigma^2$ for univariate

$\mathrm{Cov}[\mathrm{x}_1, \mathrm{x}_2] = \mathrm{E}[(\mathrm{x}_1 - \mu_1)(\mathrm{x}_2 - \mu_2)]$

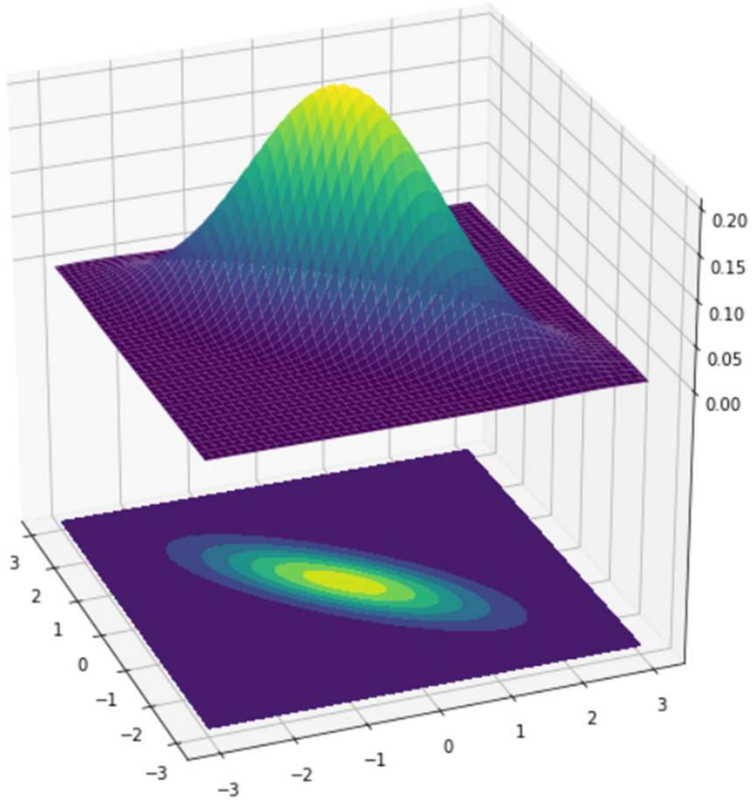
$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} = \begin{bmatrix} \mathrm{Cov}[\mathrm{x}_1, \mathrm{x}_1] & \mathrm{Cov}[\mathrm{x}_1, \mathrm{x}_2] \\ \mathrm{Cov}[\mathrm{x}_2, \mathrm{x}_1] & \mathrm{Cov}[\mathrm{x}_2, \mathrm{x}_2] \end{bmatrix}$$

$$= \begin{bmatrix} {\sigma_1}^2 & \mathrm{Cov}[\mathrm{x}_1, \mathrm{x}_2] \\ \mathrm{Cov}[\mathrm{x}_2, \mathrm{x}_1] & {\sigma_2}^2 \end{bmatrix} = \begin{bmatrix} {\sigma_1}^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & {\sigma_2}^2 \end{bmatrix}$$

Bivariate or joint Gaussian distribution

$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} 1 & -0.8 \\ -0.8 & 1 \end{bmatrix}$$

$$p(X) = p(\mathrm{x}_1, \mathrm{x}_2) = \frac{1}{\sqrt{(2\pi)^2 |\Sigma|}} e^{-\frac{1}{2}(X-\mu)^{\mathrm{T}} \Sigma^{-1} (X-\mu)}$$



$$\mathrm{E}[X] = \int_{-\infty}^{+\infty} x\, p(x) dx \,, \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = \begin{bmatrix} \mathrm{E}[\mathrm{x}_1] \\ \mathrm{E}[\mathrm{x}_2] \end{bmatrix}$$

$\mathrm{Var}[X] = \mathrm{E}[(X - \mathrm{E}[X])^2] = \sigma^2$ for univariate

$$\mathrm{Cov}[\mathrm{x}_1, \mathrm{x}_2] = \mathrm{E}[(\mathrm{x}_1 - \mu_1)(\mathrm{x}_2 - \mu_2)]$$

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} = \begin{bmatrix} \mathrm{Cov}[\mathrm{x}_1, \mathrm{x}_1] & \mathrm{Cov}[\mathrm{x}_1, \mathrm{x}_2] \\ \mathrm{Cov}[\mathrm{x}_2, \mathrm{x}_1] & \mathrm{Cov}[\mathrm{x}_2, \mathrm{x}_2] \end{bmatrix}$$

$$= \begin{bmatrix} {\sigma_1}^2 & \mathrm{Cov}[\mathrm{x}_1, \mathrm{x}_2] \\ \mathrm{Cov}[\mathrm{x}_2, \mathrm{x}_1] & {\sigma_2}^2 \end{bmatrix} = \begin{bmatrix} {\sigma_1}^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & {\sigma_2}^2 \end{bmatrix}$$

Bivariate or joint Gaussian distribution

$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} 1 & 0.6 \\ 0.6 & 1 \end{bmatrix}$$

Conditional Bivariate Gaussian distribution

$$X = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

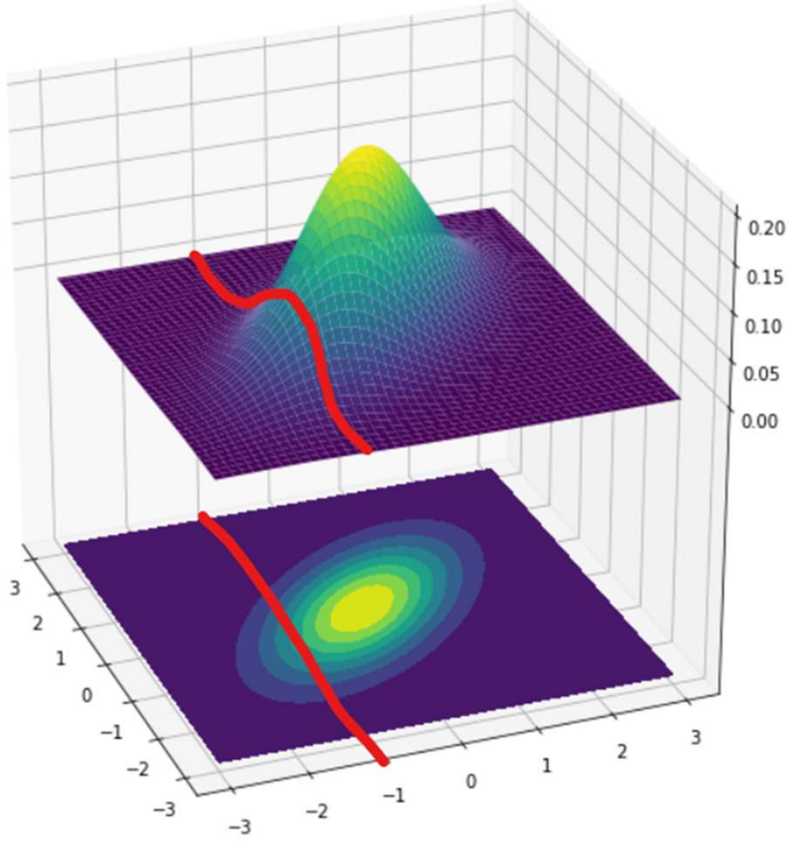
$$x_1 \sim \mathcal{N}(\mu_1 = 0, \sigma_1{}^2 = \Sigma_{11} = 1), \text{ marginal}$$

$$p(x_1) = \int p(x_1, x_2) dx_2$$

$$(x_1 | x_2 = a) \sim \mathcal{N}(\bar{\mu}, \bar{\Sigma})$$

$$\bar{\mu} = \mu_1 + \Sigma_{12} \Sigma_{22}{}^{-1} (a - \mu_2), \ \bar{\Sigma} = \Sigma_{11} - \Sigma_{12} \Sigma_{22}{}^{-1} \Sigma_{21}$$

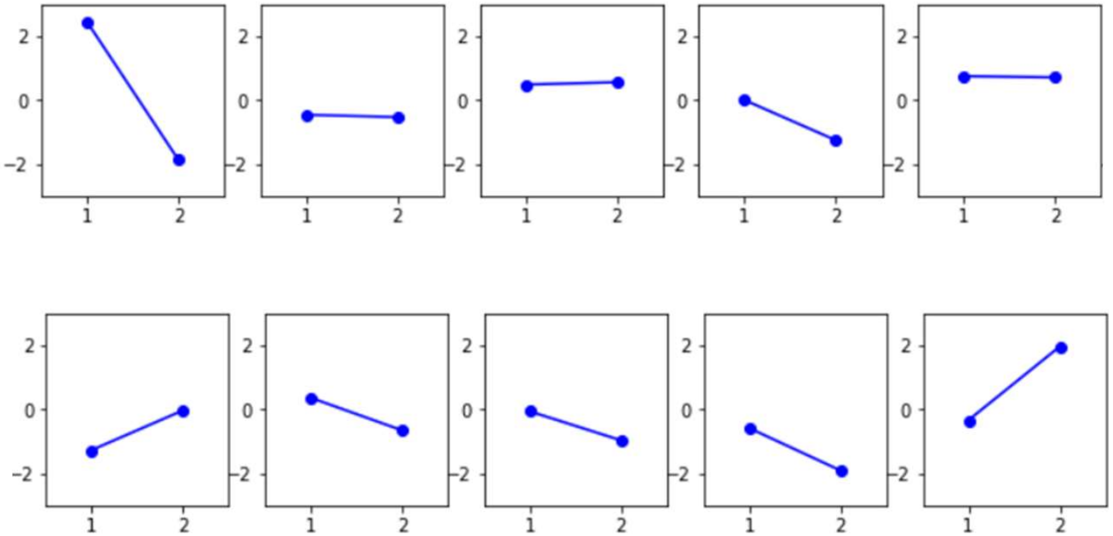
$$(x_1 | x_2 = -1) \sim \mathcal{N}(\bar{\mu} = -0.4, \bar{\Sigma} = 0.64), \text{ conditional}$$

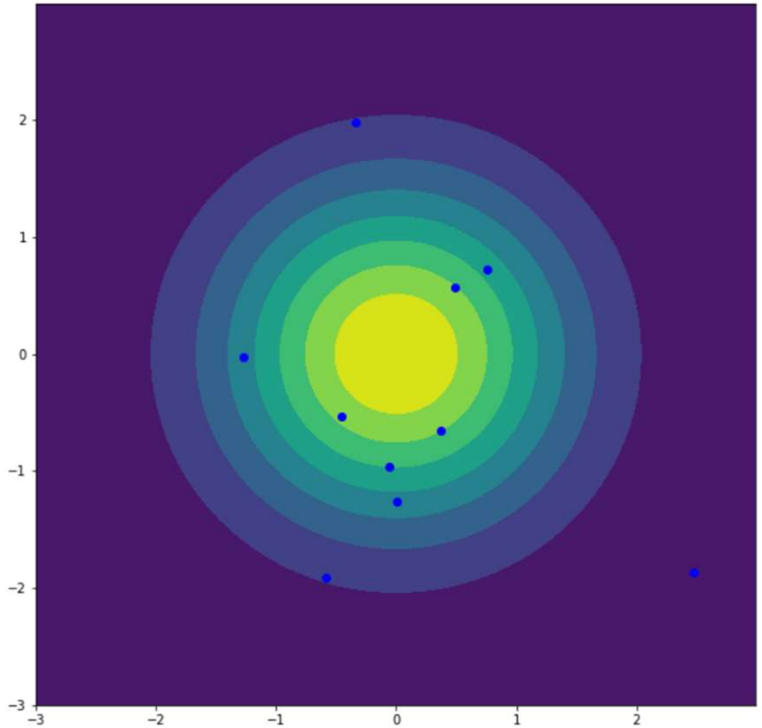


Bivariate or joint Gaussian distribution

$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Some samples, no correlation

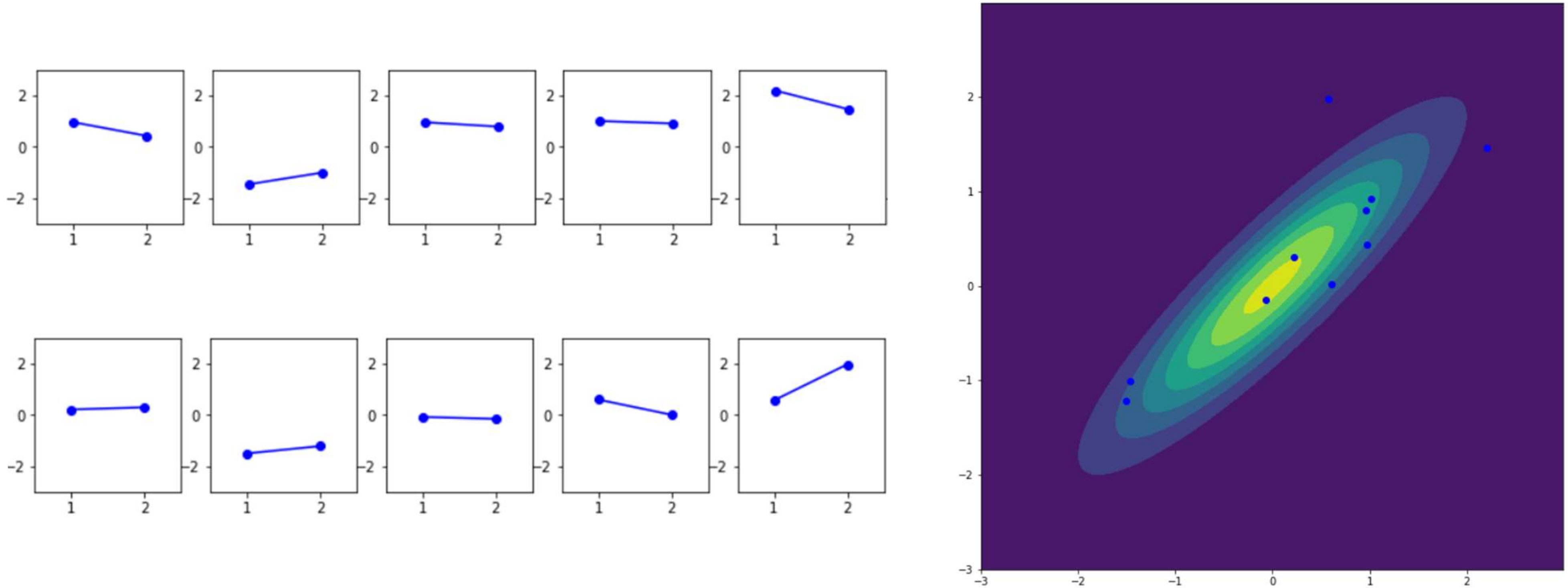




Bivariate or joint Gaussian distribution

$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}$$

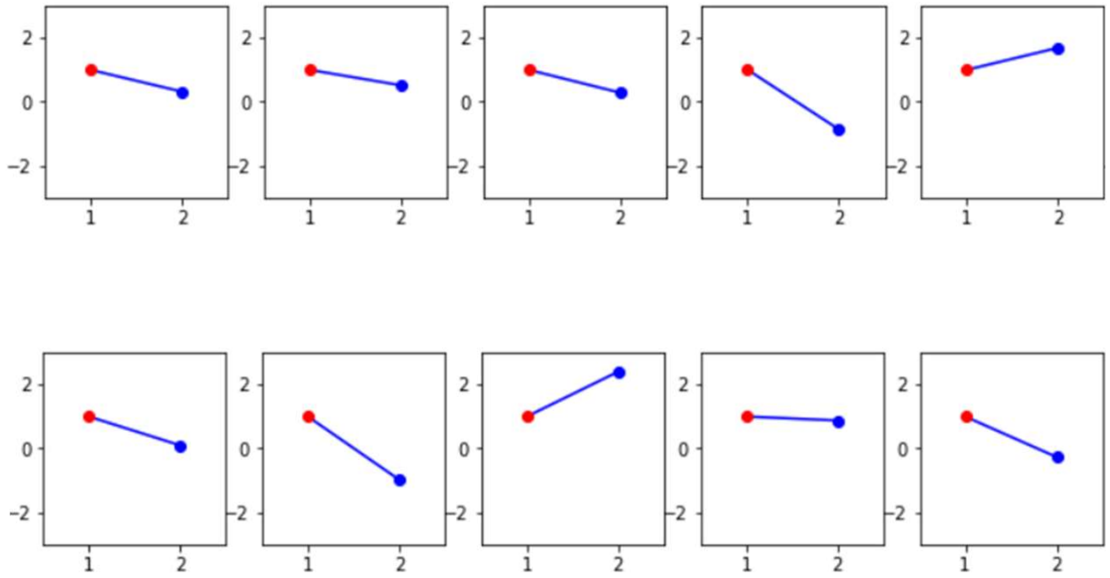
Some samples, highly correlated

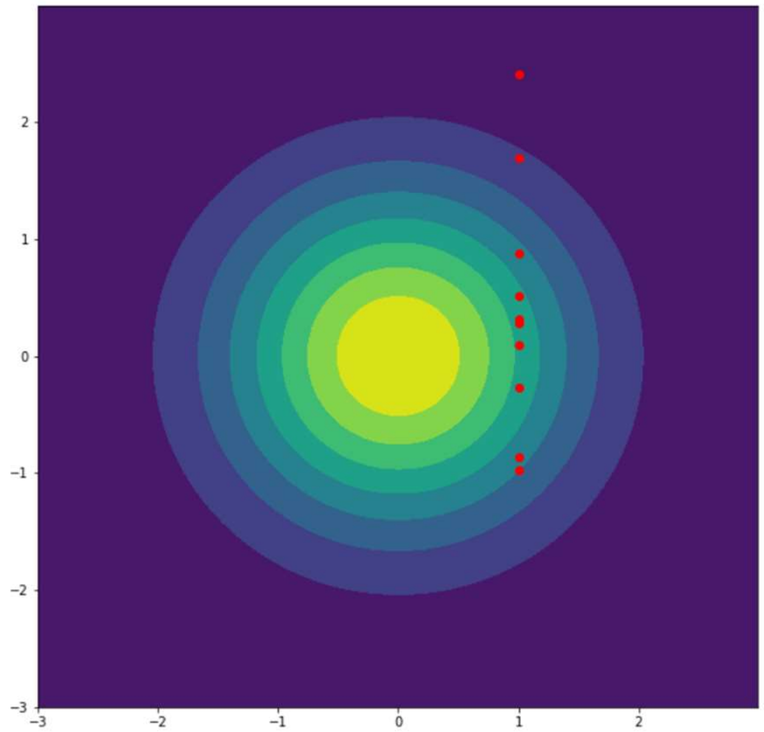


Bivariate or joint Gaussian distribution

$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Some samples, no correlation, conditional $x_1 = 1$

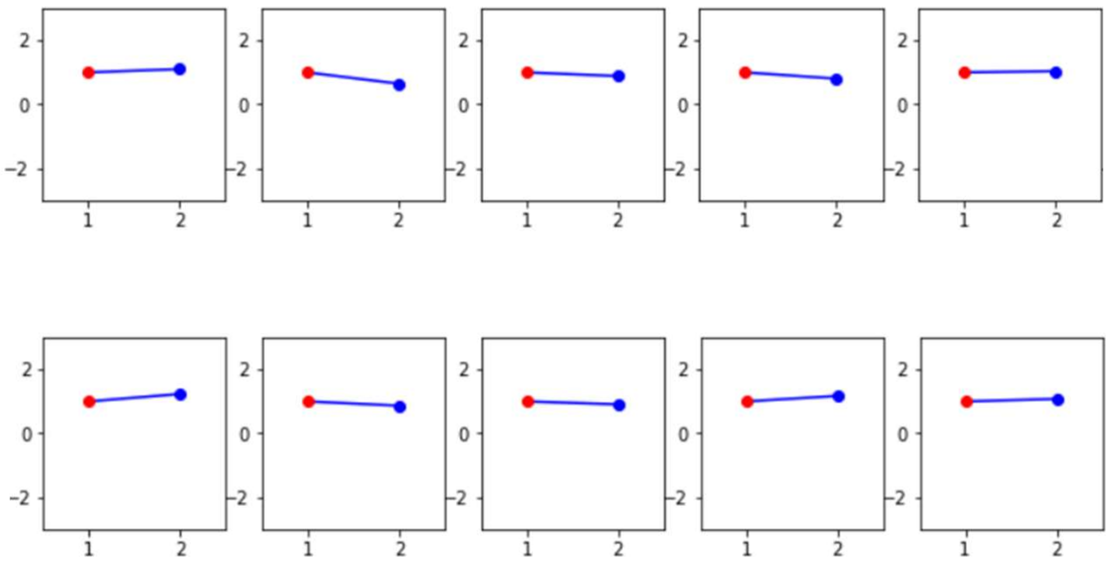


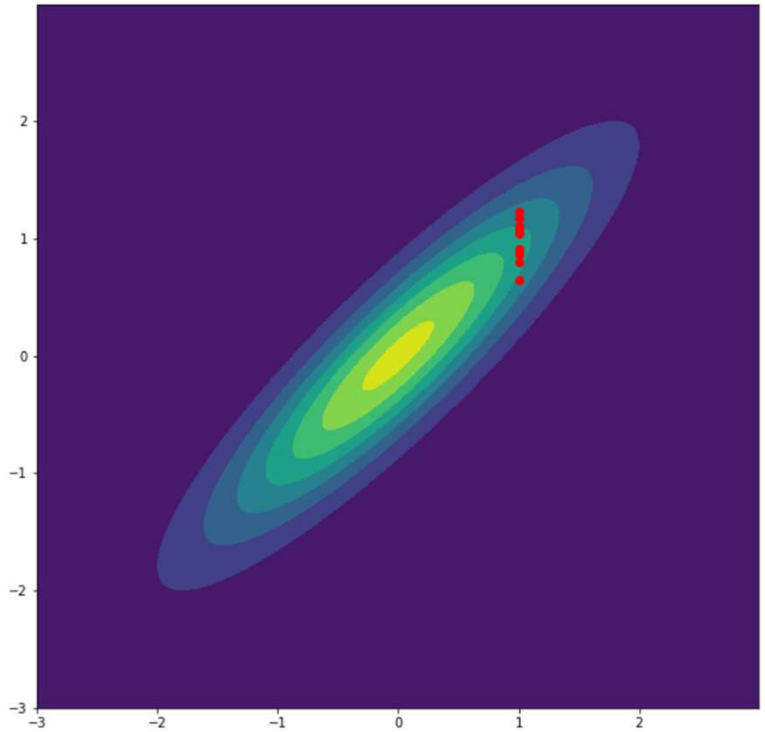


Bivariate or joint Gaussian distribution

$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}$$

Some samples, highly correlated, conditional $x_1 = 1$





Multivariate Gaussian distribution

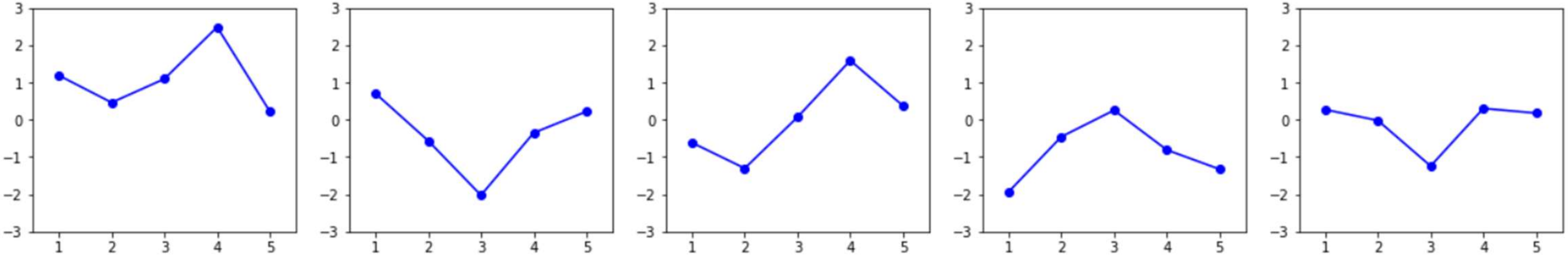
$$p(X) = p(\mathrm{x}_1, \mathrm{x}_2, \mathrm{x}_3, \mathrm{x}_4, \mathrm{x}_5) = \frac{1}{\sqrt{(2\pi)^2|\Sigma|}} e^{-\frac{1}{2}(X-\mu)^{\mathrm{T}}\Sigma^{-1}(X-\mu)}$$

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \\ \mu_5 \end{bmatrix}, \Sigma = \begin{bmatrix} {\sigma_1}^2 & \mathrm{Cov}[\mathrm{x}_1, \mathrm{x}_2] & \mathrm{Cov}[\mathrm{x}_1, \mathrm{x}_3] & \mathrm{Cov}[\mathrm{x}_1, \mathrm{x}_4] & \mathrm{Cov}[\mathrm{x}_1, \mathrm{x}_5] \\ \mathrm{Cov}[\mathrm{x}_2, \mathrm{x}_1] & {\sigma_2}^2 & \mathrm{Cov}[\mathrm{x}_2, \mathrm{x}_3] & \mathrm{Cov}[\mathrm{x}_2, \mathrm{x}_4] & \mathrm{Cov}[\mathrm{x}_2, \mathrm{x}_5] \\ \mathrm{Cov}[\mathrm{x}_3, \mathrm{x}_1] & \mathrm{Cov}[\mathrm{x}_3, \mathrm{x}_2] & {\sigma_3}^2 & \mathrm{Cov}[\mathrm{x}_3, \mathrm{x}_4] & \mathrm{Cov}[\mathrm{x}_3, \mathrm{x}_5] \\ \mathrm{Cov}[\mathrm{x}_4, \mathrm{x}_1] & \mathrm{Cov}[\mathrm{x}_4, \mathrm{x}_2] & \mathrm{Cov}[\mathrm{x}_4, \mathrm{x}_3] & {\sigma_4}^2 & \mathrm{Cov}[\mathrm{x}_4, \mathrm{x}_5] \\ \mathrm{Cov}[\mathrm{x}_5, \mathrm{x}_1] & \mathrm{Cov}[\mathrm{x}_5, \mathrm{x}_2] & \mathrm{Cov}[\mathrm{x}_5, \mathrm{x}_3] & \mathrm{Cov}[\mathrm{x}_5, \mathrm{x}_4] & {\sigma_5}^2 \end{bmatrix}$$

Multivariate Gaussian distribution

$$p(X) = p(\mathrm{x}_1, \mathrm{x}_2, \mathrm{x}_3, \mathrm{x}_4, \mathrm{x}_5) = \frac{1}{\sqrt{(2\pi)^2|\Sigma|}} e^{-\frac{1}{2}(X-\mu)^{\mathrm{T}}\Sigma^{-1}(X-\mu)}$$

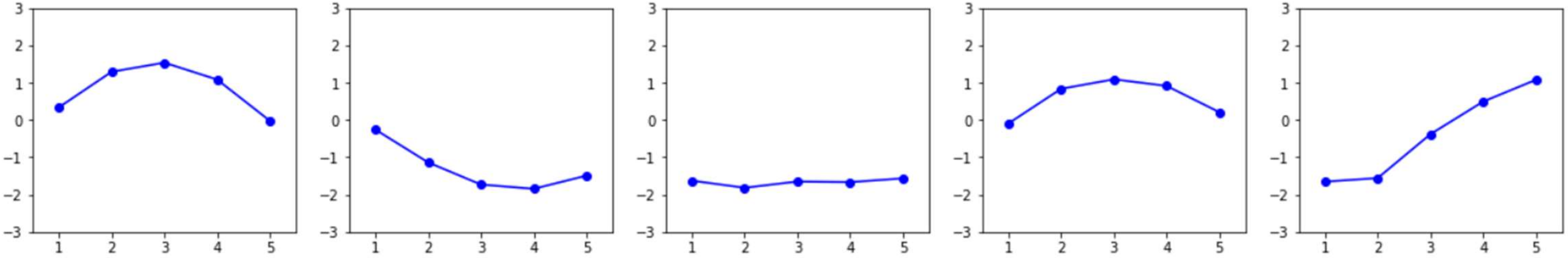
$$\mu = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$



Multivariate Gaussian distribution

$$p(X) = p(\mathrm{x}_1, \mathrm{x}_2, \mathrm{x}_3, \mathrm{x}_4, \mathrm{x}_5) = \frac{1}{\sqrt{(2\pi)^2|\Sigma|}} e^{-\frac{1}{2}(X-\mu)^{\mathrm{T}}\Sigma^{-1}(X-\mu)}$$

$$\mu = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} 1 & 0.8 & 0.5 & 0.1 & 0 \\ 0.8 & 1 & 0.8 & 0.5 & 0.1 \\ 0.5 & 0.8 & 1 & 0.8 & 0.5 \\ 0.1 & 0.5 & 0.8 & 1 & 0.8 \\ 0 & 0.1 & 0.5 & 0.8 & 1 \end{bmatrix}$$



Multivariate Gaussian distribution

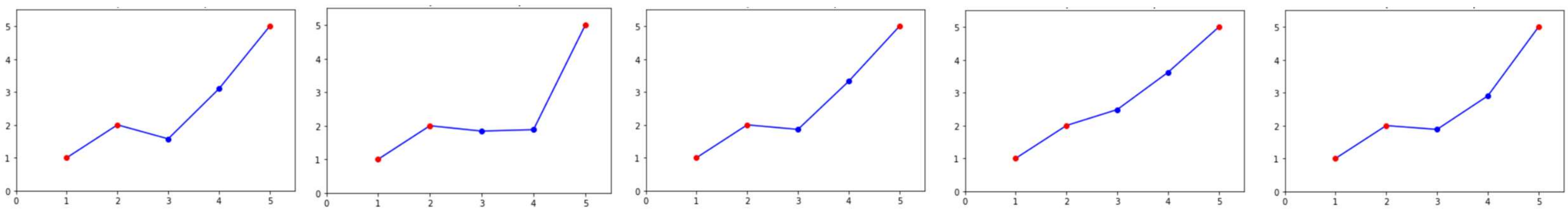
Three observations: $x_1 = 1, x_2 = 2, x_5 = 5 \Rightarrow (x_3, x_4 | x_1, x_2, x_5) \sim \mathcal{N}(\bar{\mu}, \bar{\Sigma})$

$\bar{\mu} = \mu_1 + \Sigma_{12}\Sigma_{22}{}^{-1}(x - \mu_2)$, $\bar{\Sigma} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}{}^{-1}\,\Sigma_{21}$

subscript 1 for unobserved (x_3, x_4), a.k.a. test points

subscript 2 for observations (x_1, x_2, x_5), a.k.a. data, a.k.a. training points

$\bar{\mu} = \mu_* + K_*{}^{T} K^{-1}(y - \mu)$, $\bar{\Sigma} = K_{**} - K_*{}^{T} K^{-1}\, K_*$

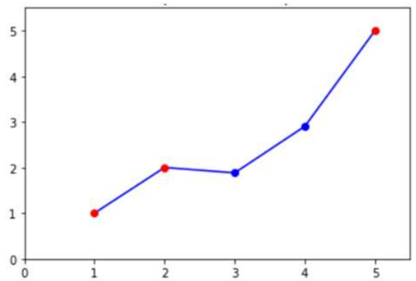


Multivariate Gaussian distribution

Three observations: $x_1 = 1, x_2 = 2, x_5 = 5 \Rightarrow (x_3, x_4 | x_1, x_2, x_5) \sim \mathcal{N}(\bar{\mu}, \bar{\Sigma})$

$$\mu_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, x = \begin{bmatrix} 1 \\ 2 \\ 5 \end{bmatrix}, \Sigma_{22} = \begin{bmatrix} 1 & 0.5 & 0 \\ 0.5 & 1 & 0.5 \\ 0 & 0.5 & 1 \end{bmatrix}, \mu_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma_{11} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$

$$\Sigma_{12} = \begin{bmatrix} 0.06 & 0 \\ 0.5 & 0.06 \\ 0.06 & 0.5 \end{bmatrix}, \Sigma_{21} = \begin{bmatrix} 0.06 & 0.5 & 0.06 \\ 0 & 0.06 & 0,5 \end{bmatrix}$$



$$\bar{\mu} = \mu_1 + \Sigma_{12} {\Sigma_{22}}^{-1} (x - \mu_2) \text{ , } \bar{\Sigma} = \Sigma_{11} - \Sigma_{12} {\Sigma_{22}}^{-1} \Sigma_{21}$$

$$\bar{\mu} = \begin{bmatrix} 1.3 \\ 2.6 \end{bmatrix}, \bar{\Sigma} = \begin{bmatrix} 0.7 & 0.4 \\ 0.4 & 0.7 \end{bmatrix}$$

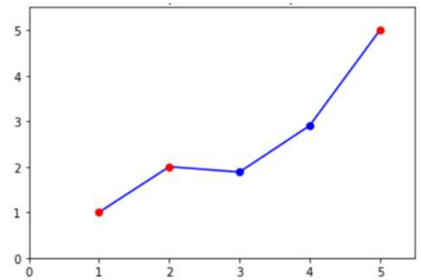
$(x_3 | x_1, x_2, x_5) \sim \mathcal{N}(\mu = 1.3 \,; \sigma^2 = 0.7), (x_4 | x_1, x_2, x_5) \sim \mathcal{N}(\mu = 2.6 \,; \sigma^2 = 0.7)$

Multivariate Gaussian distribution

Three observations: $x_1 = 1, x_2 = 2, x_5 = 5 \Rightarrow (x_3, x_4 | x_1, x_2, x_5) \sim \mathcal{N}(\bar{\mu}, \bar{\Sigma})$

$$\mu = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, y = \begin{bmatrix} 1 \\ 2 \\ 5 \end{bmatrix}, K = \begin{bmatrix} 1 & 0.5 & 0 \\ 0.5 & 1 & 0.5 \\ 0 & 0.5 & 1 \end{bmatrix}, \mu_* = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, K_{**} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$

$$K_* = \begin{bmatrix} 0.06 & 0 \\ 0.5 & 0.06 \\ 0.06 & 0.5 \end{bmatrix}, {K_*}^{\mathrm{T}} = \begin{bmatrix} 0.06 & 0.5 & 0.06 \\ 0 & 0.06 & 0{,}5 \end{bmatrix}$$



$$\bar{\mu} = \mu_* + {K_*}^{\mathrm{T}} K^{-1} (y - \mu) \,,\ \bar{\Sigma} = K_{**} - {K_*}^{\mathrm{T}} K^{-1} K_*$$

$$\bar{\mu} = \begin{bmatrix} 1.3 \\ 2.6 \end{bmatrix}, \bar{\Sigma} = \begin{bmatrix} 0.7 & 0.4 \\ 0.4 & 0.7 \end{bmatrix}$$

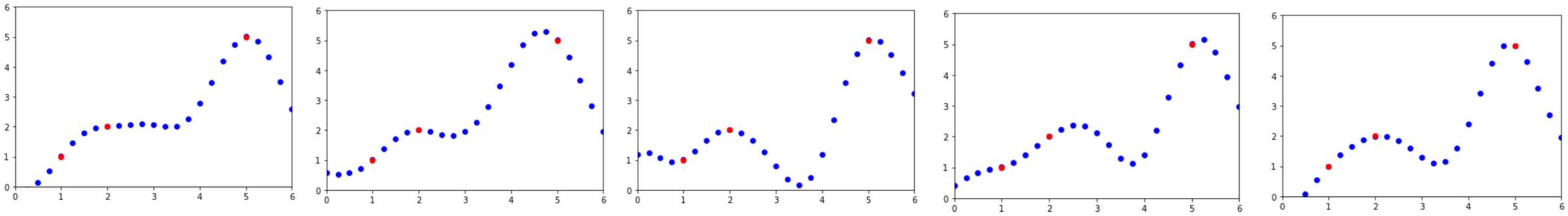
$$(x_3 | x_1, x_2, x_5) \sim \mathcal{N}(\mu = 1.3 \,;\, \sigma^2 = 0.7), (x_4 | x_1, x_2, x_5) \sim \mathcal{N}(\mu = 2.6 \,;\, \sigma^2 = 0.7)$$

Multivariate Gaussian distribution

Three observations: $x_1 = 1, x_2 = 2, x_5 = 5 \Rightarrow (x_3, x_4, x_6, \dots, x_{28} | x_1, x_2, x_5) \sim \mathcal{N}(\bar{\mu}, \bar{\Sigma})$

3 data points, 25 test points $\Rightarrow$ 28 dimensional Gaussian distribution

5 samples

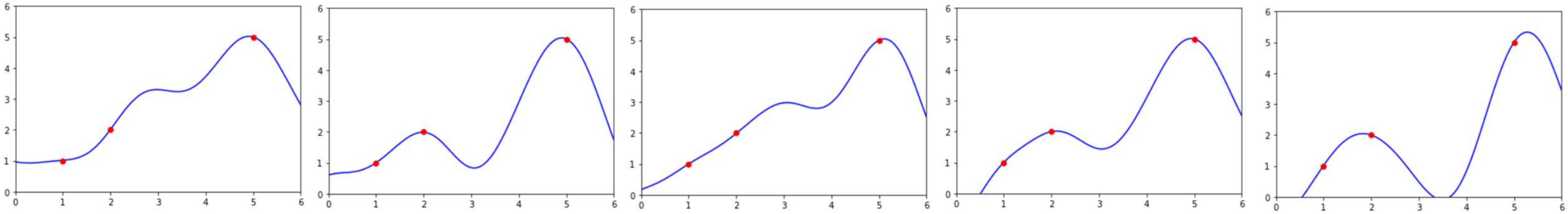


Gaussian Process

Three observations: $x_1 = 1, x_2 = 2, x_5 = 5 \Rightarrow (x_3, x_4, x_6, \ldots, x_{28}, \ldots | x_1, x_2, x_5) \sim \mathcal{N}(\bar{\mu}, \bar{\Sigma})$

Infinite amount of test points, $\bar{\mu}$ and $\bar{\Sigma}$ become functions

5 samples $f \sim \mathcal{GP}(\mu(x), K(x, x'))$

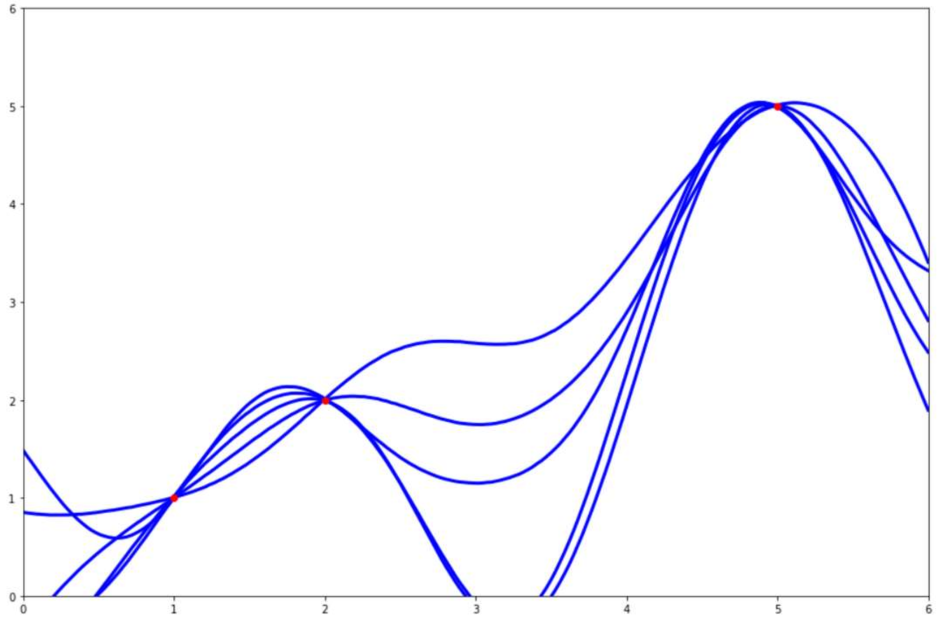


Gaussian Process

Three observations: $x_1 = 1, x_2 = 2, x_5 = 5 \Rightarrow (x_3, x_4, x_6, \ldots, x_{28}, \ldots | x_1, x_2, x_5) \sim \mathcal{N}(\bar{\mu}, \bar{\Sigma})$

Infinite amount of test points, $\bar{\mu}$ and $\bar{\Sigma}$ become functions

5 samples $f \sim \mathcal{GP}(\mu(x), K(x, x'))$

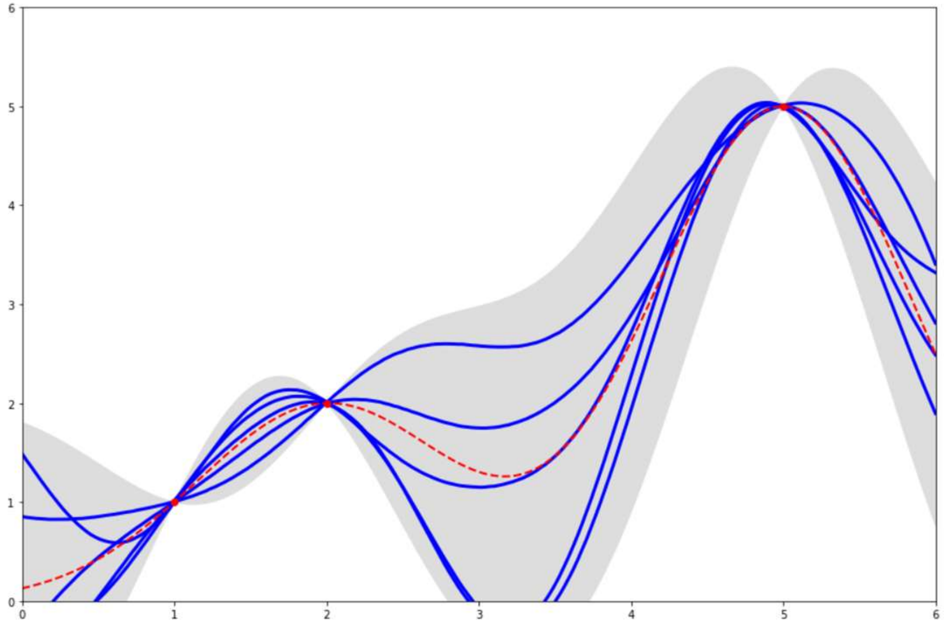


Gaussian Process

Three observations: $x_1 = 1, x_2 = 2, x_5 = 5 \Rightarrow (x_3, x_4, x_6, \dots, x_{28}, \dots | x_1, x_2, x_5) \sim \mathcal{N}(\bar{\mu}, \bar{\Sigma})$

Infinite amount of test points, $\bar{\mu}$ and $\bar{\Sigma}$ become functions

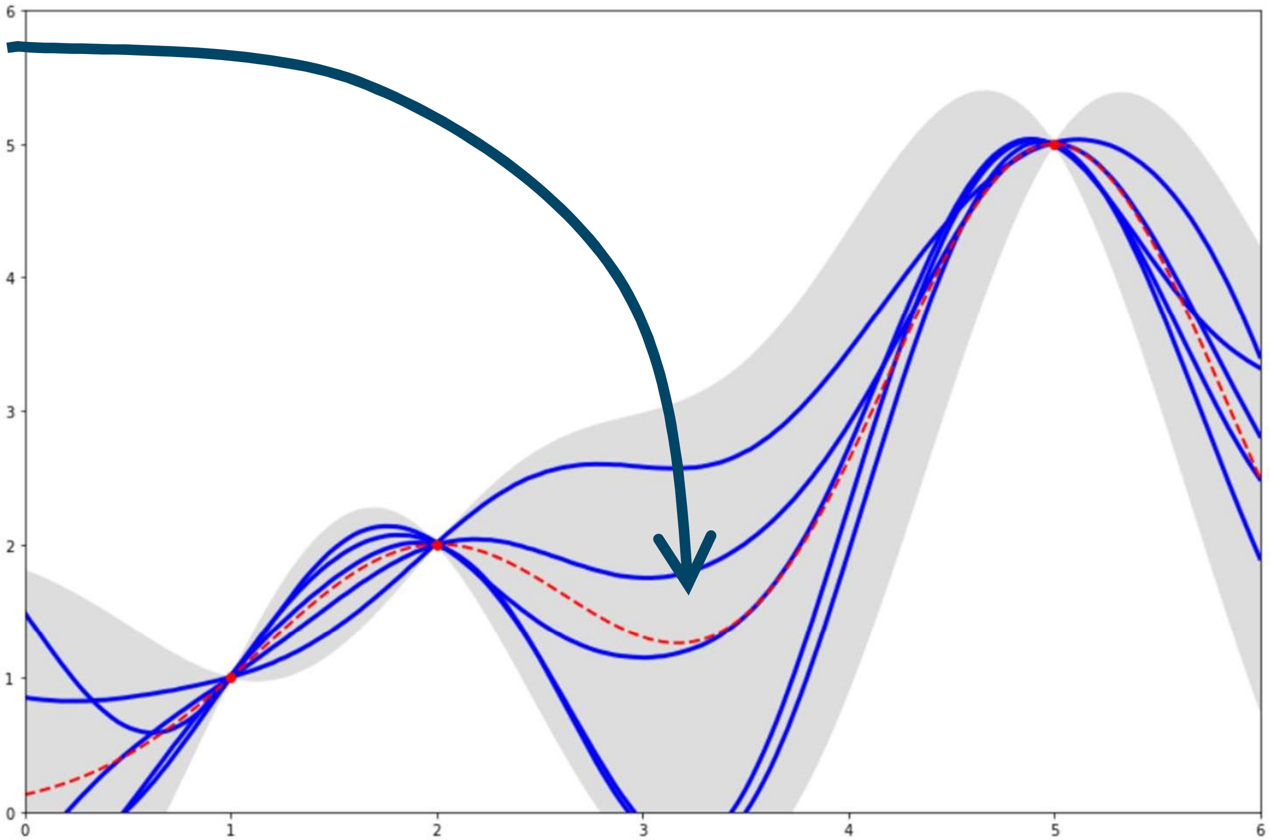
5 samples $f \sim \mathcal{GP}(\mu(x), K(x, x'))$



Gaussian Process

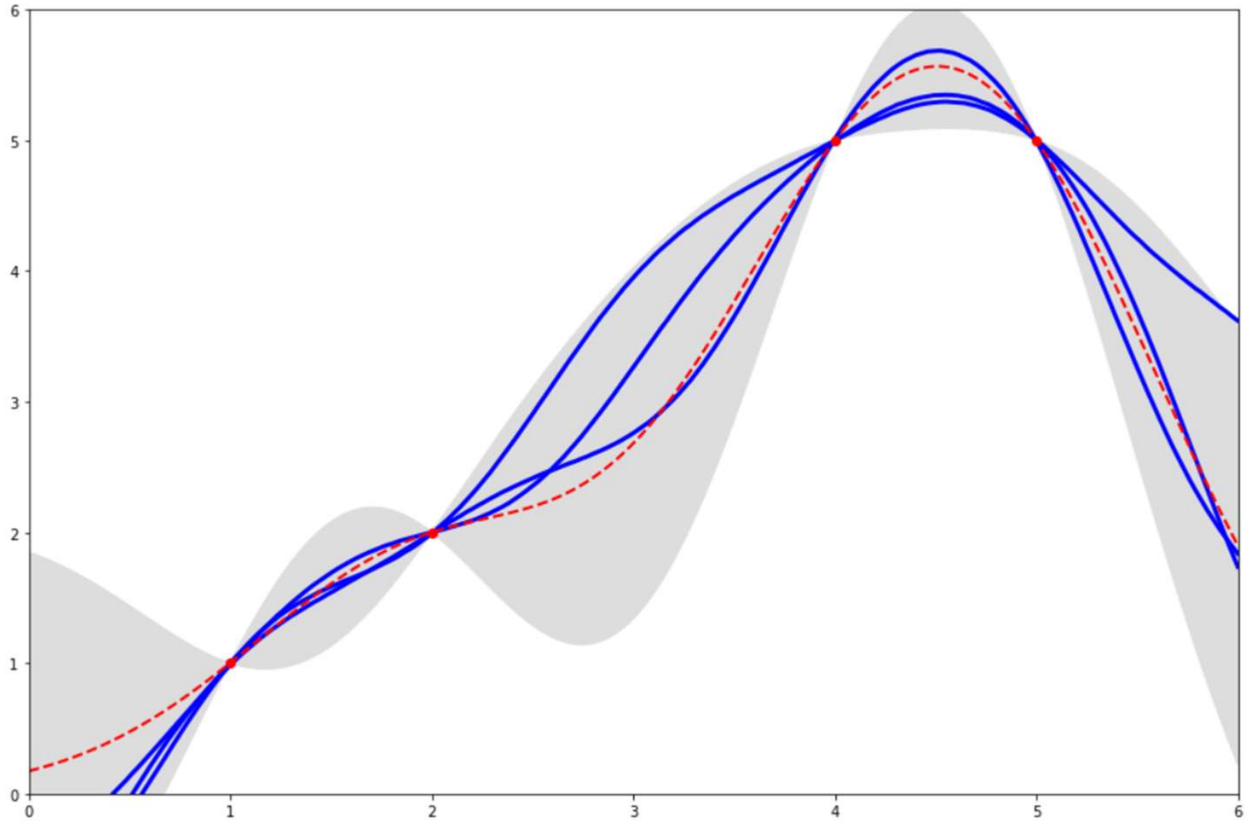
Three observations: $x_1 = 1, x_2 = 2, x_5 = 5 \Rightarrow (x_3, x_4, x_6, \dots, x_{28}, \dots | x_1, x_2, x_5) \sim \mathcal{N}(\bar{\mu}, \bar{\Sigma})$

Why this dip?



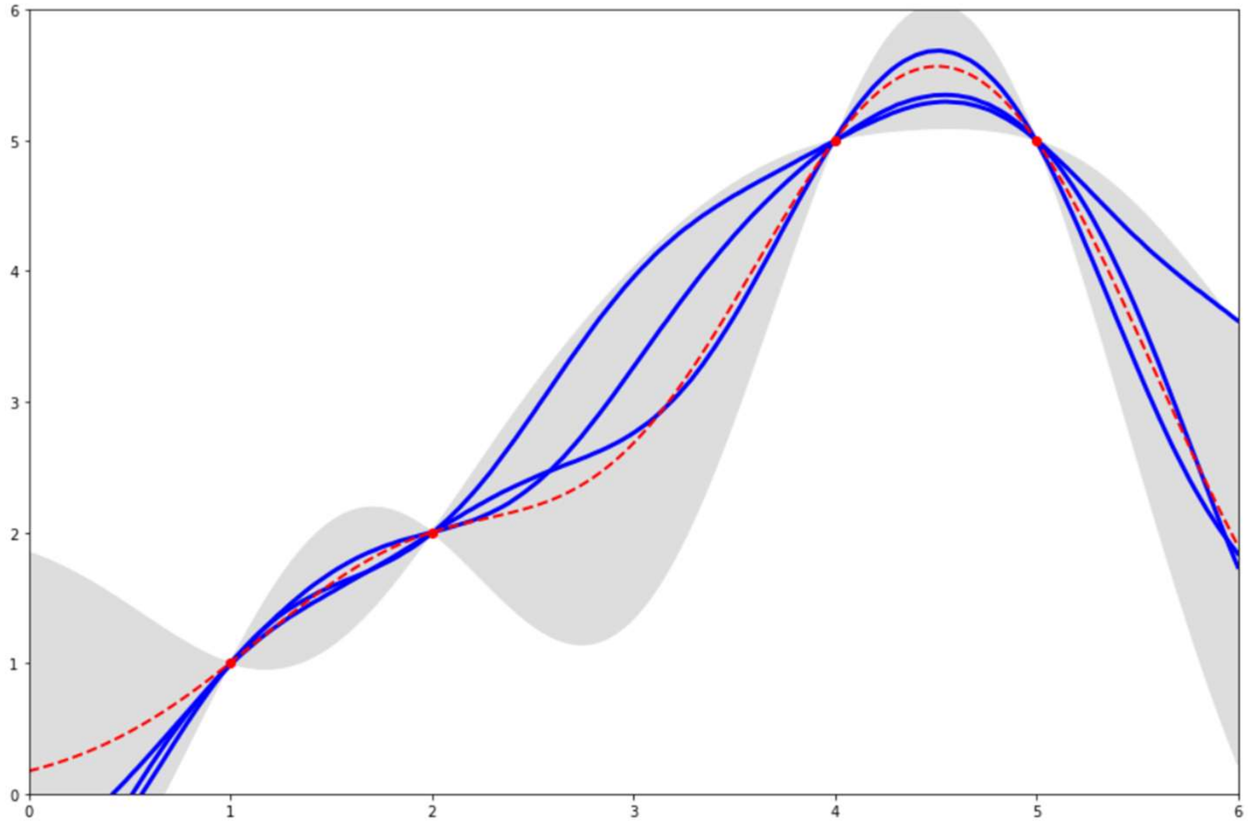
Gaussian Process

Four observations: $x_1 = 1, x_2 = 2, x_4 = 5, x_5 = 5 \Rightarrow (x_3, \dots | x_1, x_2, x_4, x_5) \sim \mathcal{N}(\bar{\mu}, \bar{\Sigma})$



Gaussian Process

Gaussian process is a stochastic process (a collection of random variables), such that every finite collection of those random variables has a multivariate normal distribution. It's a distribution over functions.



Gaussian Process pros and cons

Pros:

- Non-linear regression
- Non-parametric, let the data speak for itself
- Bonus: uncertainty interval

Cons:

- $\mathcal{O}(n^3)$ in complexity
- $\mathcal{O}(n^2)$ in storage
- Kernel tuning

Bayesian Inference

Sum rule: $p(A) = \int p(A, B)\, dB$

Product rule: $p(A, B) = p(B, A) = p(A|B)p(B) = p(B|A)p(A)$

Bayes rule: $p(A|B) = \dfrac{p(B|A)p(A)}{p(B)} = \dfrac{p(B|A)p(A)}{\int p(B,A)d} = \dfrac{p(B|A)p(A)}{\int p(B|A)p(A)dA}$

Hypothesis or model comparison:

$$p(\mathcal{H}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathcal{H})p(\mathcal{H})}{p(\mathcal{D})} = \frac{p(\mathcal{D}|\mathcal{H})p(\mathcal{H})}{\int p(\mathcal{D}|\mathcal{H})p(\mathcal{H})d\mathcal{H}}$$

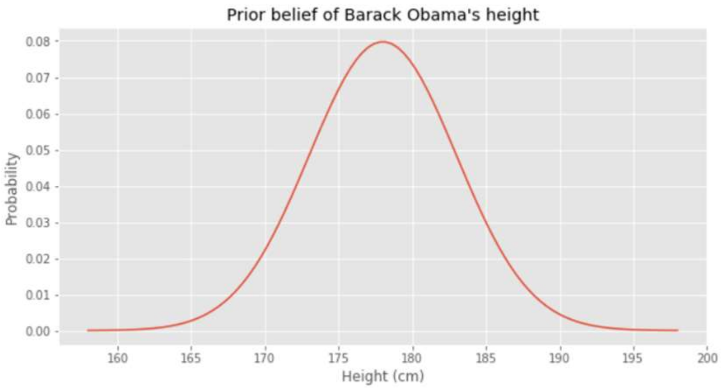
$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{marginal likelihood (evidence)}}$$

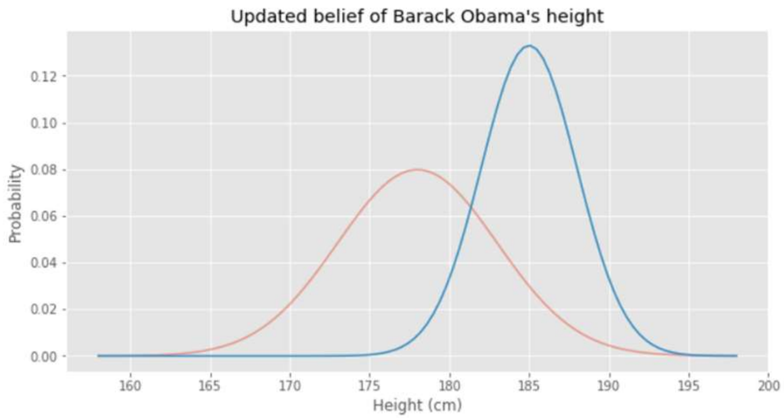
Thomas Bayes (1702 - 1761)
English mathematician
Presbyterian minister

Rediscovered by
Pierre-Simon Laplace 1774

Bayesian Inference

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{marginal likelihood (evidence)}}$$





Blue: updated belief. Red: old belief

Gaussian distribution is auto-conjugate

https://towardsdatascience.com/an-intuitive-guide-to-gaussian-processes-ec2f0b45c71d

Bayesian Inference

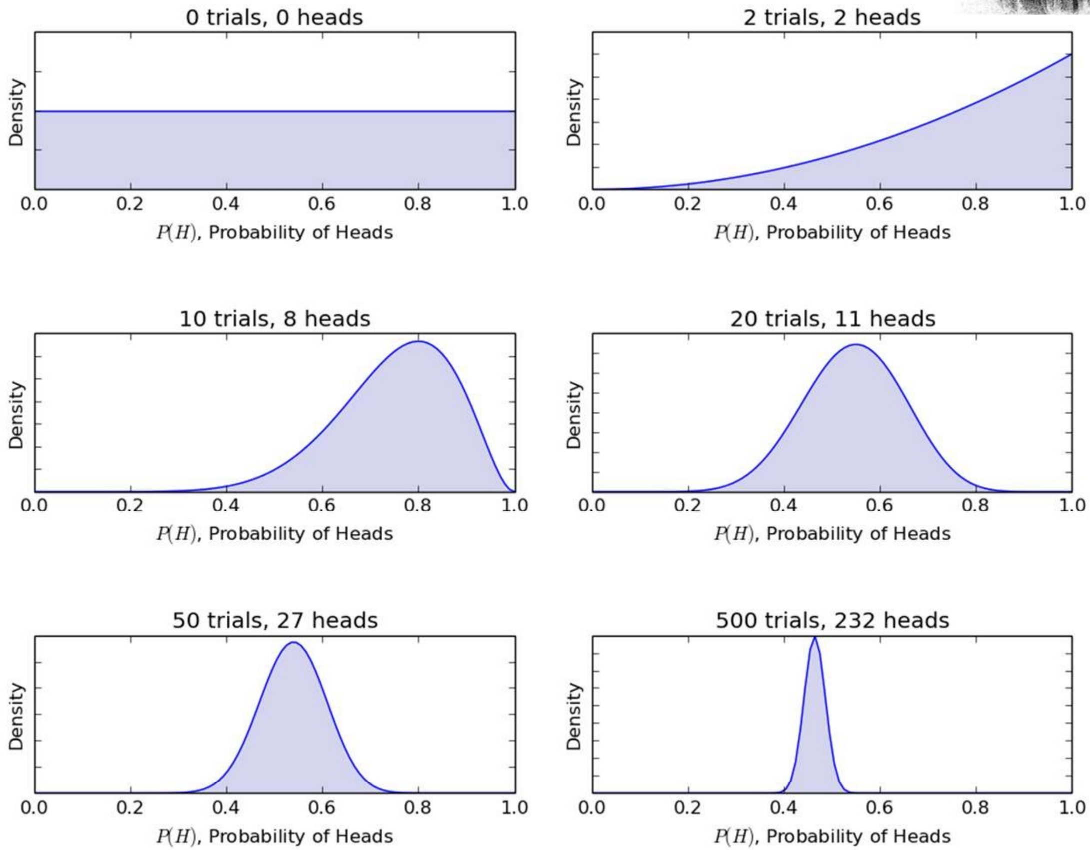
$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{marginal likelihood (evidence)}}$$

$$\text{Beta}\left(\alpha + \sum_{i=1}^{n} X_i, \beta + n - \sum_{i=1}^{n} X_i\right)$$

$$\propto \text{Bin}(\theta) \times \text{Beta}(\alpha, \beta)$$

$$\propto \theta^{k}(1-\theta)^{n-k} \times \theta^{\alpha-1}(1-\theta)^{\beta-1}$$
$$\propto \theta^{\alpha+k-1}(1-\theta)^{\beta+n-k-1}$$

Beta distribution is a conjugate prior to the Binomial (Bernouilli) distribution



https://www.analyticsvidhya.com/blog/2016/06/bayesian-statistics-beginners-simple-english/

Bayesian Inference

$$\text{posterior} = \frac{\text{likelihoo} \quad \times \text{prior}}{\text{marginal likelihood (evidence)}}$$

When likelihood function is a discrete distribution [edit]

Likelihood	Model parameters	Conjugate prior distribution	Prior hyperparameters	Posterior hyperparameters	Interpretation of hyperparameters[note 1]	Posterior predictive[note 2]
Bernoulli	p (probability)	Beta	$\alpha,\, \beta$	$\alpha + \sum_{i=1}^n x_i,\, \beta + n - \sum_{i=1}^n x_i$	$\alpha - 1$ successes, $\beta - 1$ failures[note 1]	$p(\tilde{x}=1) = \frac{\alpha'}{\alpha'+\beta'}$
Binomial	p (probability)	Beta	$\alpha,\, \beta$	$\alpha + \sum_{i=1}^n x_i,\, \beta + \sum_{i=1}^n N_i - \sum_{i=1}^n x_i$	$\alpha - 1$ successes, $\beta - 1$ failures[note 1]	$\operatorname{BetaBin}(\tilde{x}\mid\alpha',\beta')$ (beta-binomial)
Negative binomial with known failure number, r	p (probability)	Beta	$\alpha,\, \beta$	$\alpha + \sum_{i=1}^n x_i,\, \beta + rn$	$\alpha - 1$ total successes, $\beta - 1$ failures[note 1] (i.e., $\frac{\beta-1}{r}$ experiments, assuming r stays fixed)	$\operatorname{BetaNegBin}(\tilde{x}\mid\alpha',\beta')$ (beta-negative binomial)
Poisson	λ (rate)	Gamma	$k,\, \theta$	$k + \sum_{i=1}^n x_i,\, \frac{\theta}{n\theta+1}$	k total occurrences in $\frac{1}{\theta}$ intervals	$\operatorname{NB}(\tilde{x}\mid k',\theta')$ (negative binomial)
			$\alpha,\, \beta$[note 3]	$\alpha + \sum_{i=1}^n x_i,\, \beta + n$	α total occurrences in β intervals	$\operatorname{NB}\left(\tilde{x}\mid \alpha', \frac{1}{1+\beta'}\right)$ (negative binomial)
Categorical	$\mathbf{p}$ (probability vector), k (number of categories; i.e., size of $\mathbf{p}$)	Dirichlet	$\boldsymbol\alpha$	$\boldsymbol\alpha + (c_1,\ldots,c_k)$, where c_i is the number of observations in category i	$\alpha_i - 1$ occurrences of category i[note 1]	$p(\tilde{x}=i) = \frac{\alpha_i'}{\sum_i \alpha_i'} = \frac{\alpha_i + c_i}{\sum_i \alpha_i + n}$
Multinomial	$\mathbf{p}$ (probability vector), k (number of categories; i.e., size of $\mathbf{p}$)	Dirichlet	$\boldsymbol\alpha$	$\boldsymbol\alpha + \sum_{i=1}^n \mathbf{x}_i$	$\alpha_i - 1$ occurrences of category i[note 1]	$\operatorname{DirMult}(\tilde{\mathbf{x}}\mid\boldsymbol\alpha')$ (Dirichlet-multinomial)
Hypergeometric with known total population size, N	M (number of target members)	Beta-binomial[4]	$n = N, \alpha,\, \beta$	$\alpha + \sum_{i=1}^n x_i,\, \beta + \sum_{i=1}^n N_i - \sum_{i=1}^n x_i$	$\alpha - 1$ successes, $\beta - 1$ failures[note 1]	
Geometric	p_0 (probability)	Beta	$\alpha,\, \beta$	$\alpha + n,\, \beta + \sum_{i=1}^n x_i$	$\alpha - 1$ experiments, $\beta - 1$ total failures[note 1]	

When likelihood function is a continuous distribution [edit]

Likelihood	Model parameters	Conjugate prior distribution	Prior hyperparameters	Posterior hyperparameters	Interpretation of hyperparameters	Posterior predictive[note 4]
Normal with known variance σ^2	μ (mean)	Normal	$\mu_0,\, \sigma_0^2$	$\frac{1}{\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}}\left(\frac{\mu_0}{\sigma_0^2}+\frac{\sum_{i=1}^n x_i}{\sigma^2}\right), \left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right)^{-1}$	mean was estimated from observations with total precision (sum of all individual precisions) $1/\sigma_0^2$ and with sample mean μ_0	$\mathcal{N}(\tilde{x}\mid\mu_0', {\sigma_0^2}' + \sigma^2)$[5]
Normal with known precision τ	μ (mean)	Normal	$\mu_0,\, \tau_0$	$\frac{\tau_0\mu_0 + \tau\sum_{i=1}^n x_i}{\tau_0 + n\tau},\, \tau_0 + n\tau$	mean was estimated from observations with total precision (sum of all individual precisions) τ_0 and with sample mean μ_0	$\mathcal{N}\left(\tilde{x}\mid\mu_0', \frac{1}{\tau_0'} + \frac{1}{\tau}\right)$[5]
Normal with known mean μ	σ^2 (variance)	Inverse gamma	$\alpha,\, \beta$ [note 5]	$\alpha + \frac{n}{2},\, \beta + \frac{\sum_{i=1}^n (x_i-\mu)^2}{2}$	variance was estimated from 2α observations with sample variance β/α (i.e. with sum of squared deviations 2β, where deviations are from known mean μ)	$t_{2\alpha'}(\tilde{x}\mid\mu, \sigma^2 = \beta'/\alpha')$[5]
Normal with known mean μ	σ^2 (variance)	Scaled inverse chi-squared	$\nu,\, \sigma_0^2$	$\nu + n,\, \frac{\nu\sigma_0^2 + \sum_{i=1}^n (x_i-\mu)^2}{\nu+n}$	variance was estimated from ν observations with sample variance σ_0^2	$t_{\nu'}(\tilde{x}\mid\mu, {\sigma_0^2}')$[5]
Normal with known mean μ	τ (precision)	Gamma	$\alpha,\, \beta$[note 3]	$\alpha + \frac{n}{2},\, \beta + \frac{\sum_{i=1}^n (x_i-\mu)^2}{2}$	precision was estimated from 2α observations with sample variance β/α (i.e. with sum of squared deviations 2β, where deviations are from known mean μ)	$t_{2\alpha'}(\tilde{x}\mid\mu, \sigma^2 = \beta'/\alpha')$[5]

https://en.wikipedia.org/wiki/Conjugate_prior

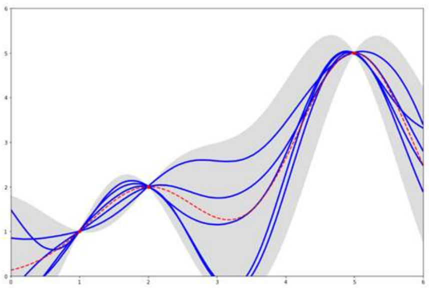
Gaussian Processes and Bayesian Inference

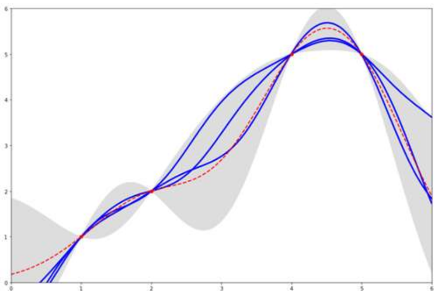
$$p(\mathbf{f}|\mathbf{y}, X, \theta) = \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|X,\theta)}{p(\mathbf{y}|X,\theta)} = \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|X,\theta)}{\int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|X,\theta)d\mathbf{f}} = \frac{\text{likelihood} \times \text{prior}}{\text{marginal likelihood (evidence)}} = \text{posterior}$$

$$p(\mathbf{f}, \mathbf{f}_*|X_*, X, \theta) = \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} K & K_* \\ {K_*}^{\mathrm{T}} & K_{**} \end{bmatrix}\right)$$

$$p(\mathbf{f}_*|X_*, X, \mathbf{y}, \theta) = \int p(\mathbf{f}_*, \mathbf{f}|X_*, X, \mathbf{y})\, d\mathbf{f} = \int p(\mathbf{f}_*|\mathbf{f}, X_*, X, \theta)p(\mathbf{f}|\mathbf{y}, X, \theta)\, d\mathbf{f}$$

$$p(\mathbf{f}_*|X_*, X, \mathbf{y}, \theta) = \mathcal{N}\left({K_*}^{\mathrm{T}} K^{-1}\mathbf{y}\, ,\, K_{**} - {K_*}^{\mathrm{T}} K^{-1}\, K_*\right)$$





Kernels

Squared Exponential

a.k.a. SE

a.k.a. RBF

a.k.a. EQ

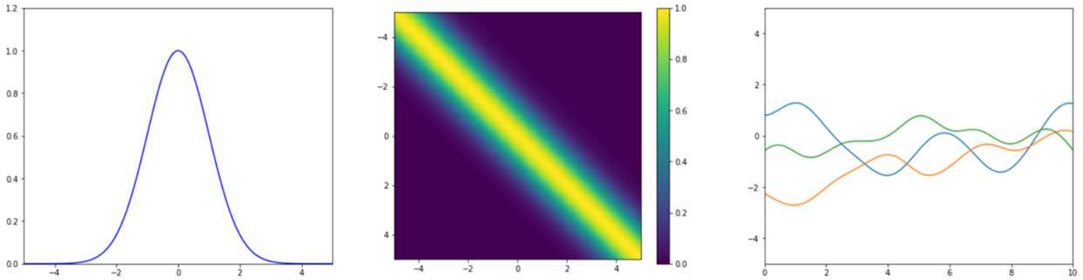
a.k.a. Gaussian

$$k(x, x') = \sigma^2 e^{-\frac{1}{2}\left(\frac{|x-x'|}{l}\right)^2}$$

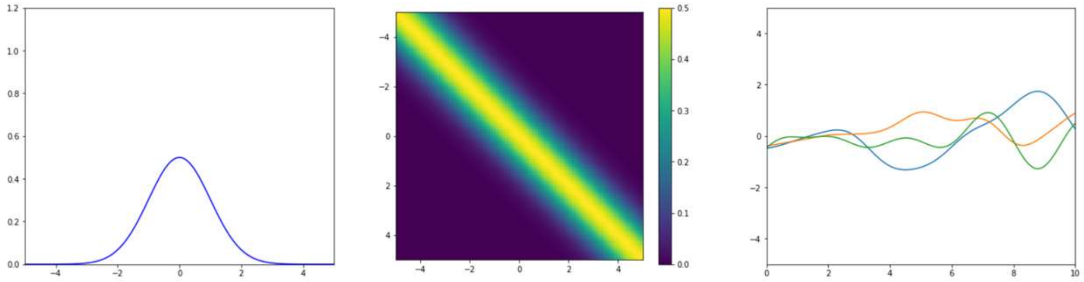
$$p(X) = p(\mathrm{x}_1, \mathrm{x}_2, \mathrm{x}_3, \mathrm{x}_4, \mathrm{x}_5) = \frac{1}{\sqrt{(2\pi)^2|\Sigma|}} e^{-\frac{1}{2}(X-\mu)^{\mathrm{T}}\Sigma^{-1}(X-\mu)}$$

$$\mu = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} 1 & 0.8 & 0.5 & 0.1 & 0 \\ 0.8 & 1 & 0.8 & 0.5 & 0.1 \\ 0.5 & 0.8 & 1 & 0.8 & 0.5 \\ 0.1 & 0.5 & 0.8 & 1 & 0.8 \\ 0 & 0.1 & 0.5 & 0.8 & 1 \end{bmatrix}$$

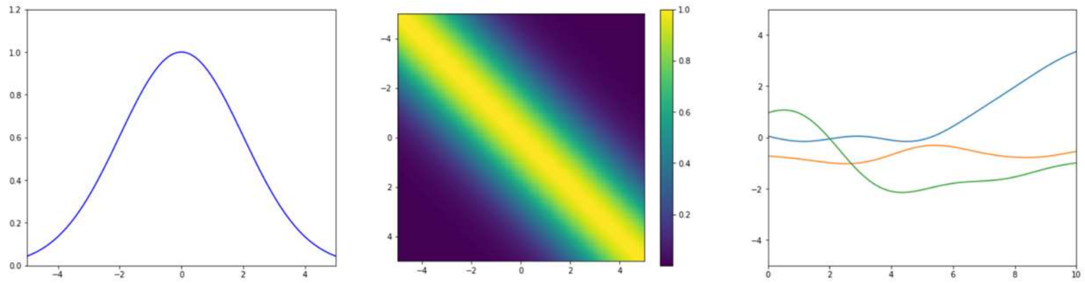
$\sigma^2 = 1$
$l = 1$



$\sigma^2 = 0.5$
$l = 1$



$\sigma^2 = 1$
$l = 2$



Kernels

Rational Quadratic

a.k.a. RQ

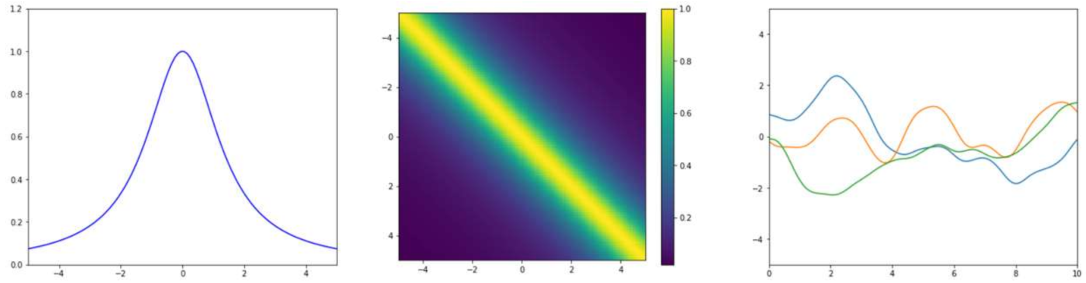
adding SE kernels

with different lengthscales

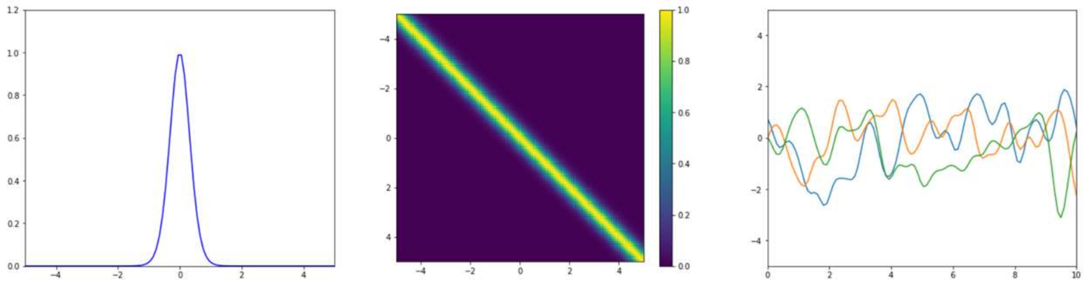
$$k(x, x') = \sigma^2 \left(1 + \frac{|x - x'|}{2\alpha l^2}\right)^{-\alpha}$$

$$k(x, x')_{\alpha \to \infty} \approx k(x, x')_{SE}$$

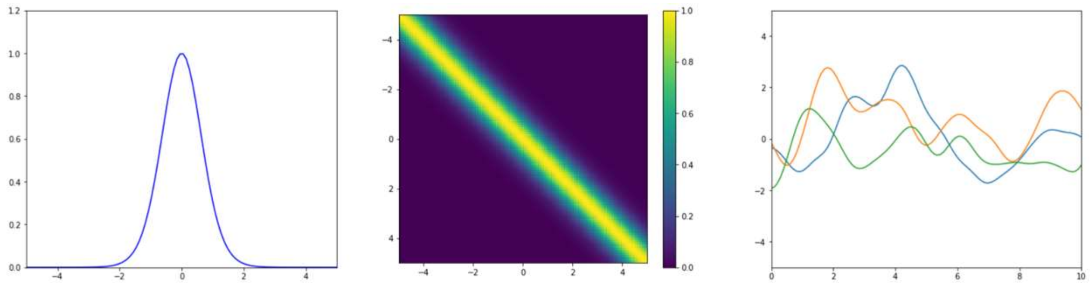
$\sigma^2 = 1$
$l = 1$
$\alpha = 1$



$\sigma^2 = 1$
$l = 1$
$\alpha = 10$



$\sigma^2 = 1$
$l = 2$
$\alpha = 10$



Kernels

Matérn family

$$k(x, x') = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}|x - x'|}{l} \right)^{\nu} K_{\nu} \left(\frac{\sqrt{2\nu}|x - x'|}{l} \right)$$

$$\nu = p + \frac{1}{2}$$

$$k(x, x')_{1/2} = \sigma^2 e^{\frac{-|x-x'|}{l}}$$

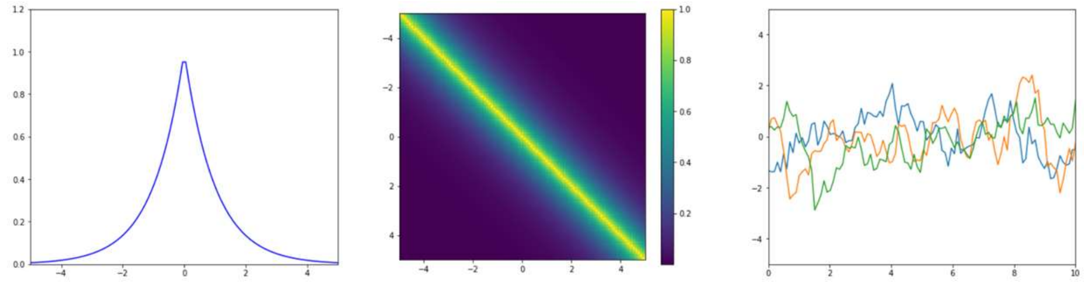
a.k.a. exponential, a.k.a. Ornstein–Uhlenbeck

$$k(x, x')_{3/2} = \sigma^2 \left(1 + \frac{\sqrt{3}|x - x'|}{l} \right) e^{\frac{-\sqrt{3}|x-x'|}{l}}$$

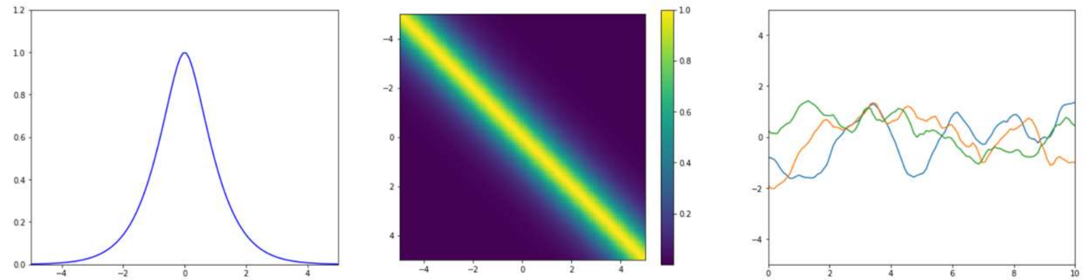
$$k(x, x')_{5/2} = \sigma^2 \left(1 + \frac{\sqrt{5}|x - x'|}{l} + \frac{5|x - x'|^2}{3l^2} \right) e^{\frac{-\sqrt{5}|x-x'|}{l}}$$

$$k(x, x')_{>7/2} \approx k(x, x')_{SE}$$

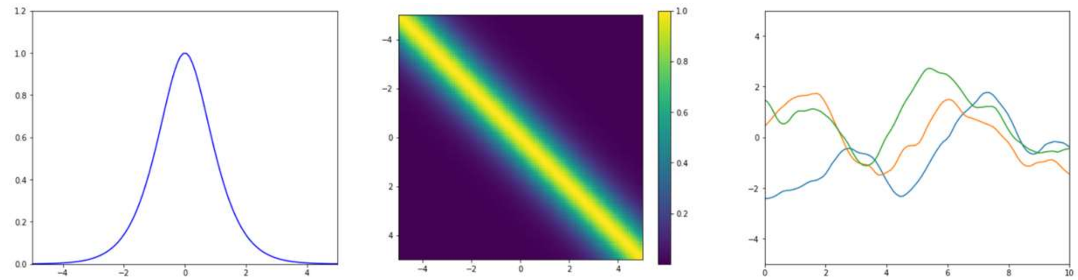
$\sigma^2 = 1$
$l = 1$
$\nu = 1/2$



$\sigma^2 = 1$
$l = 1$
$\nu = 3/2$



$\sigma^2 = 1$
$l = 1$
$\nu = 5/2$



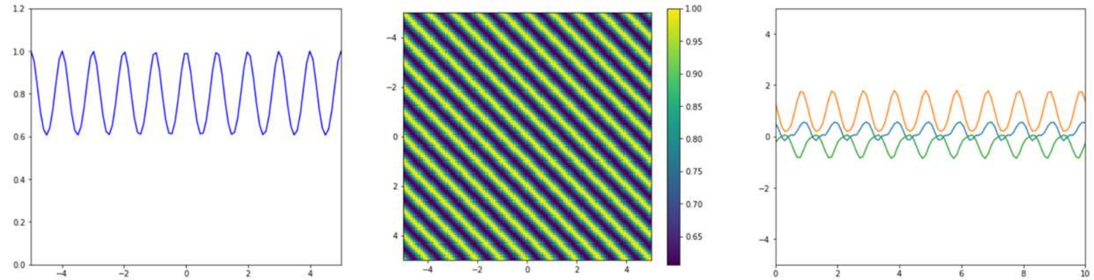
Kernels

Periodic

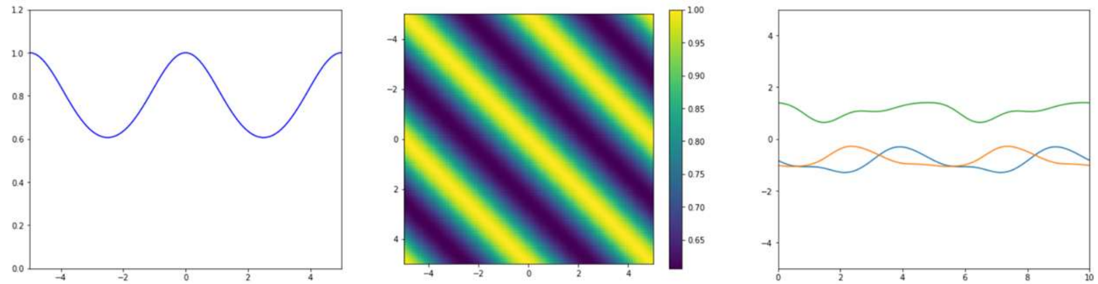
a.k.a. PER

$$k(x, x') = \sigma^2 e^{-\frac{2}{l^2}\sin^2\left(\frac{2\pi|x-x'|}{p}\right)}$$

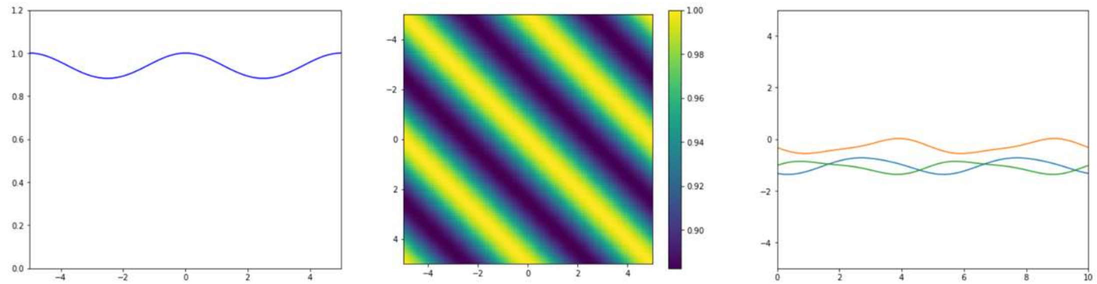
$$\sigma^2 = 1$$
$$l = 1$$
$$p = 1$$



$$\sigma^2 = 1$$
$$l = 1$$
$$p = 5$$

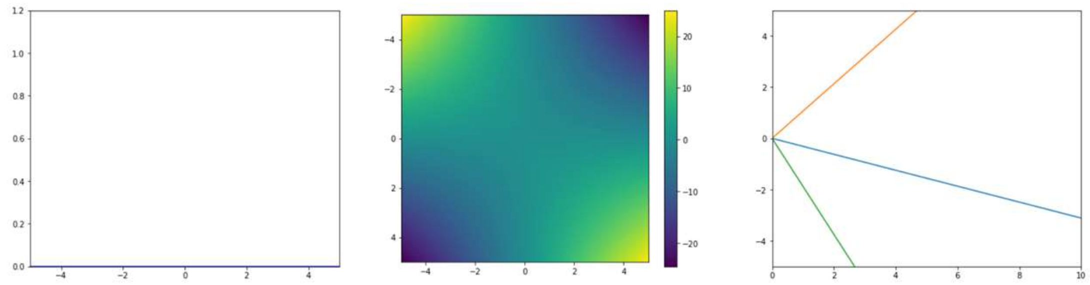


$$\sigma^2 = 1$$
$$l = 2$$
$$p = 5$$

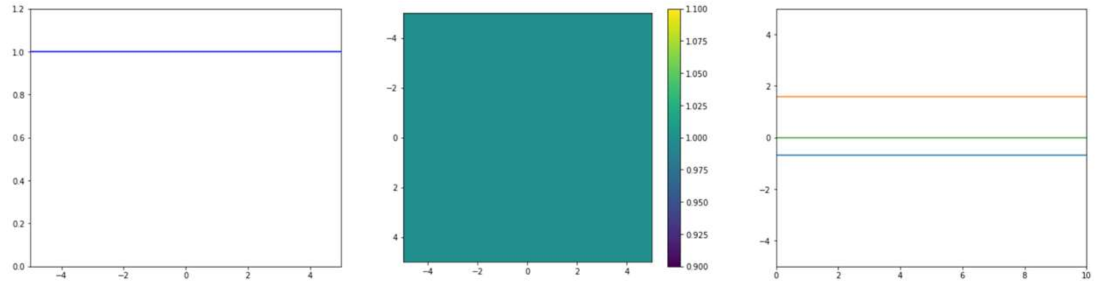


Kernels

Linear: $k(x, x') = \sum_{d=1}^{D} \sigma^2 x_d x'_d$



Constant a.k.a. Bias: $k(x, x') = \sigma^2$

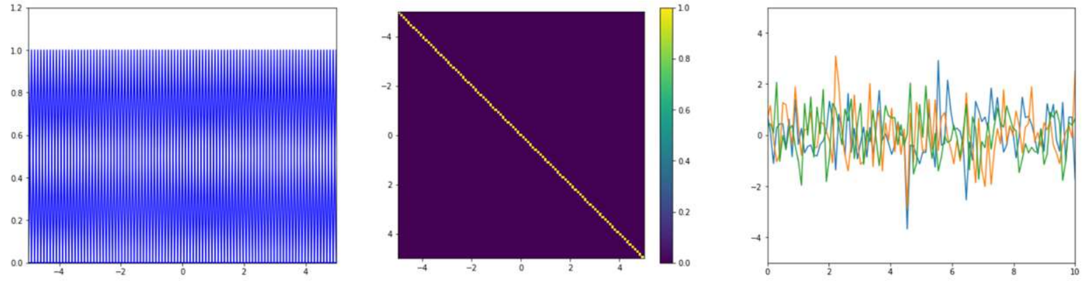


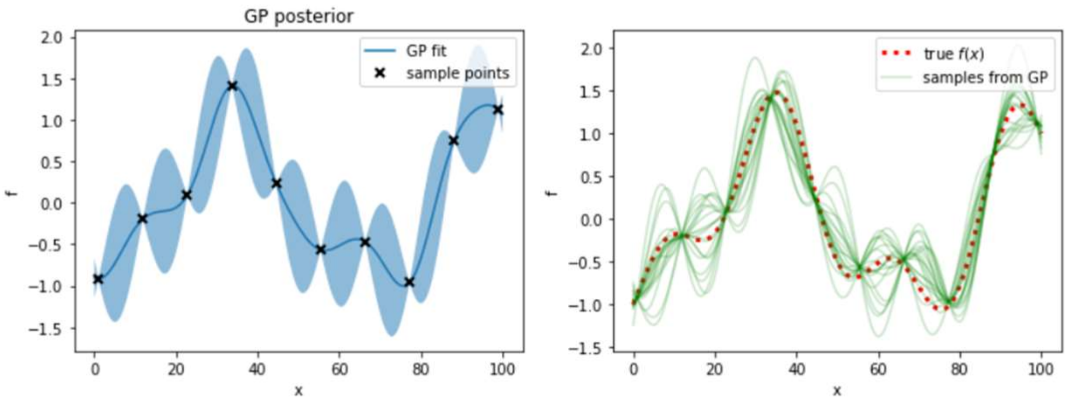
Many more: Laplace, cosine, neural net, polynomial, γ-exponential, …

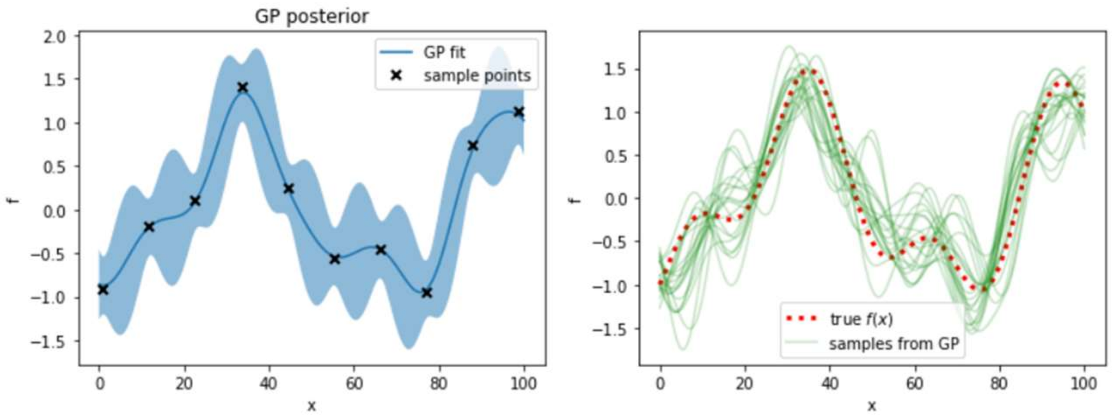
Kernels

White Noise: $k(x, x') = \sigma^2 \delta_{x,x'}$

$k(x, x')_{noise} = k(x, x') + \sigma_n^2 \delta_{x,x'}$
$K_{noise} = K + \sigma_n^2 I$

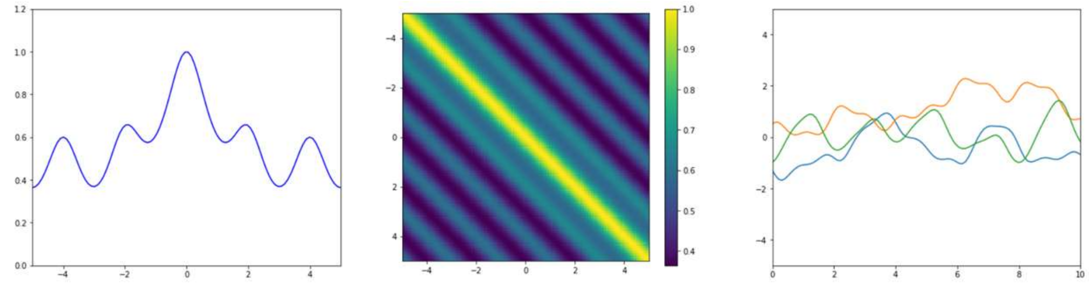




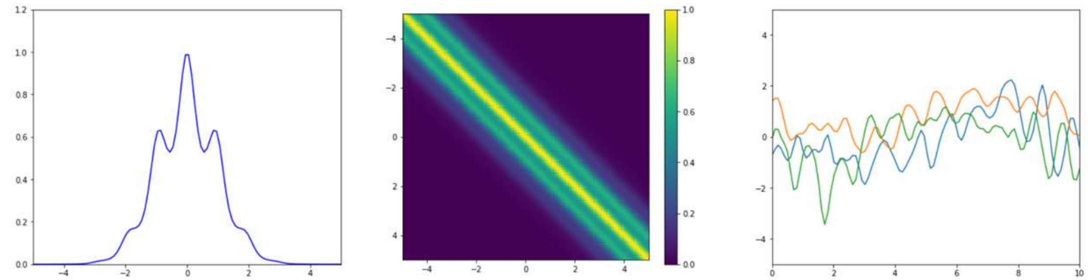


Kernels

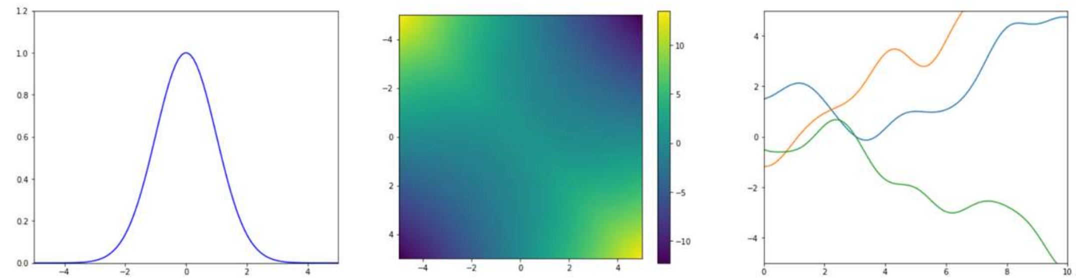
SE + PER



SE x PER

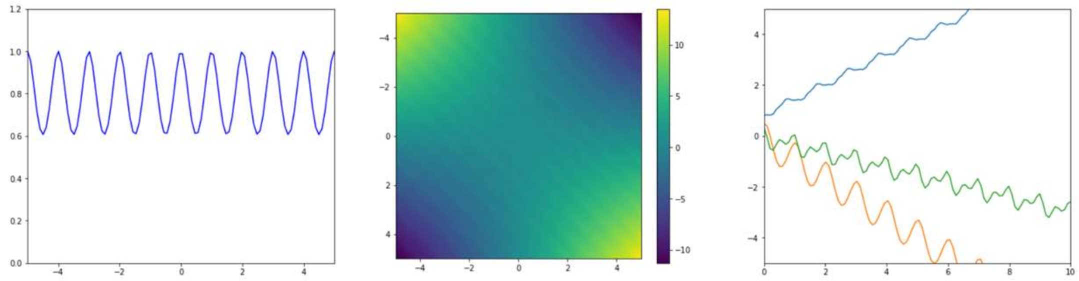


SE + LIN

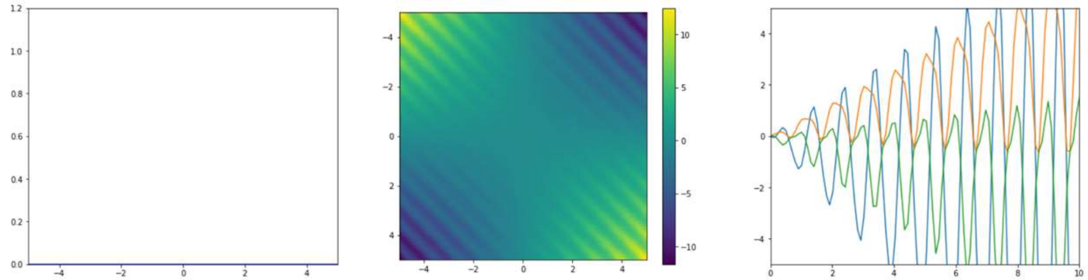


Kernels

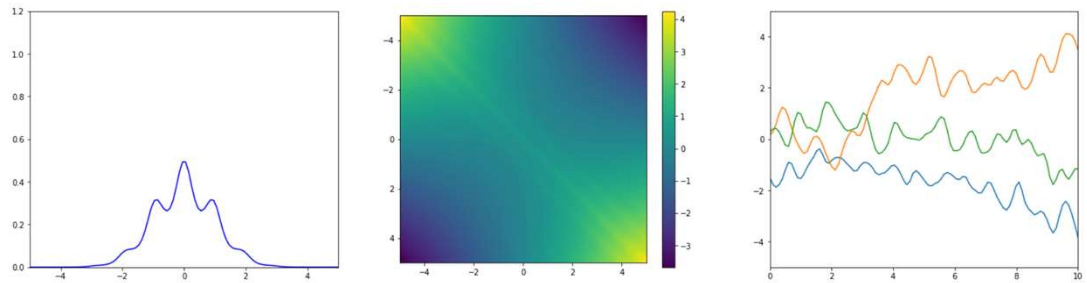
PER + LIN



PER x LIN



LIN + (PE x SE)



Kernels

a.k.a. covariance function a.k.a. kernel function a.k.a. covariance kernel, a.k.a. Gram matrix

x and x': vectors in a Euclidean space, graphs, images, discrete or categorical inputs, text, …

Stationary when depending on |x-x'|

Isotropic when same in all dimensions

Kernels

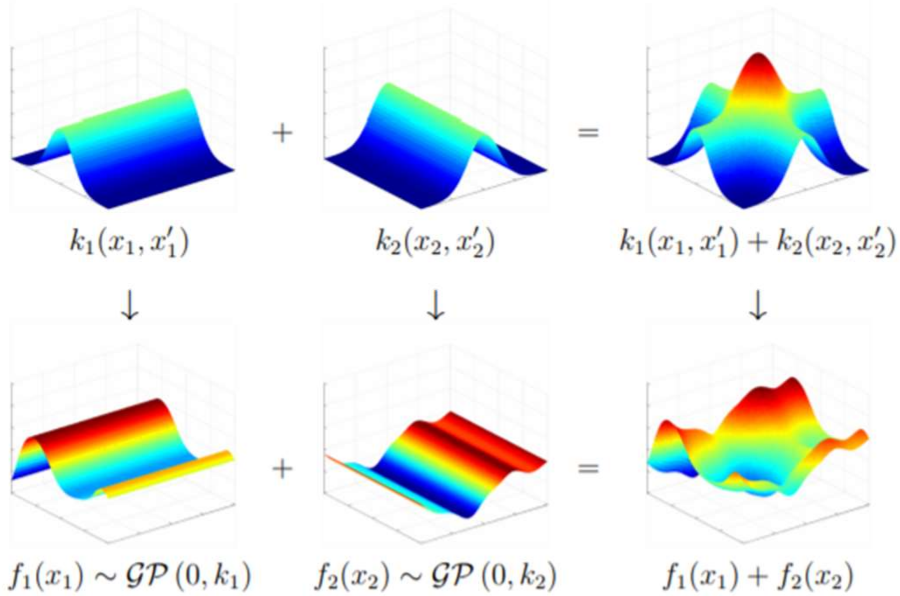
Additive across multiple dimensions, ARD

$f(x) \sim \mathcal{GP}(0, k_1(x,x') + k_2(x,x'))$

$f(x_1, x_2) \sim \mathcal{GP}(0, k_1(x_1,x_1') + k_2(x_2,x_2'))$

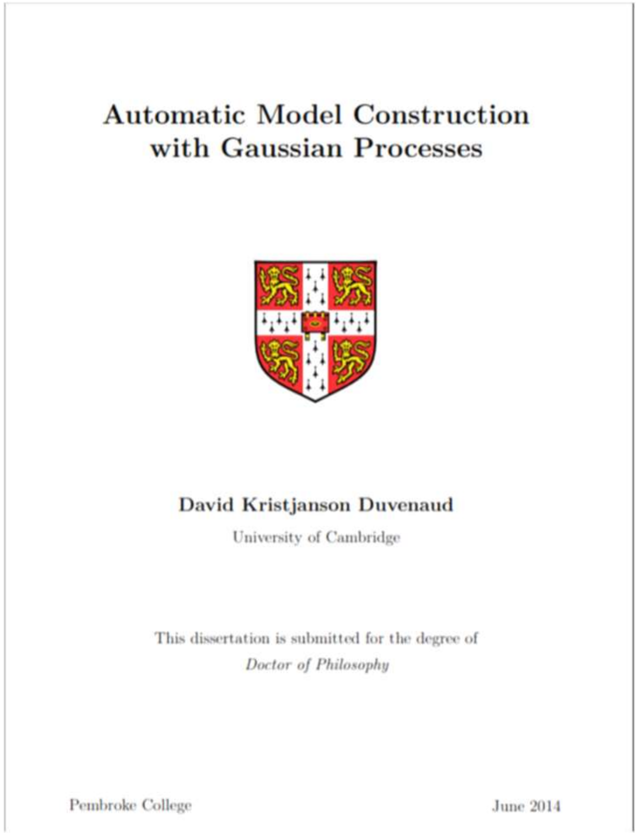
$f(x) \sim \mathcal{GP}(0, k_1(x,x') \times k_2(x,x'))$

$f(x_1, x_2) \sim \mathcal{GP}(0, k_1(x_1,x_1') \times k_2(x_2,x_2'))$



David Kristjanson Duvenaud, "Automatic Model Construction with Gaussian Processes", https://www.cs.toronto.edu/~duvenaud/thesis.pdf

Kernels


Automatic Model Construction
with Gaussian Processes

David Kristjanson Duvenaud
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Pembroke College
June 2014

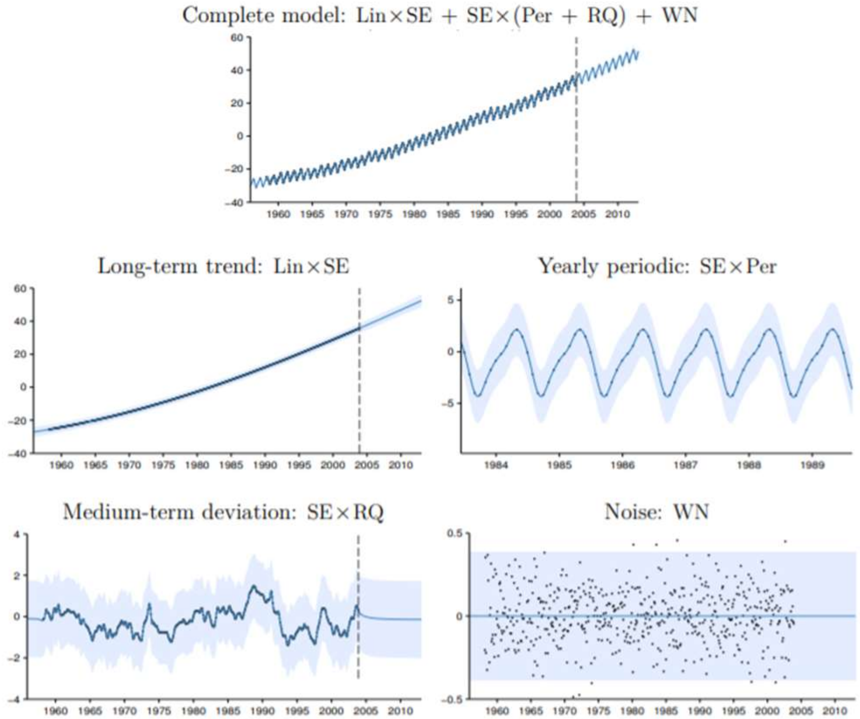


Figure 3.4: *First row:* The full posterior on the Mauna Loa dataset, after a search of depth 10. *Subsequent rows:* The automatic decomposition of the time series. The model is a sum of long-term, yearly periodic, medium-term components, and residual noise, respectively. The yearly periodic component has been rescaled for clarity.

Kernels

Positive definite matrix (function), so it is **invertible**

$$\begin{aligned}&\text{a symmetric, n} \times \text{n, real matrix } K \text{ is positive definite}\\ &\Leftrightarrow z^T K z > 0 \text{ for all } z \in \mathbb{R}^n \backslash 0\\ &\Leftrightarrow \text{all eigenvalues} > 0\end{aligned}$$

$$K = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}, z^T K z > 0$$

$$K = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}, z^T K z = -2 < 0, \text{for } z = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

In practice use Cholesky decomposition, add noise across diagonal and solve system of linear equations

$$K = LL^T, K^{-1} = (L^T)^{-1} L^{-1}$$

Kernels

Mercer's Theorem (1909):

Any positive-definite kernel can be represented as the inner product between a fixed set of features, evaluated at x and at x'.

$$k(\mathrm{x}, \mathrm{x}') = h(\mathrm{x})^{\mathrm{T}} h(\mathrm{x}')$$

Kernels

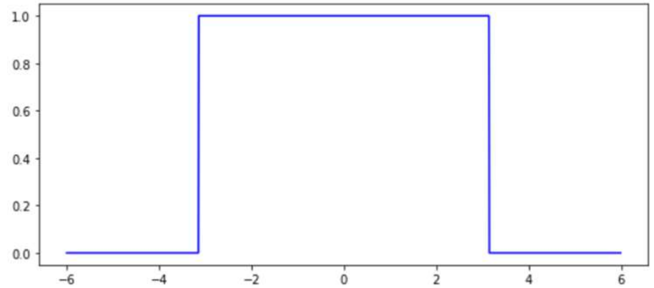
Bochner's Theorem (1959):

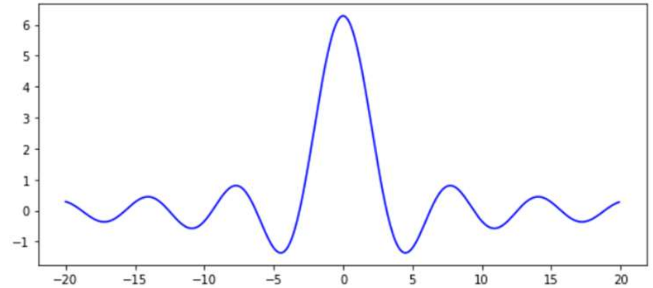
A continuous stationary function k(x, x') = k(|x − x'|) is positive definite if and only if k is the Fourier transform of a finite positive measure :

$$k[t] = \int_{\mathbb{R}} e^{-i\omega} \; d\mu(\omega)$$

Useful to prove the positive definiteness of stationary functions.

$$k(x, x') = \frac{\sin|x - x'|}{|x - x'|}$$





Kernels

Hyperparameter tuning via log marginal likelihood

$$p\,(\mathbf{f}|\mathbf{y}, X, \theta) = \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|X,\theta)}{p(\mathbf{y}|X,\theta)} = \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|X,\theta)}{\int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|X,\theta)d\mathbf{f}}$$

$$= \frac{\text{likelihood} \times \text{prior}}{\text{marginal likelihood (evidence)}} = \text{posterior}$$

$$\log p(\mathbf{y}|X, \theta) = -\frac{1}{2}\mathbf{y}^{\mathrm{T}}K_y^{-1}\,\mathbf{y} - \frac{1}{2}\log\left|K_y\right| - \frac{n}{2}\log 2\pi$$

Combination of a data fit term and complexity penalty (Occam's Razor)

Cross validation (k-fold or LOO-CV)

Kernels

Hyperparameter

Broyden–
Fletcher–
Goldfarb–
Shanno
(l-BFGS)

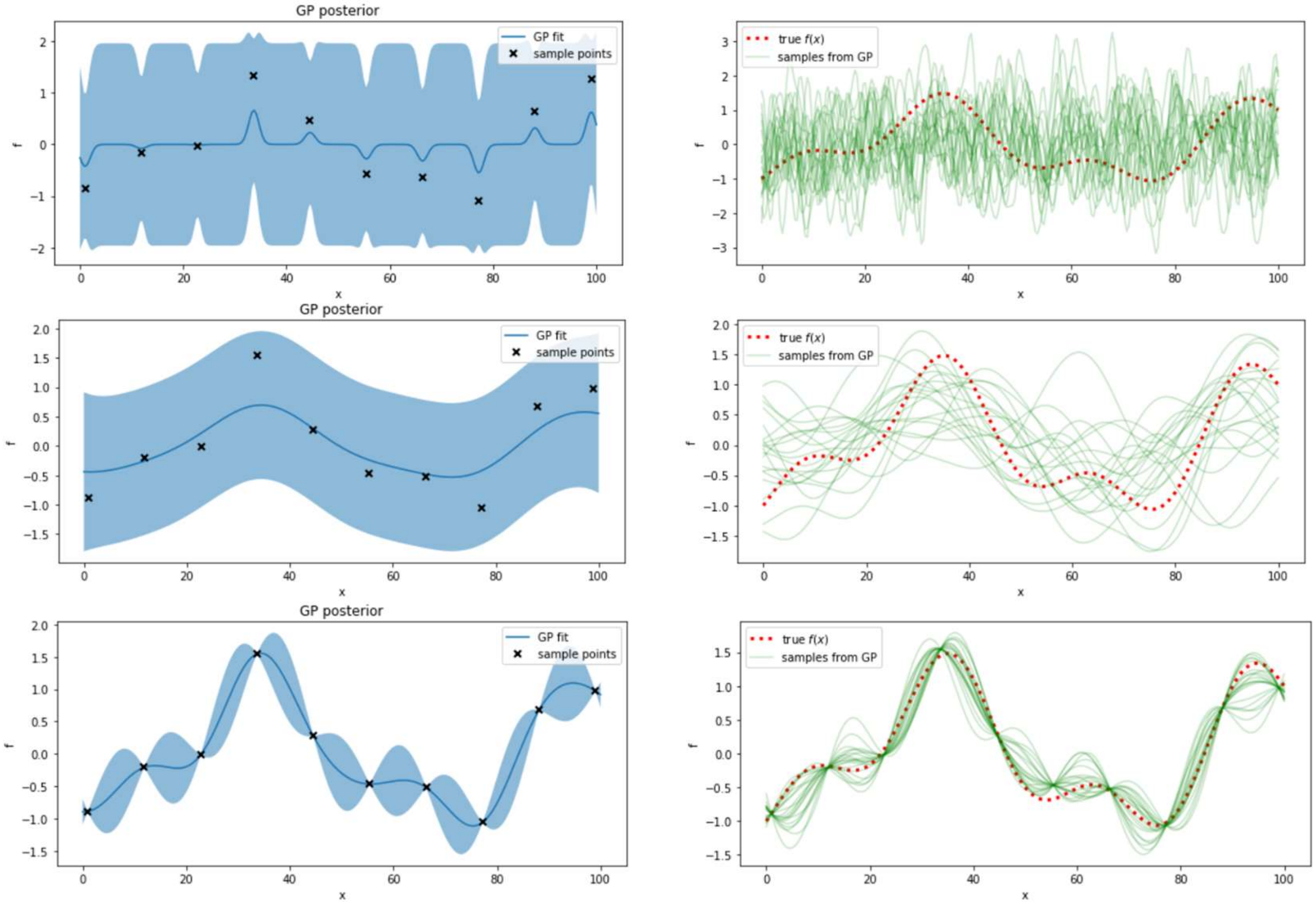
DIRECT

CMA-ES

$$\sigma^2 = 1$$
$$l = 1$$

$$\sigma^2 = 1$$
$$l = 10$$

$$\sigma^2 = 0.6$$
$$l = 7.1$$

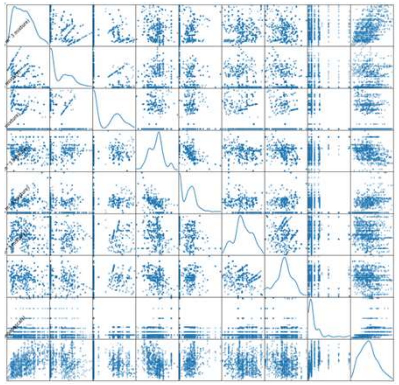


Kernels

Kaggle example: Predicting Compressive Strength of Concrete

https://www.kaggle.com/pavanraj159/concrete-compressive-strength-data-set

	Cement	Blast Furnace Slag	Fly Ash	Water	Superplasticizer	Coarse Aggregate	Fine Aggregate	Age	Concrete compressive strength
0	540.0	0.0	0.0	162.0	2.5	1040.0	676.0	28	79.99
1	540.0	0.0	0.0	162.0	2.5	1055.0	676.0	28	61.89
2	332.5	142.5	0.0	228.0	0.0	932.0	594.0	270	40.27
3	332.5	142.5	0.0	228.0	0.0	932.0	594.0	365	41.05
4	198.6	132.4	0.0	192.0	0.0	978.4	825.5	360	44.30



Standard SE x Matérn3/2: RMSE 14.4

Optimized SE x Matérn3/2: RMSE 6.8

Standard SE + Matérn3/2 , with ARD: RMSE 8.1

Optimized SE + Matérn3/2, with ARD: RMSE 6.1

Standard RQ, but each one only active in one dimension: RMSE 6.8

Optimized RQ, but each one only active in one dimension: RMSE 4.95

Variance, length scale and power plus overall noise optimized, 8 x 3 + 1 = 25 parameters

	Model	RMSE	R Squared
0	Linear Regression	10.278166	0.625282
1	Ridge Regression	10.276215	0.625366
2	Lasso Regression	10.871002	0.580476
3	K Neighbors Regressor	9.144185	0.701923
4	Decision Tree Regressor	7.974669	0.778044
5	Random Forest Regressor	5.595289	0.889286
6	Gradient Boosting Regressor	5.184161	0.904415
7	Adaboost Regressor	7.719348	0.783844

Classification and Non Gaussian Likelihoods

$$p(\mathbf{f}|\mathbf{y}, X, \theta) = \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|X,\theta)}{p(\mathbf{y}|X,\theta)} = \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|X,\theta)}{\int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|X,\theta)d\mathbf{f}} = \frac{\text{likelihoo} \quad \times \text{prior}}{\text{marginal likelihood (evidence)}} = \text{posterior}$$

$$p(\mathbf{f}_*|X_*, X, \mathbf{y}, \theta) \sim \mathcal{N}\left(\mathbf{f}_*|{K_*}^{\mathrm{T}}K^{-1}\mathbf{y}\,,\, K_{**} - {K_*}^{\mathrm{T}}K^{-1}\, K_*\right)$$

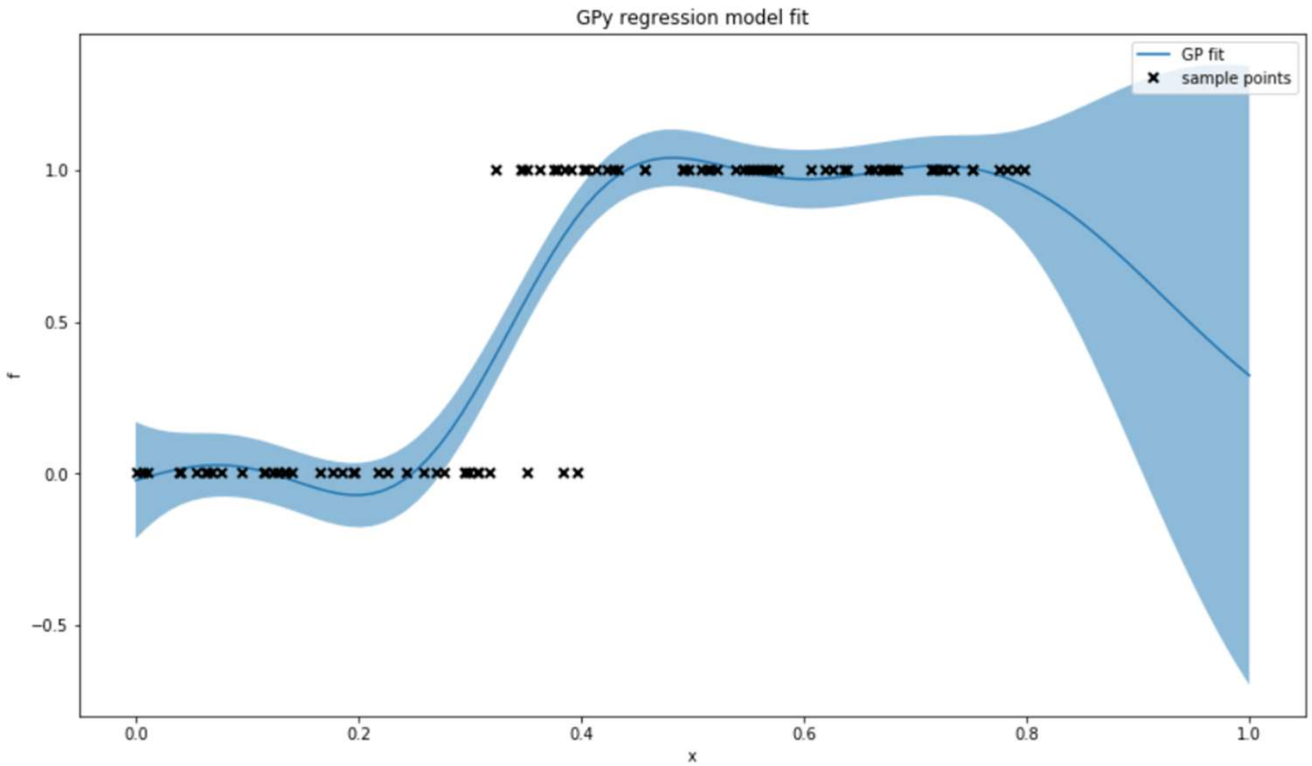
This assumes Gaussian likelihood (Gaussian distribution auto-conjugate)
in order to be analytically tractable (closed form solution)

$$p(\mathbf{f}|\mathbf{y}, X, \theta) = \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|X,\theta)}{p(\mathbf{y}|X,\theta)} = \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|X,\theta)}{\int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|X,\theta)d\mathbf{f}}$$

Classification and Non Gaussian Likelihoods

Binairy classification ~ Bernouilli likelihood

Classification as regression is called least-squares classification



Classification – Non Gaussian Likelihoods

Binairy classification ~ Bernouilli likelihood

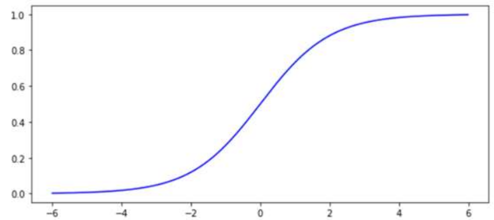
Fit GP over latent f, then squash with logit or probit

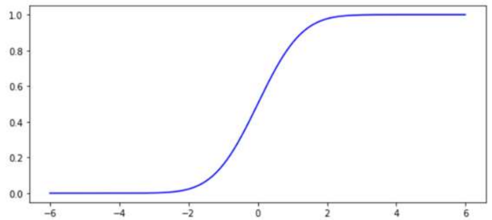
$$p(f_*|X_*, X, \mathbf{y}) = \int p(f_*|X_*, X, \mathbf{f})\, p\,(\mathbf{f}|X, \mathbf{y}) d\mathbf{f}$$

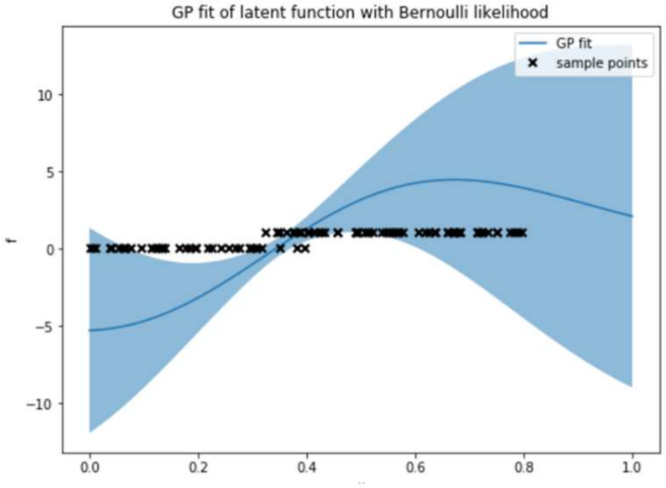
$$\pi_* \triangleq p\,(y_* = +1|X_*, X, \mathbf{y}) = \int \sigma(f_*)\, p\,(f_*|X_*, X, \mathbf{y}) df_*$$

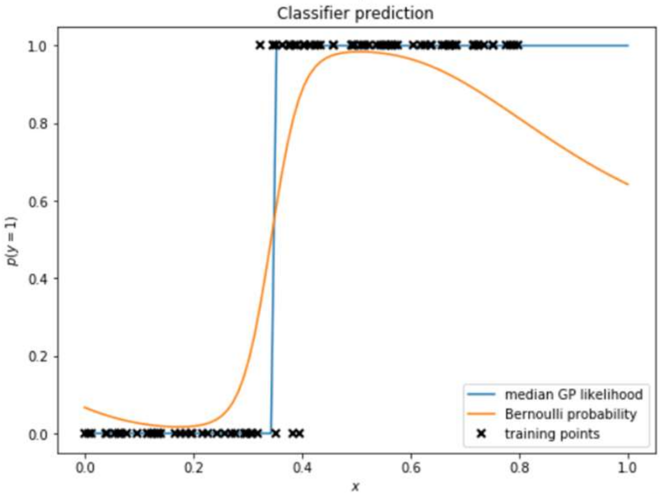
logit $\sigma(z) = \frac{1}{1+e^{-z}}$

probit $\sigma(z) = \Phi(z) = \int_{-\infty}^{z} \mathcal{N}(x|0,1)dx$







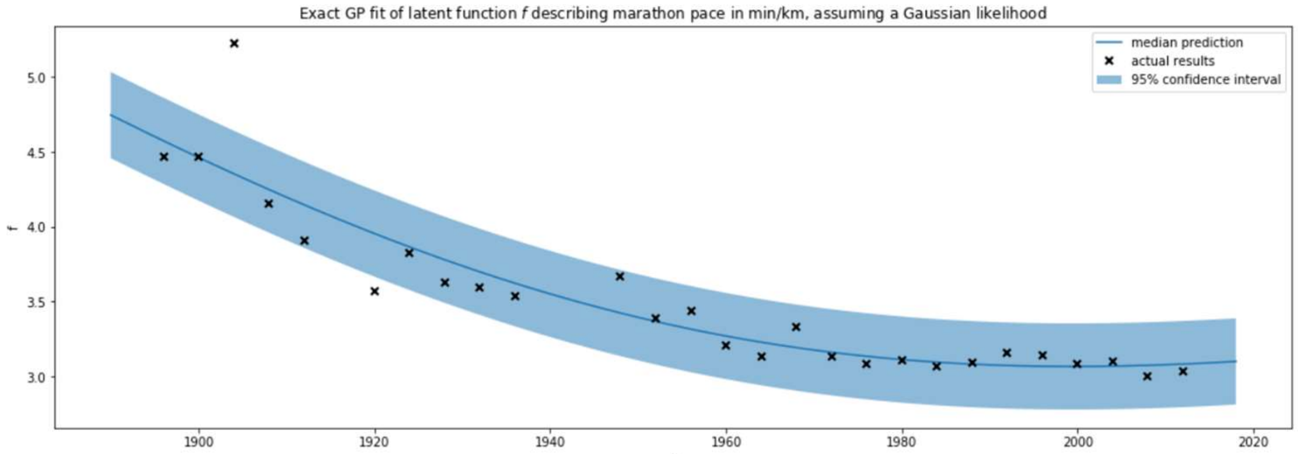


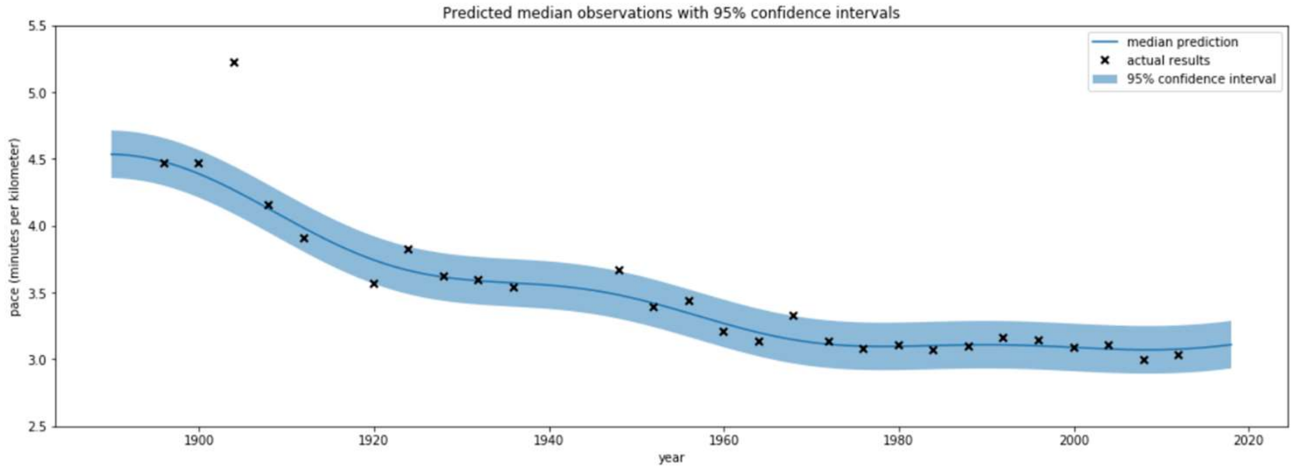
Classification and Non Gaussian Likelihoods

Olympic mens' marathon gold medal winning times from 1896 to 2012 (recurrent GPSS example)

Outlier at 1904

Student’s t-distribution likelihood

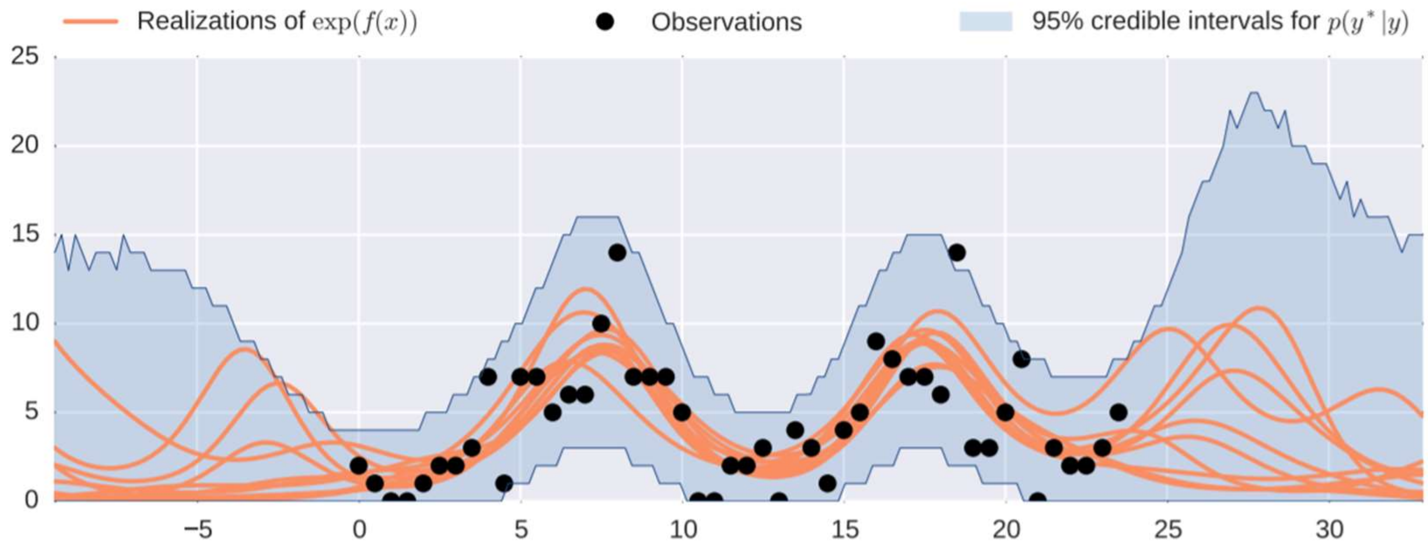




Classification and Non Gaussian Likelihoods

Count data

$$\mathbf{y}_i \sim \text{Poisson}(\mathbf{y}_i|\lambda_i = \lambda(\mathbf{f}_i)) \qquad \text{Poisson}(\mathbf{y}_i|\lambda_i) = \frac{\lambda_i^{\mathbf{y}_i}}{!\mathbf{y}_i} e^{-\lambda_i}$$



Alan Saul, "*Non-Gaussian likelihoods for Gaussian Processes*", GPSS15

Classification and Non Gaussian Likelihoods

$$p(\mathbf{f}|\mathbf{y}, X, \theta) = \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|X,\theta)}{p(\mathbf{y}|X,\theta)} = \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|X,\theta)}{\int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|X,\theta)d\mathbf{f}}$$

Approximations

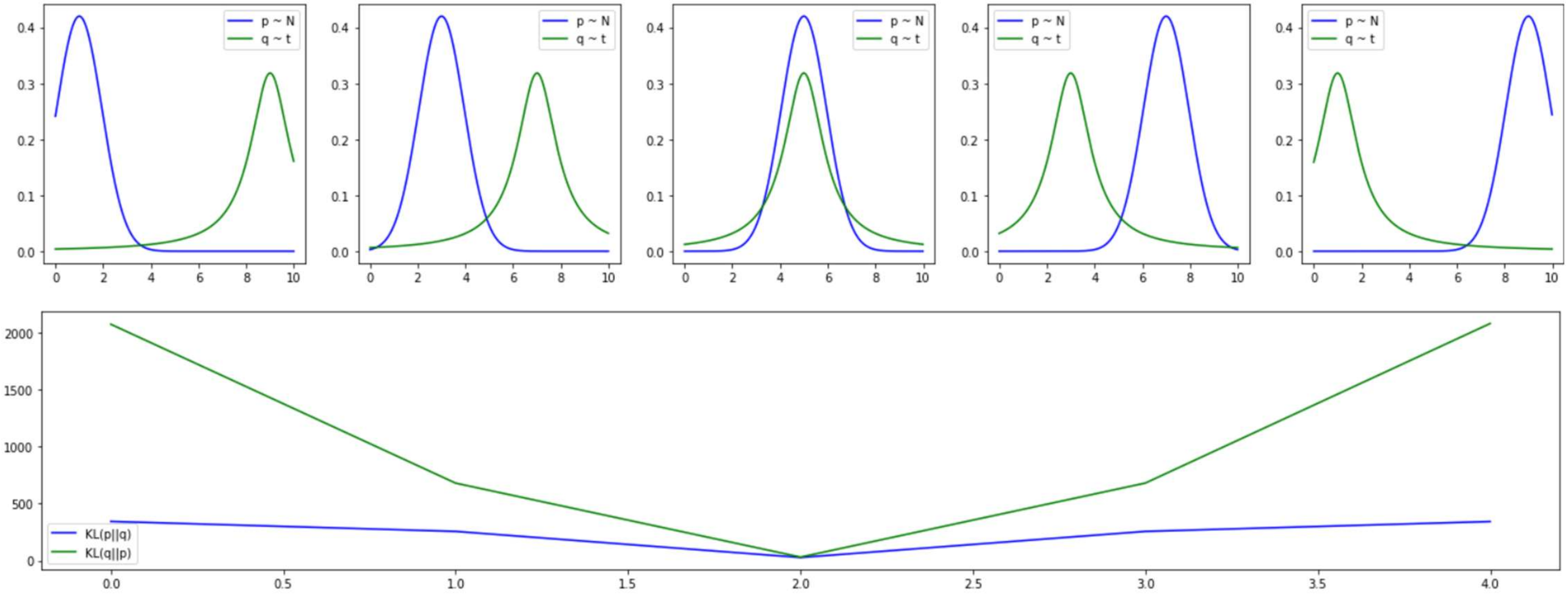
- KL-Divergence Minimization, a.k.a. Variational Bayes, a.k.a. KL-Method
- Laplace Approximation
- Expectation Propagation
- Variational Bounds (KL-method on each likelihood term)
- …

Sampling

- Rejection
- Importance Sampling
- MCMC
  - Metropolis (Hastings)
  - Gibbs
  - Hamiltonian
- …

KL-Divergence (a.k.a. relative entropy)

$$\mathrm{KL}(p(\mathbf{f})||q(\mathbf{f})) = \int p(\mathbf{f}) \ln \frac{p(\mathbf{f})}{q(\mathbf{f})} d\mathbf{f}$$



KL-Divergence Minimization a.k.a. Variational Bayes

$$p(\mathbf{f}|\mathbf{y}, X, \theta) = \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|X,\theta)}{p(\mathbf{y}|X,\theta)} = \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|X,\theta)}{\int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|X,\theta)d\mathbf{f}} = \frac{p(\mathbf{f}|X,\theta)\prod_{i=1}^{n} p(y_i|fi)}{Z}$$

$$\approx q(\mathbf{f}|\mathbf{y}, X, \theta) = \mathcal{N}(\mathbf{f}|\mu, \Sigma)$$

$$\mathrm{KL}(q(\mathbf{f})||p(\mathbf{f}|\mathbf{y})) = \int q(\mathbf{f}) \ln \frac{q(\mathbf{f})}{p(\mathbf{f}|\mathbf{y})} d\mathbf{f} = \left\langle \ln \frac{q(\mathbf{f})}{p(\mathbf{f}|\mathbf{y})} \right\rangle_{q(\mathbf{f})} = \int q(\mathbf{f}) \ln \frac{q(\mathbf{f})p(\mathbf{y})}{p(\mathbf{y}|\mathbf{f})p(\mathbf{f})} d\mathbf{f}$$

$$\ln p(\mathbf{y}) = \langle \ln p(\mathbf{y}|\mathbf{f}) \rangle_{q(\mathbf{f})} - \mathrm{KL}(q(\mathbf{f})||p(\mathbf{f})) + \mathrm{KL}(q(\mathbf{f})||p(\mathbf{f}|\mathbf{y}))$$

$$\ln p(\mathbf{y}) \geq \langle \ln p(\mathbf{y}|\mathbf{f}) \rangle_{q(\mathbf{f})} - \mathrm{KL}(q(\mathbf{f})||p(\mathbf{f}))$$

Fixed when varying *q* make big as possible becomes very small

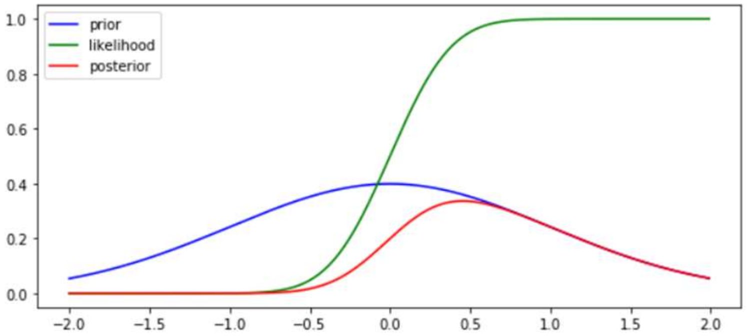
Laplace approximation

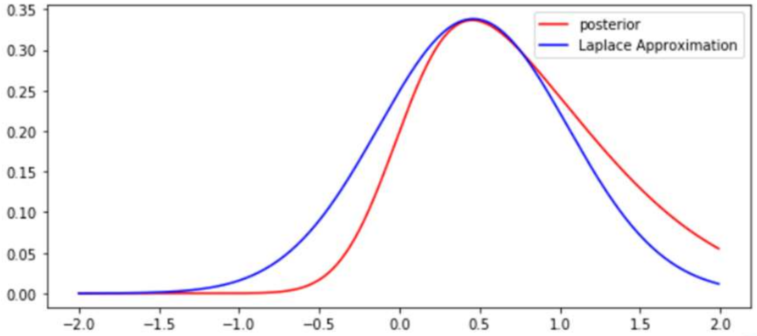
Second order Taylor expansion around posterior mode

$$p(\mathbf{f}|\mathbf{y}, X, \theta) = \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|X,\theta)}{p(\mathbf{y}|X,\theta)} = \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|X,\theta)}{\int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|X,\theta)d\mathbf{f}}$$

$$p(\mathbf{f}|\mathbf{y}, X, \theta) \approx q(\mathbf{f}|\mathbf{y}, X, \theta) = \mathcal{N}(\mathbf{f}|\mu, \Sigma)$$

$$\mu = \underset{\mathbf{f}\in\mathbb{R}^n}{\operatorname{argmax}}\, p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|X, \theta), \Sigma = \left(K^{-1} + -\frac{\partial^2 \ln p(\mathbf{y}|\mathbf{f})}{\partial\mathbf{f}\partial\mathbf{f}^{\mathrm{T}}}\bigg|_{\mathbf{f}=\mu}\right)^{-1}$$





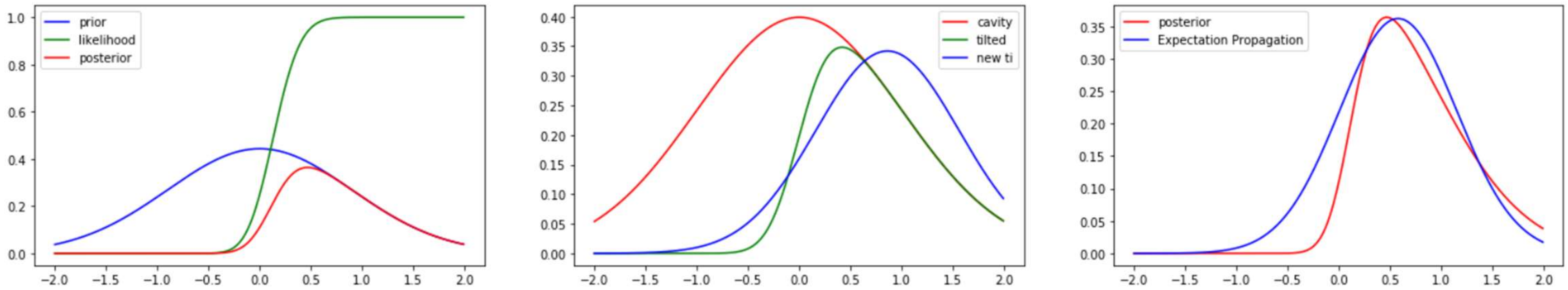
Expectation Propagation

Iteratively replace likelihood factors one by one with unnormalized Gaussian

At each step minimize KL-divergence (for Gaussian match moments)

$$p(\mathbf{f}|\mathbf{y},X,\theta) = \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|X,\theta)}{p(\mathbf{y}|X,\theta)} = \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|X,\theta)}{\int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|X,\theta)d\mathbf{f}} = \frac{p(\mathbf{f}|X,\theta)\prod_{i=1}^{n} p(y_i|fi)}{Z}$$

$$\approx q(\mathbf{f}|\mathbf{y},X,\theta) = \mathcal{N}(\mathbf{f}|\mu,\Sigma) = \frac{p(\mathbf{f}|X,\theta)\prod_{i=1}^{n} t_i\left(f_i|\tilde{Z}_i,\tilde{\mu}_i,\tilde{\sigma}_i\right)}{Z_{EP}}$$

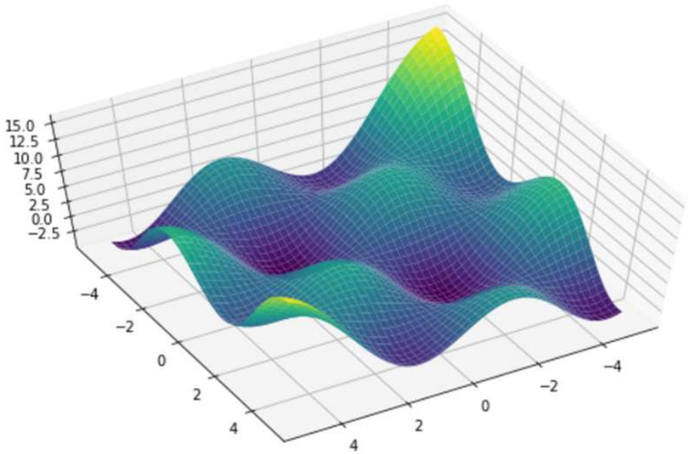


https://tminka.github.io/papers/ep/roadmap.html

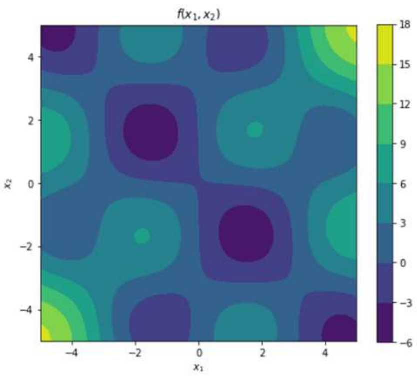
Comparison

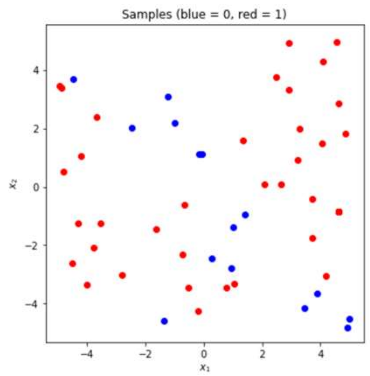
	Pros	Cons	When
VB	Directly lowering divergence	Can become overly confident !!!	Wide range
LA	Simple, fast	Poor when mode does not describe posterior well (e.g. Bernouilli)	When mode does describe posterior well (e.g. Poisson)
EP	Lends itself to sparse approximations	Slow, might not converge	Binary data

LA vs EP

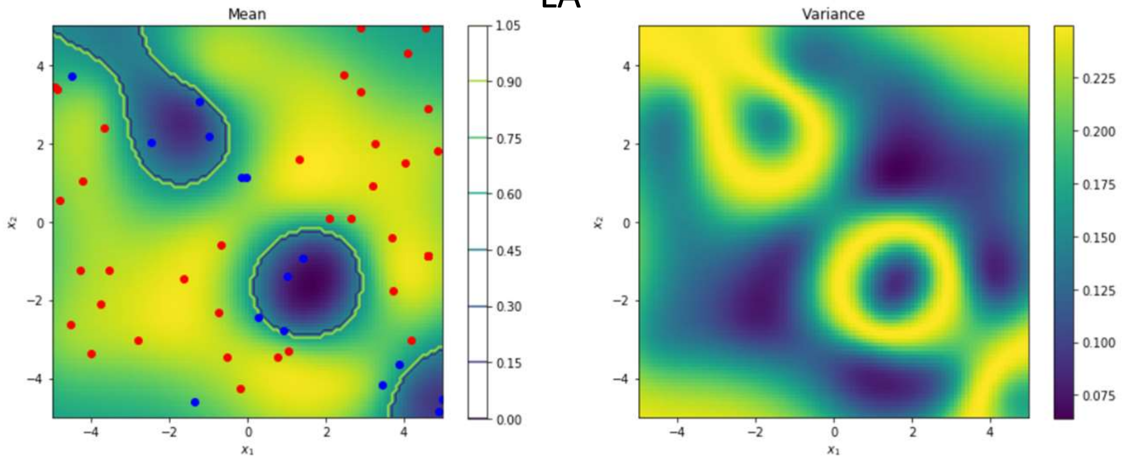


$$f(x_1, x_2) = 5 \times \sin x_1 \times \sin x_2 + \ 0.05 \times (2 \times x_1 + x_2)^2 - 0.1$$

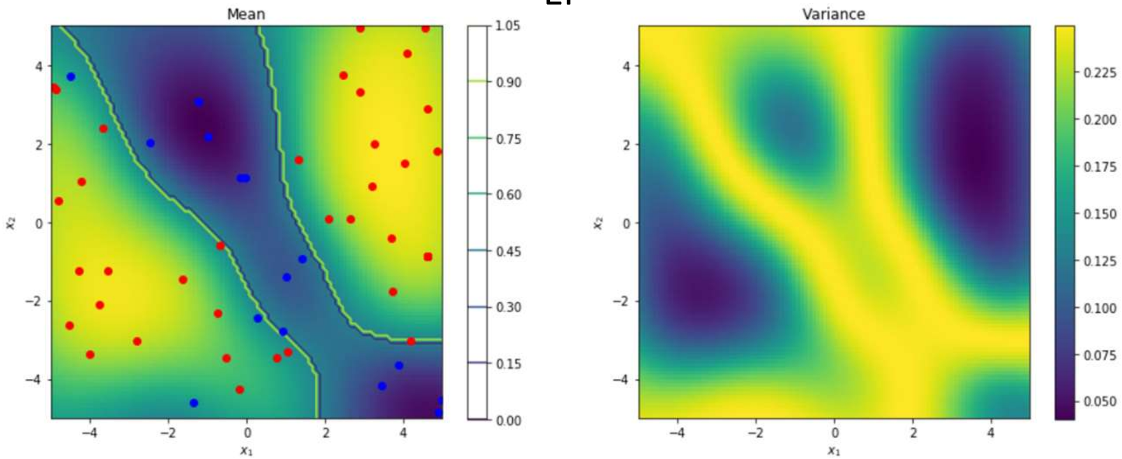




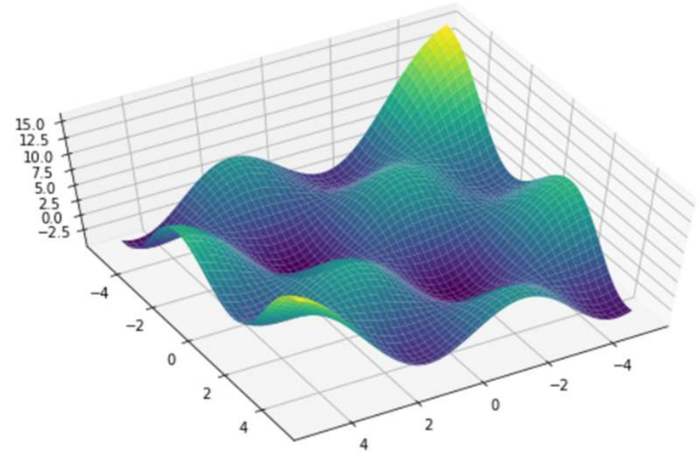
LA



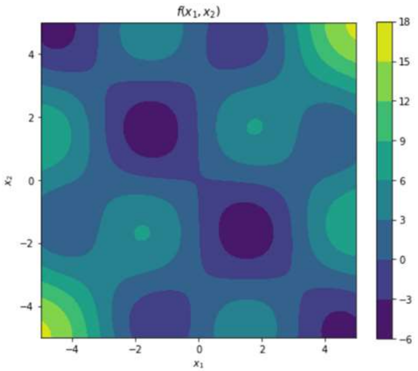
EP

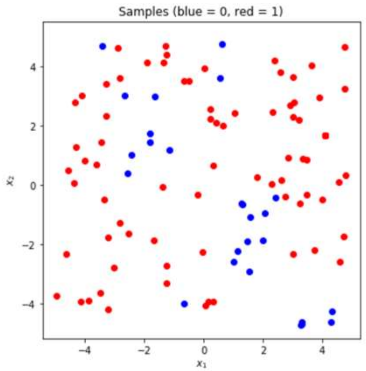


LA vs EP

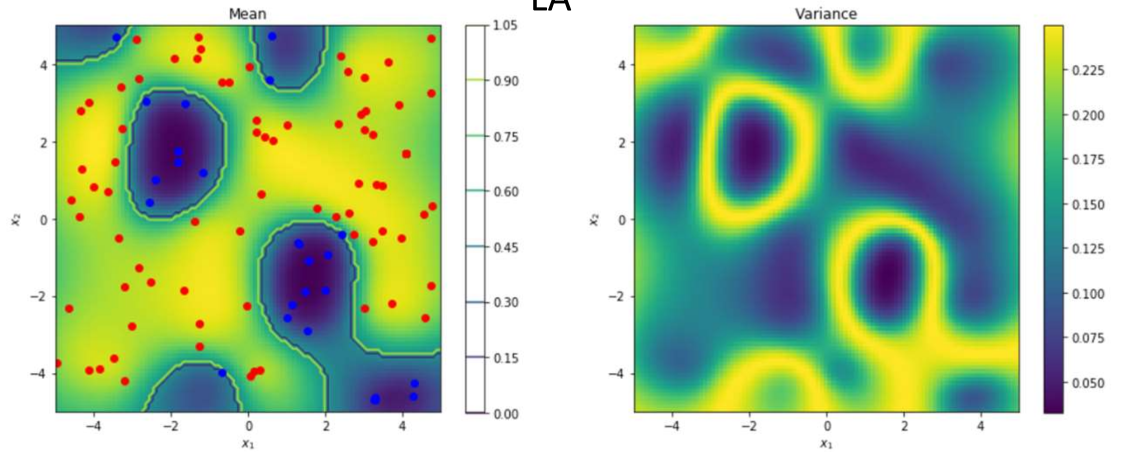


$$f(x_1, x_2) = 5 \times \sin x_1 \times \sin x_2 + \ 0.05 \times (2 \times x_1 + x_2)^2 - 0.1$$

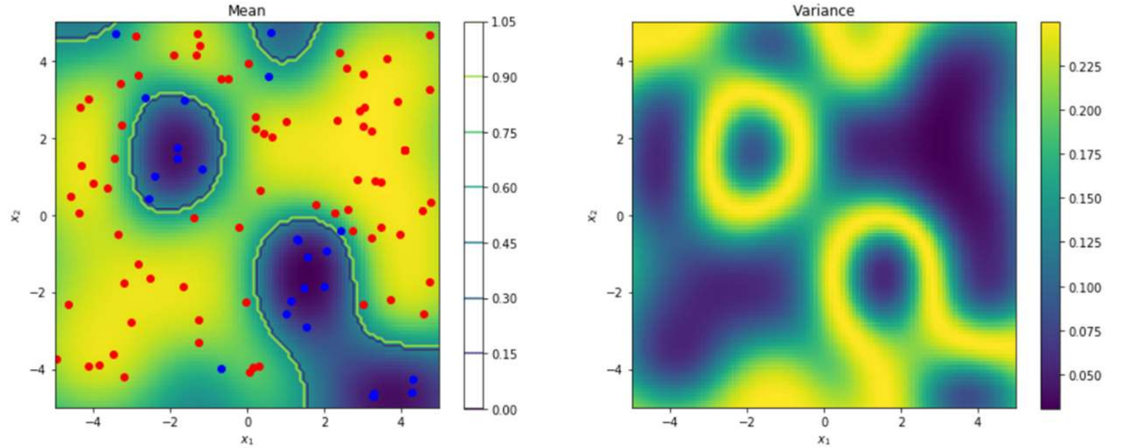




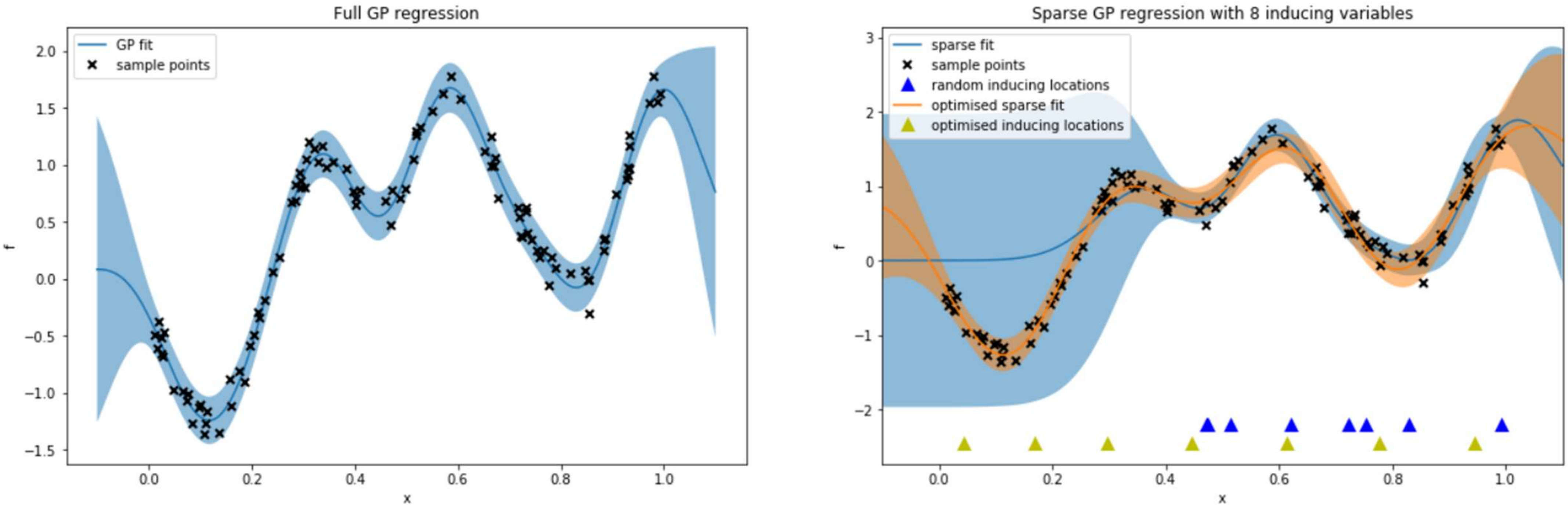
LA



EP



Approximations for Large Datasets a.k.a. Sparse GP

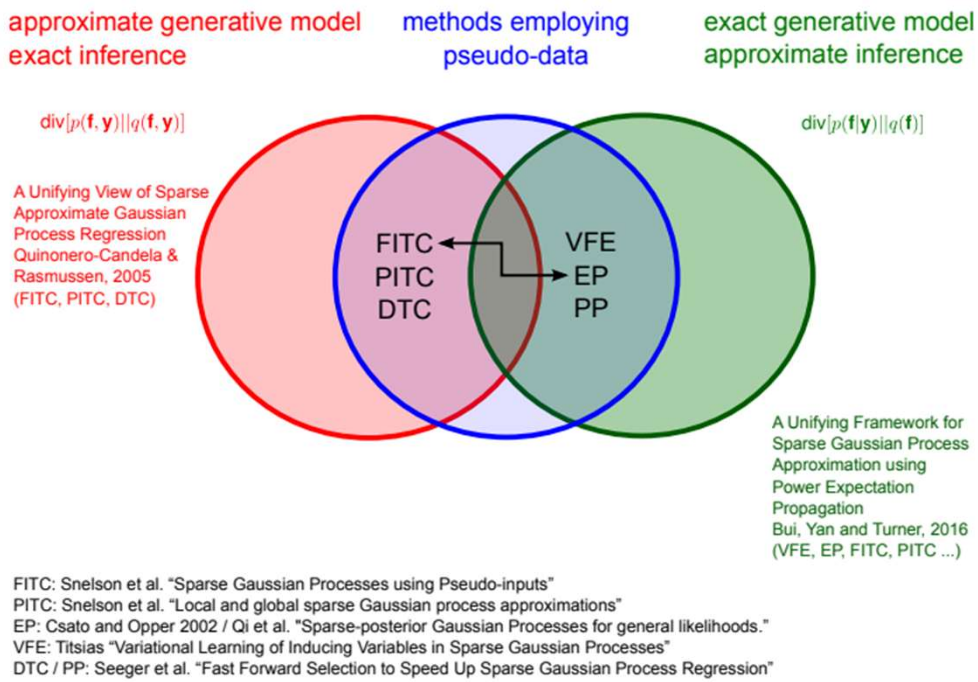


Approximations for Large Datasets a.k.a. Sparse GP

GP are $\mathcal{O}(n^3)$ in complexity and $\mathcal{O}(n^2)$ in storage

Family of approximations:

- Nyström approximation, subset of random data
- Fully independent training conditional (FITC), pseudo data, gradient optimization
- Variational sparse GP combines FITC with (power) EP



Richard E. Turner, "Sparse Gaussian Process Approximations", GPSS17, http://gpss.cc/gpss17/slides/gp-approx-new.pdf

Bayesian Optimization – Active Learning

Bayesian Optimization: find min (or max) in unknown function …

Active Learning: find whole picture of unknow function …

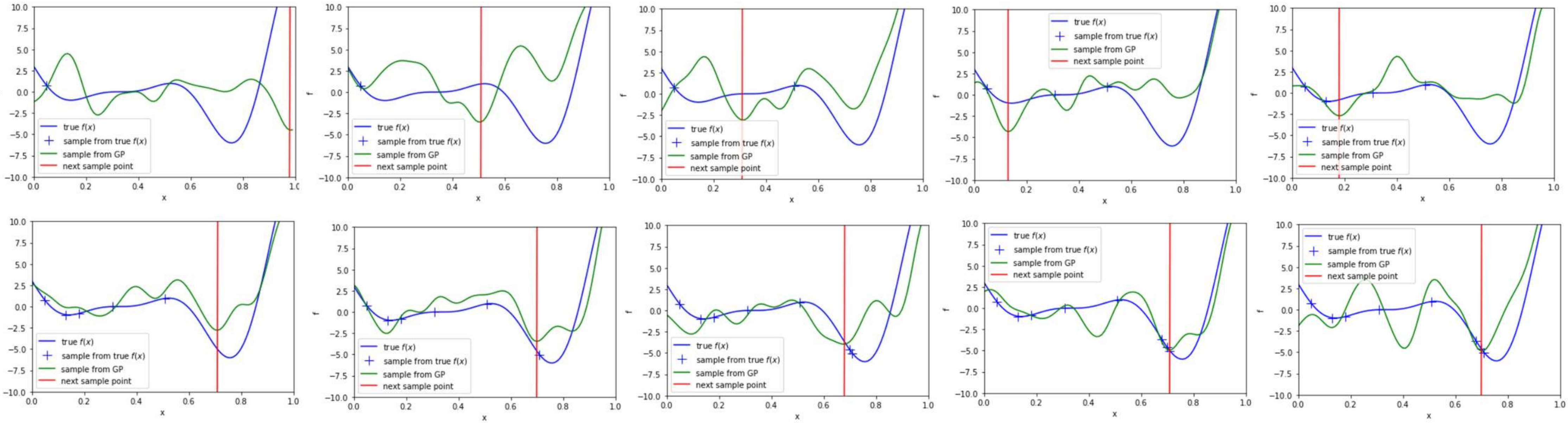
… in as few steps as possible.

Exploration / exploitation trade-off

Kriging in geostatistics

Danie Krige learning the trade, 1939.

Bayesian Optimization – Thompson sampling

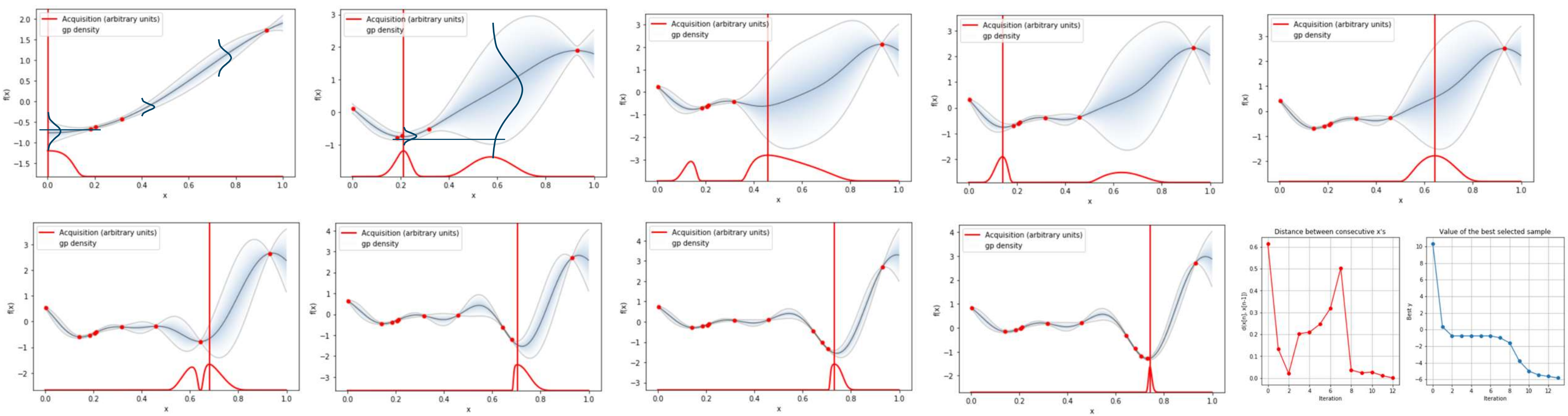


Forrester function: $f(x) = (6x-2)^2 \sin(12x-4)$

Bayesian Optimization

Aquisition function: Probability of Improvement

$$x = \underset{x}{\operatorname{argmax}} \, \mathrm{PI}(x), \mathrm{PI}(x) = p(f(x) \geq \mu_{max} + \varepsilon) = \Phi(Z), Z = \frac{\mu(x) - \mu_{max} - \varepsilon}{\sigma(x)}$$



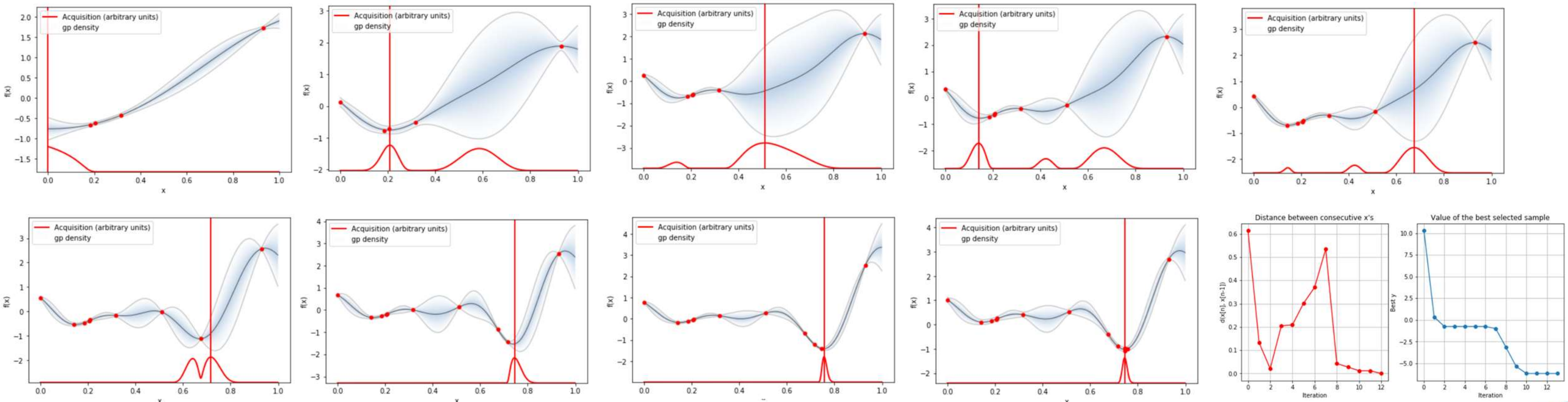
Forrester function: $f(x) = (6x-2)^2 \sin(12x-4)$

Bayesian Optimization

Aquisition function: Expected Improvement

$$x = \underset{x}{\operatorname{argmax}}\, \mathrm{E}(max\{0, f_{n+1}(x) - f_{max}\}|D_n)$$

$$\mathrm{EI}(x) = \begin{cases} (\mu(x) - \mu_{max} - \varepsilon)\Phi(Z) + \sigma(x)\phi(Z), if\ \sigma(x) > 0 \\ 0, if\ \sigma(x) = 0 \end{cases}, Z = \frac{\mu(x) - \mu_{max} - \varepsilon}{\sigma(x)}$$



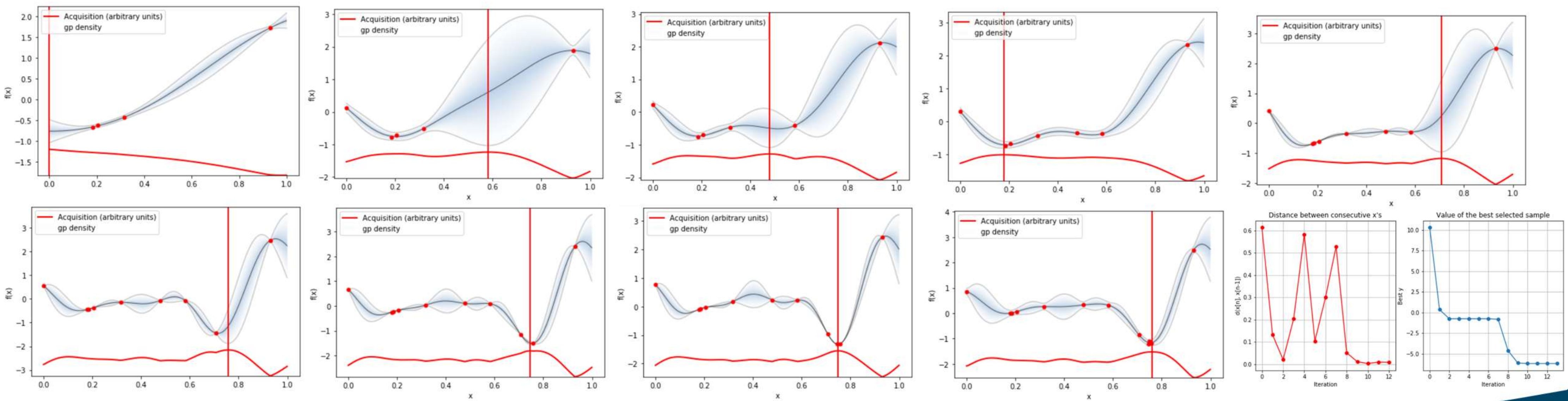
Forrester function: $f(x) = (6x - 2)^2 \sin(12x - 4)$

Bayesian Optimization

Aquisition function: Upper (Lower) Confidence Bounds

regret $r(x) = f_{max} - f_x$, cumulative regret $R_T = r(x_1) + r(x_2) + \; \ldots r(x_T)$

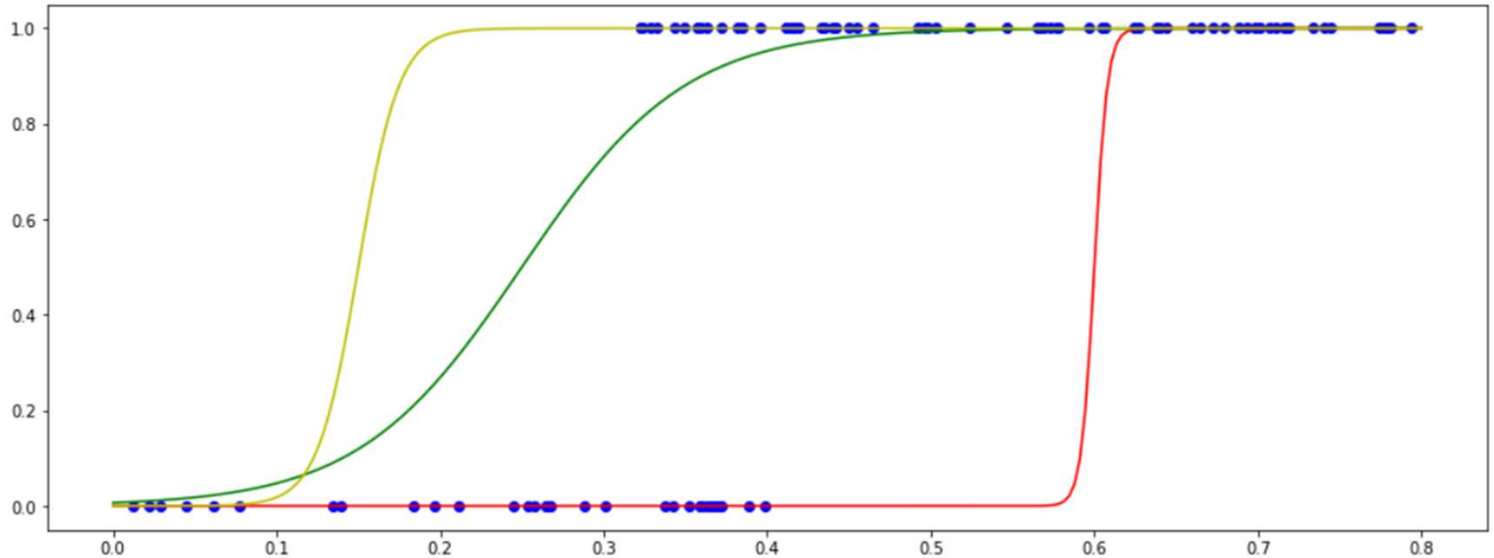
$\mathrm{GP} - \mathrm{UCB}(x) = \mu(x) + \sqrt{\nu\beta_t}\sigma(x)$, exploitation / exploration trade-off



Forrester function: $f(x) = (6x - 2)^2 \sin(12x - 4)$

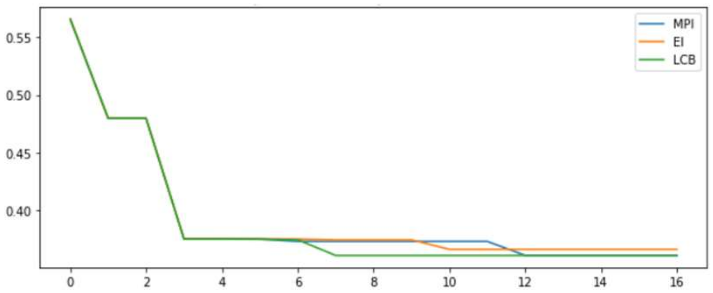
Bayesian Optimization Example: logistic regression parameters

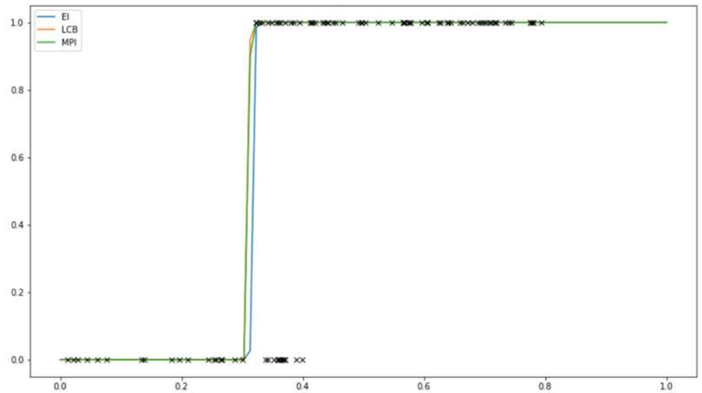
$$h_\theta(x) = \frac{1}{1 + e^{-z}}, z = \theta_0 + \ \theta_1 x$$

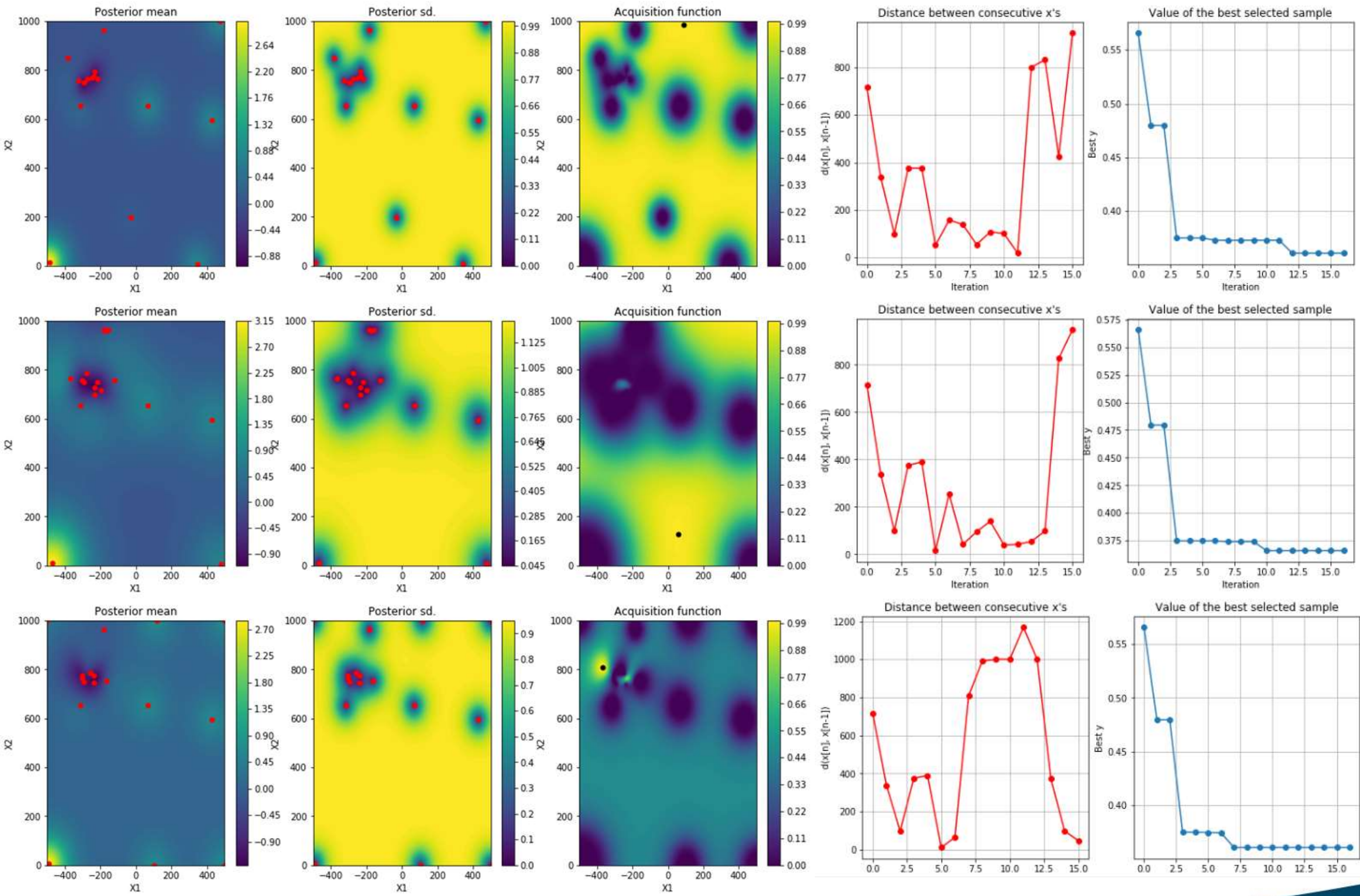


Bayesian Optimization Example: logistic regression parameters

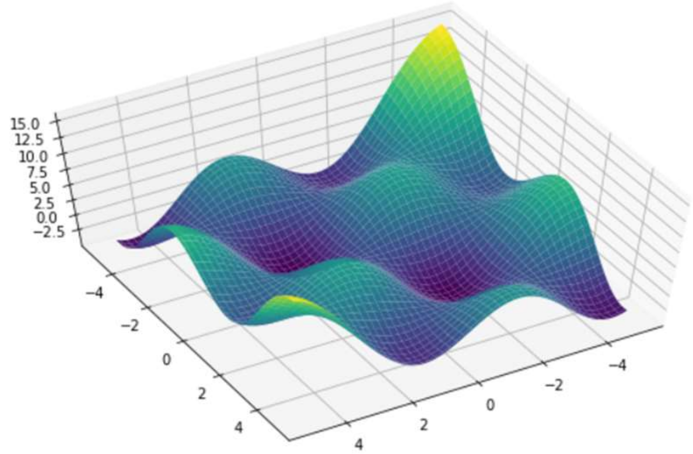
$$h_\theta(x) = \frac{1}{1+e^{-z}}, z = \theta_0 + \ \theta_1 x$$



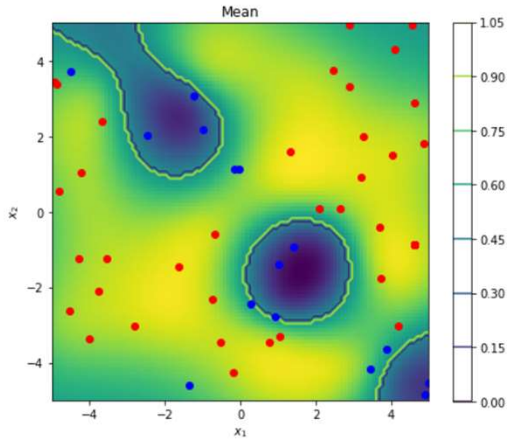


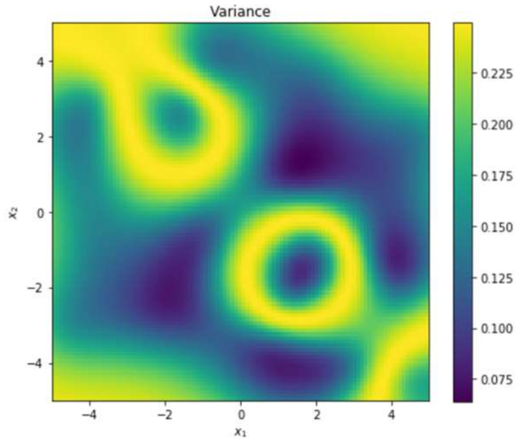


Active Learning in classification

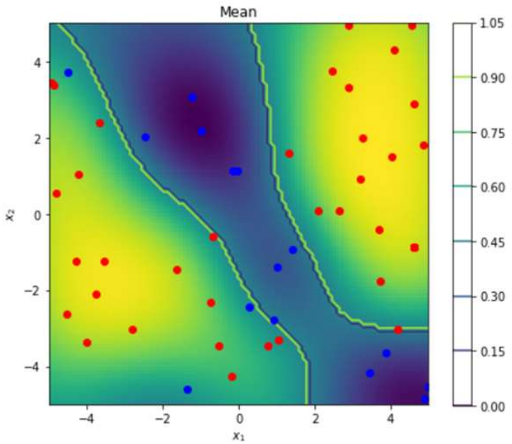


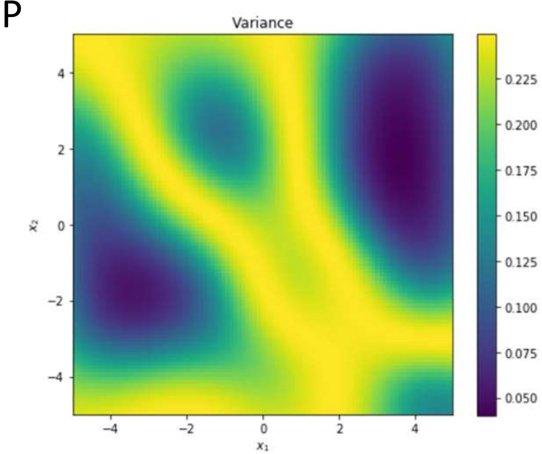
LA



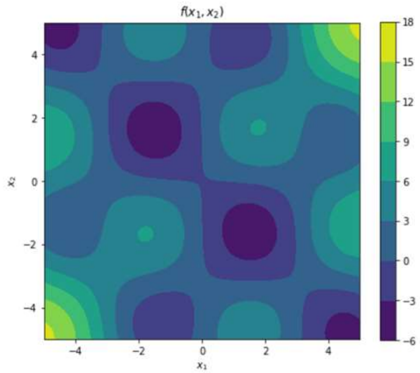


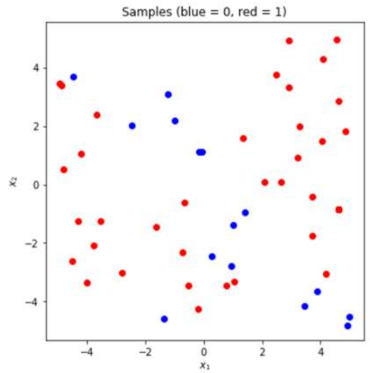
EP





$$f(x_1, x_2) = 5 \times \sin x_1 \times \sin x_2 + \; 0.05 \times (2 \times x_1 + x_2)^2 - 0.1$$

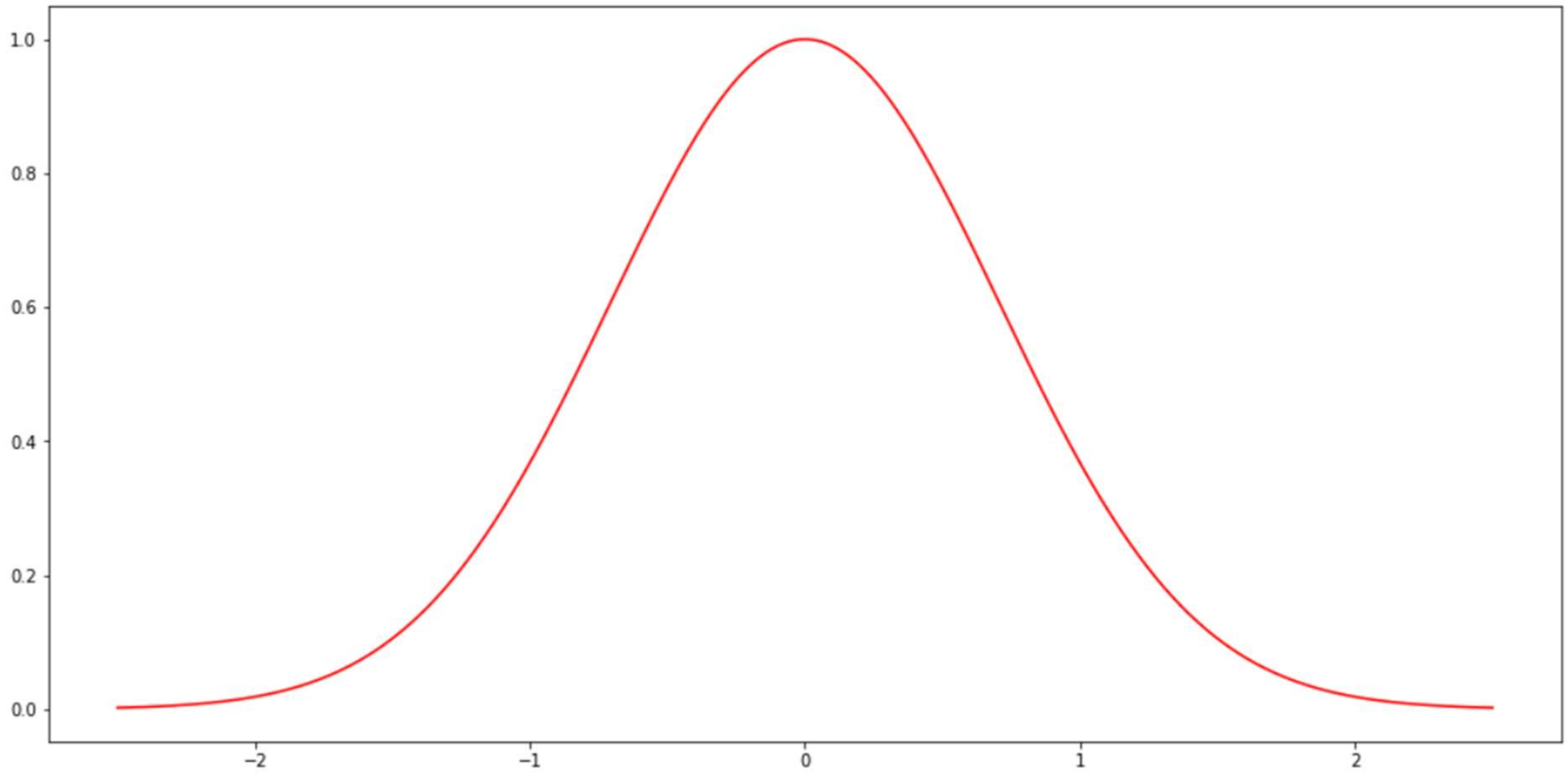




Nickisch H., Rasmussen C. E., Approximations for Binary Gaussian Process Classification, Journal of Machine Learning Research 9 (2008) 2035-2078, http://www.jmlr.org/papers/volume9/nickisch08a/nickisch08a.pdf

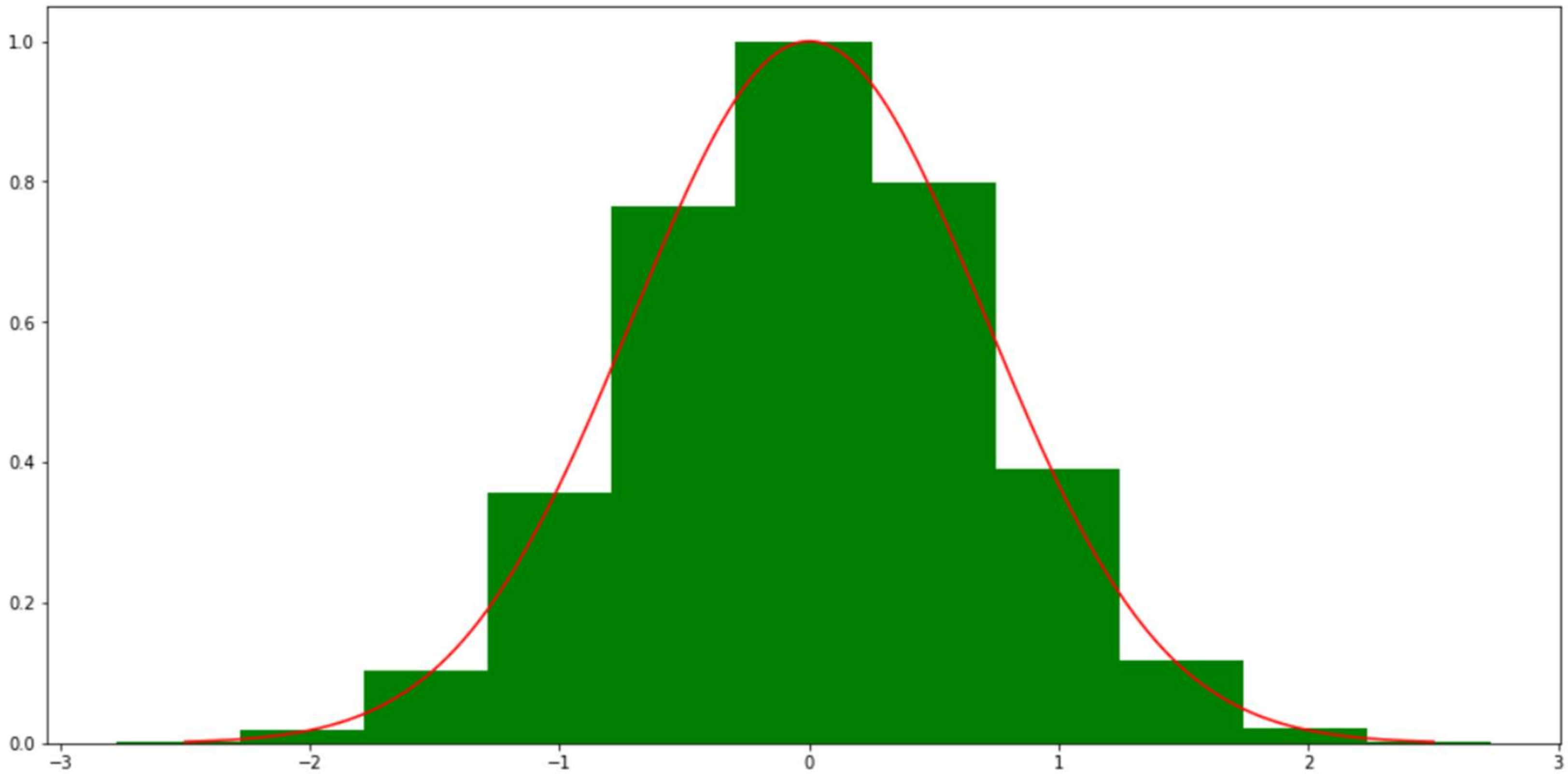
Bayesian Quadrature example

$$\int_{-\infty}^{+\infty} e^{-x^2}\,dx \approx ?$$



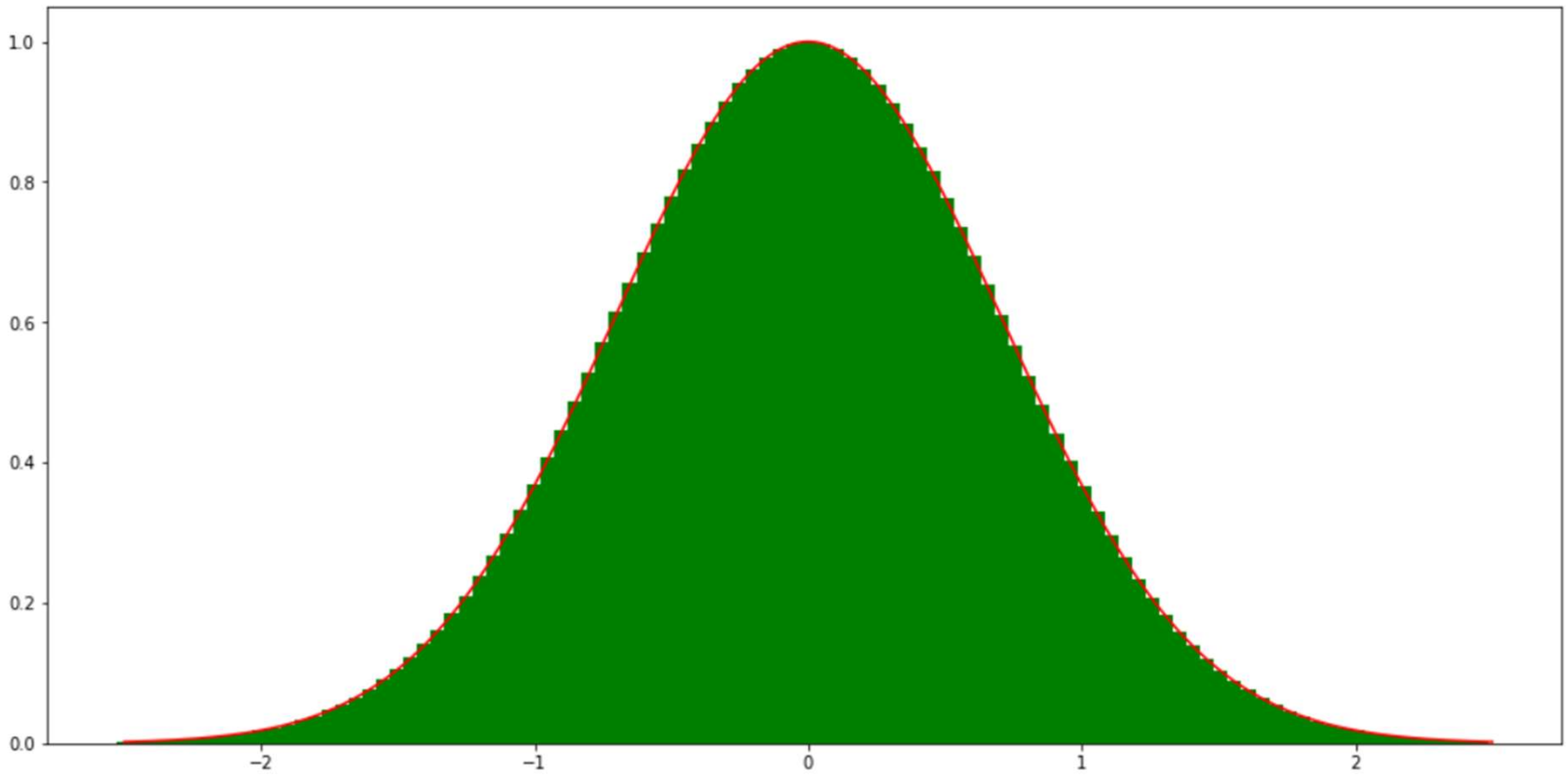
Bayesian Quadrature example

$$\int_{-\infty}^{+\infty} e^{-x^2}\,dx \approx 1{,}626$$



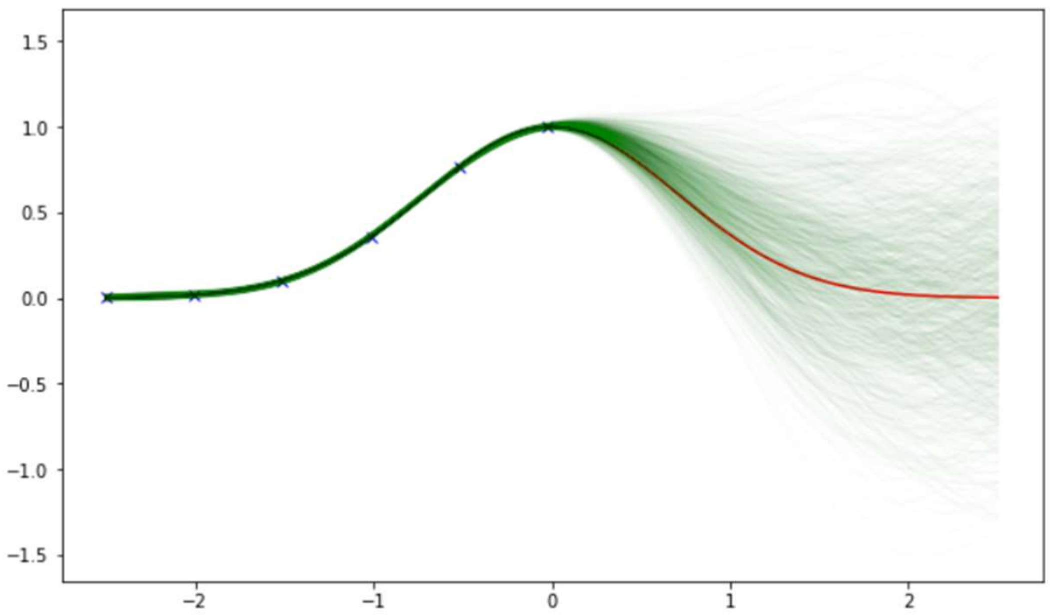
Bayesian Quadrature example

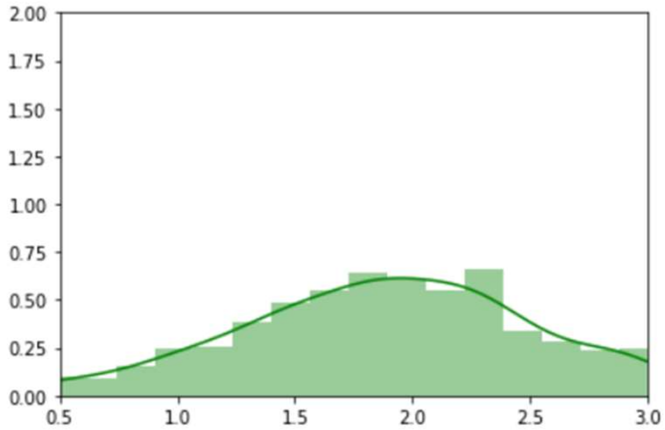
$$\int_{-\infty}^{+\infty} e^{-x^2}\,dx \approx 1{,}756$$



Bayesian Quadrature example

$$\int_{-\infty}^{+\infty} e^{-x^2}\,dx \approx \mu\ 1.92, \sigma^2\ 0.46$$

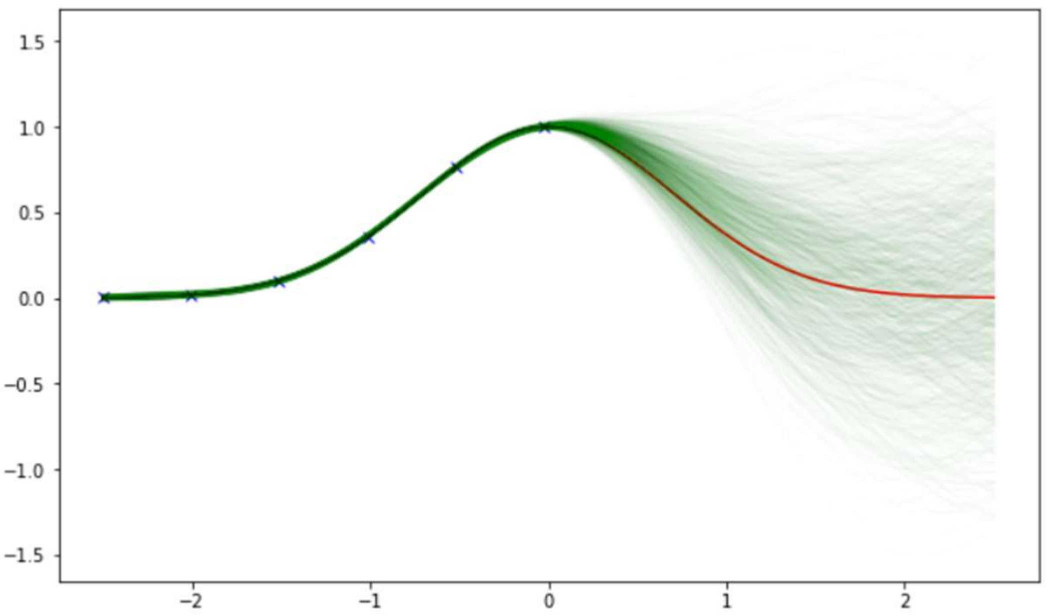


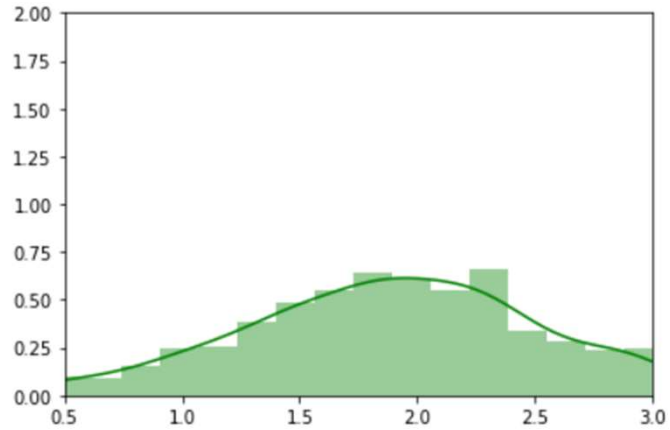


Bayesian Quadrature example

$$\int_{-\infty}^{+\infty} e^{-x^2}\,dx \approx \mu\ 1.92, \sigma^2\ 0.46$$

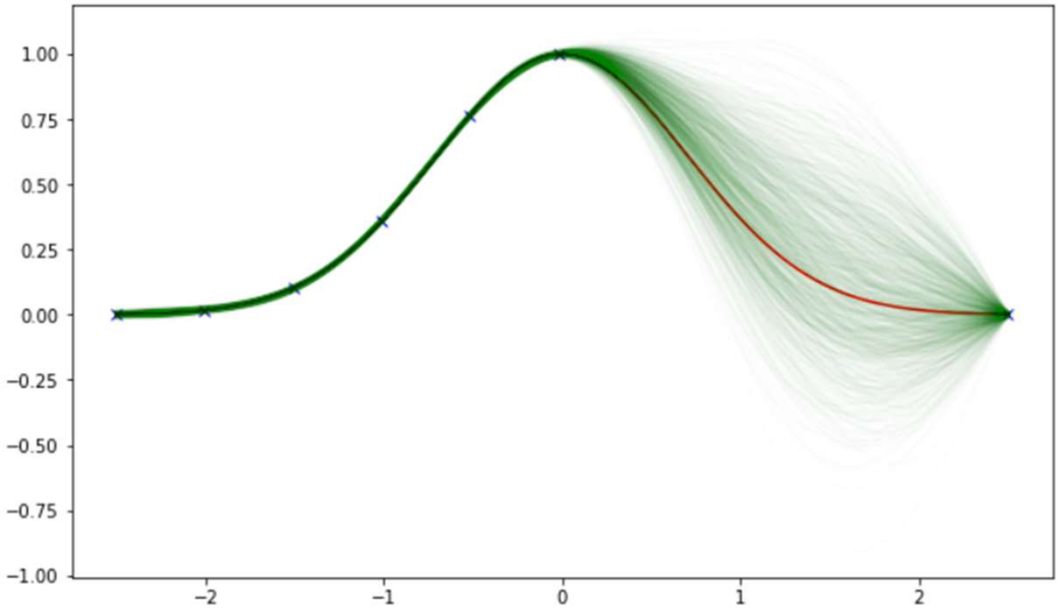
Where to sample next?

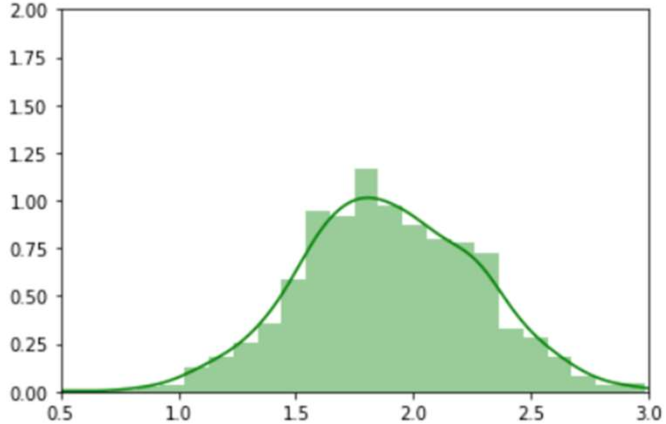




Bayesian Quadrature example

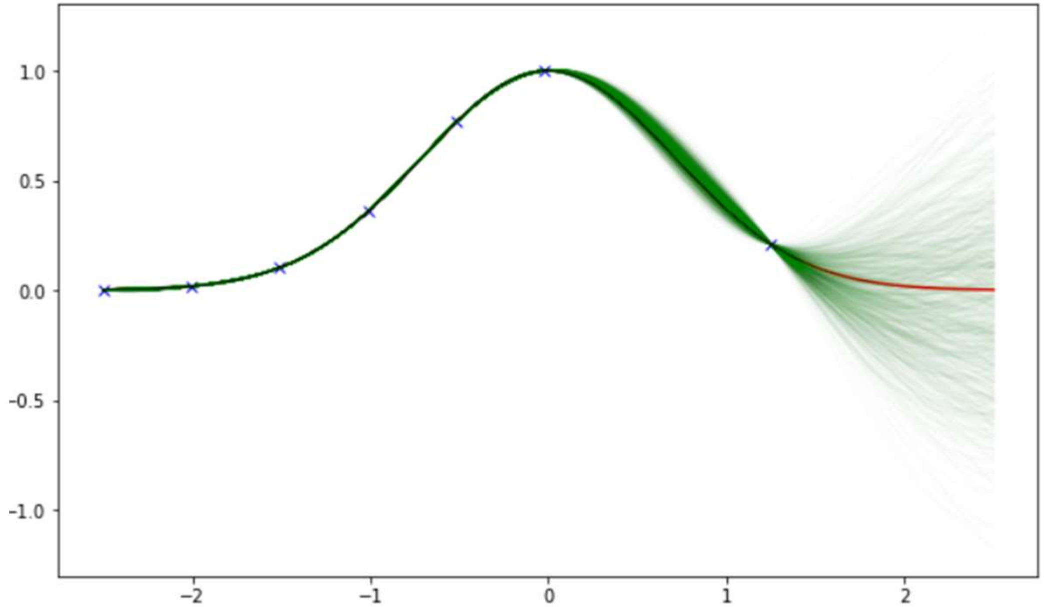
$$\int_{-\infty}^{+\infty} e^{-x^2}\,dx \approx \mu\ 1.89, \sigma^2\ 0.14$$

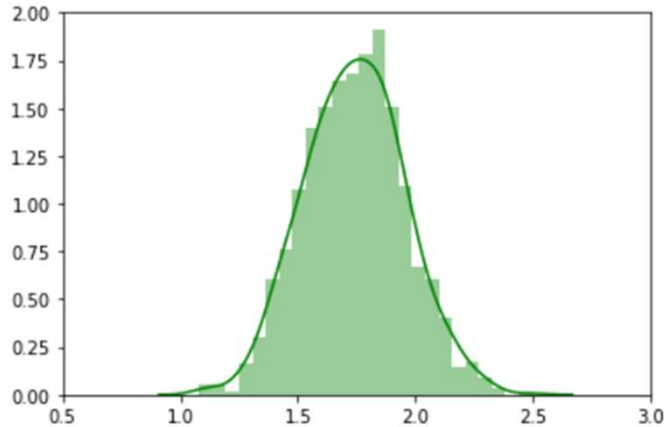




Bayesian Quadrature example

$$\int_{-\infty}^{+\infty} e^{-x^2}\,dx \approx \mu\ 1.74, \sigma^2\ 0.05$$

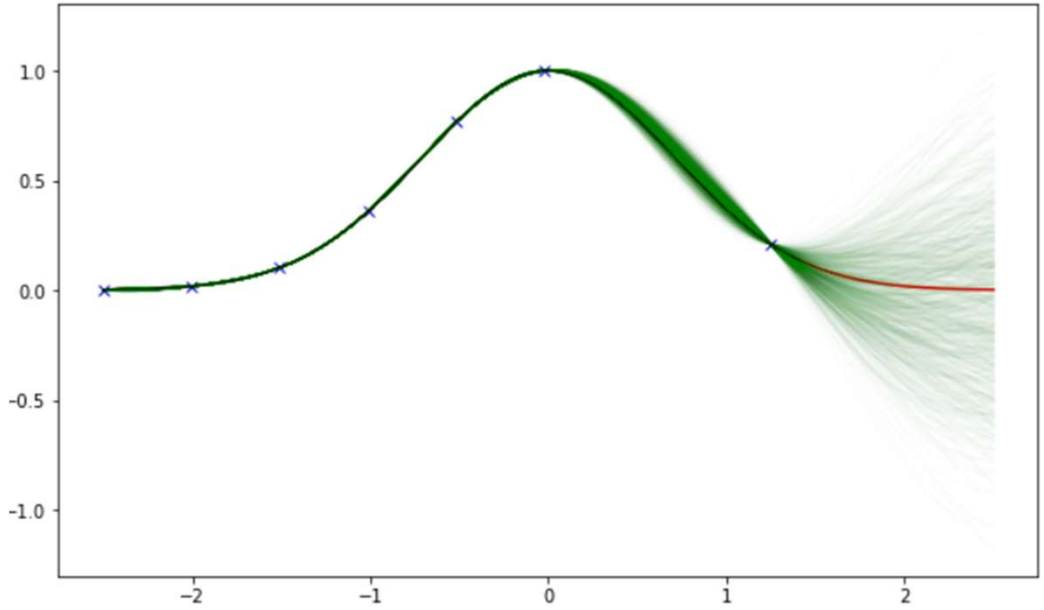


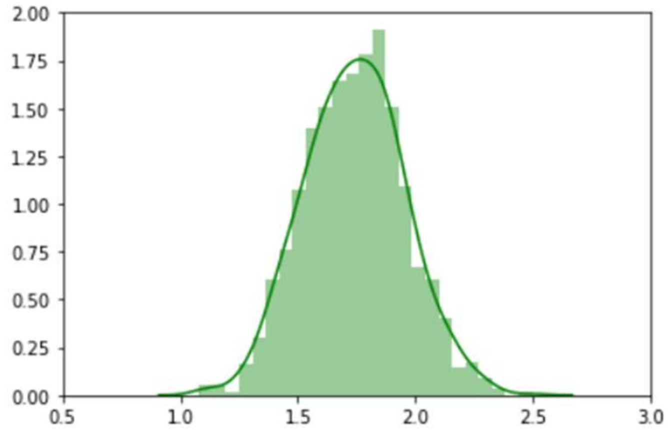


Bayesian Quadrature example

$$\int_{-\infty}^{+\infty} e^{-x^2}\,dx \approx \mu\ 1.74, \sigma^2\ 0.05$$

Aquisition functions: sample to reduce integral or integrand uncertainty, …

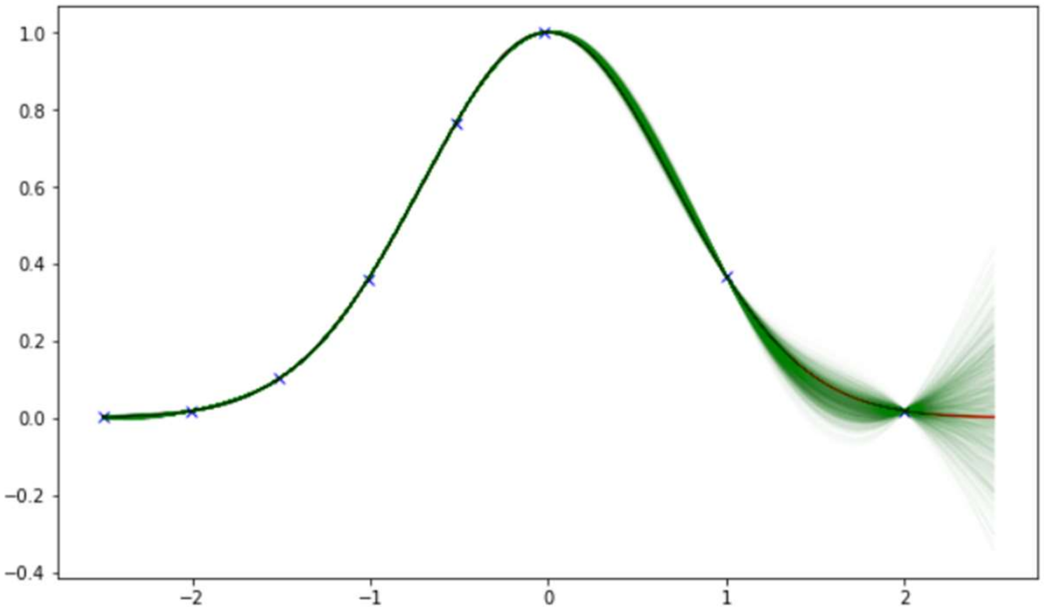


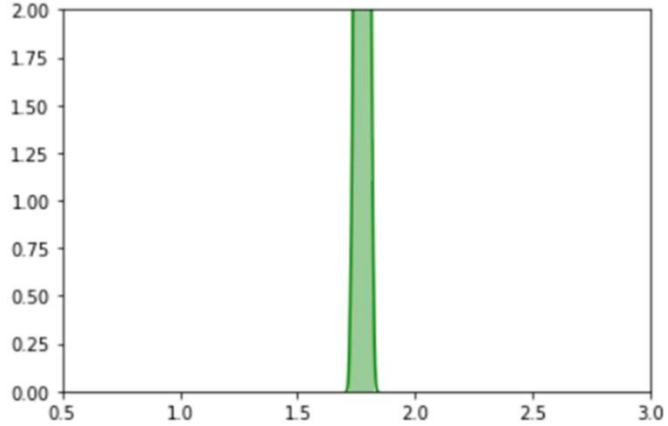


Bayesian Quadrature example

$$\int_{-\infty}^{+\infty} e^{-x^2}\,dx \approx \mu\ 1.7737, \sigma^2\ 0.0003$$

$$\sqrt{\pi} = 1.7725$$





Gaussian Processes and Neural Nets

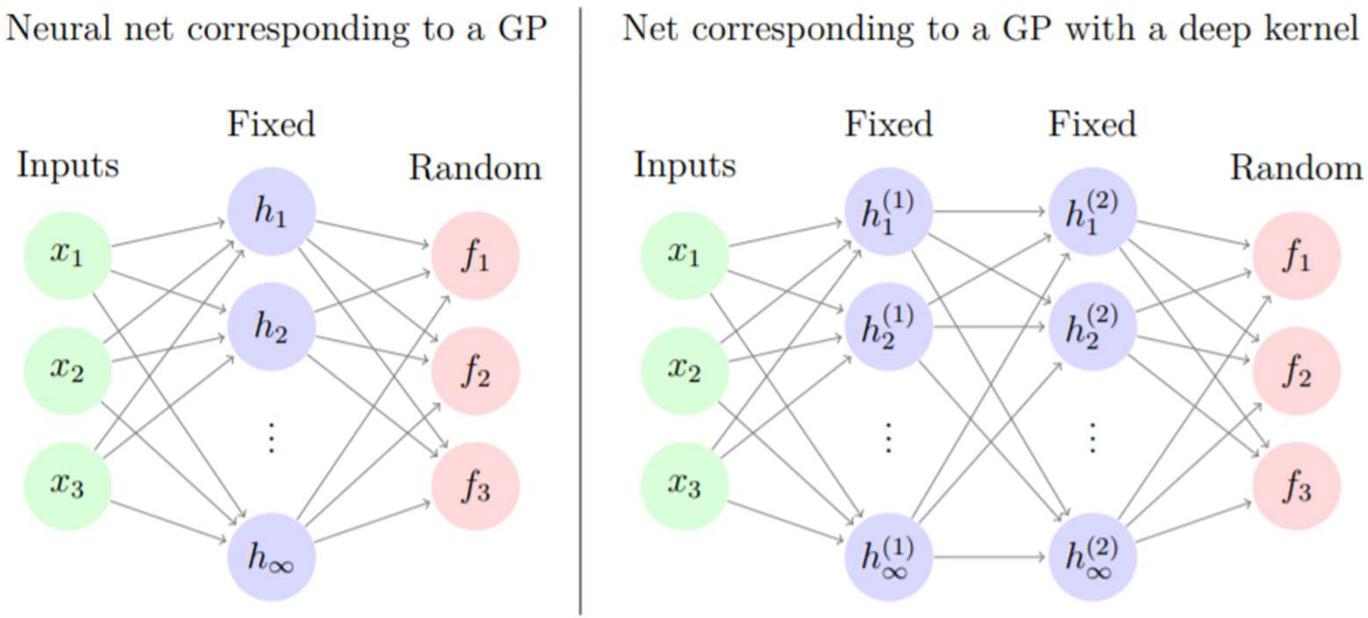


Figure 5.1: *Left:* GPs can be derived as a one-hidden-layer MLP with infinitely many fixed hidden units having unknown weights. *Right:* Multiple layers of fixed hidden units gives rise to a GP with a deep kernel, but not a deep GP.

David Kristjanson Duvenaud, “Automatic Model Construction with Gaussian Processes”, https://www.cs.toronto.edu/~duvenaud/thesis.pdf

Deep GP

$y = f(x) = f_1(f_2(x)), f_1 \sim \mathcal{GP}$ and $f_2 \sim \mathcal{GP}, f: \mathbf{x} \xrightarrow[f_2]{} \mathbf{z} \xrightarrow[f_1]{} y$

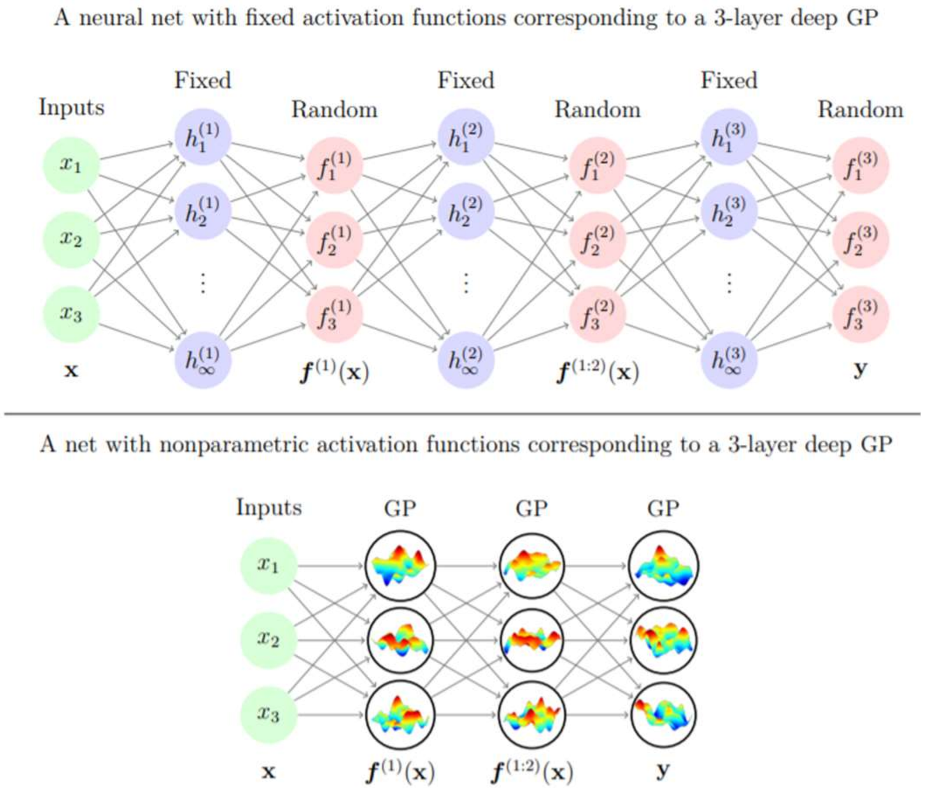
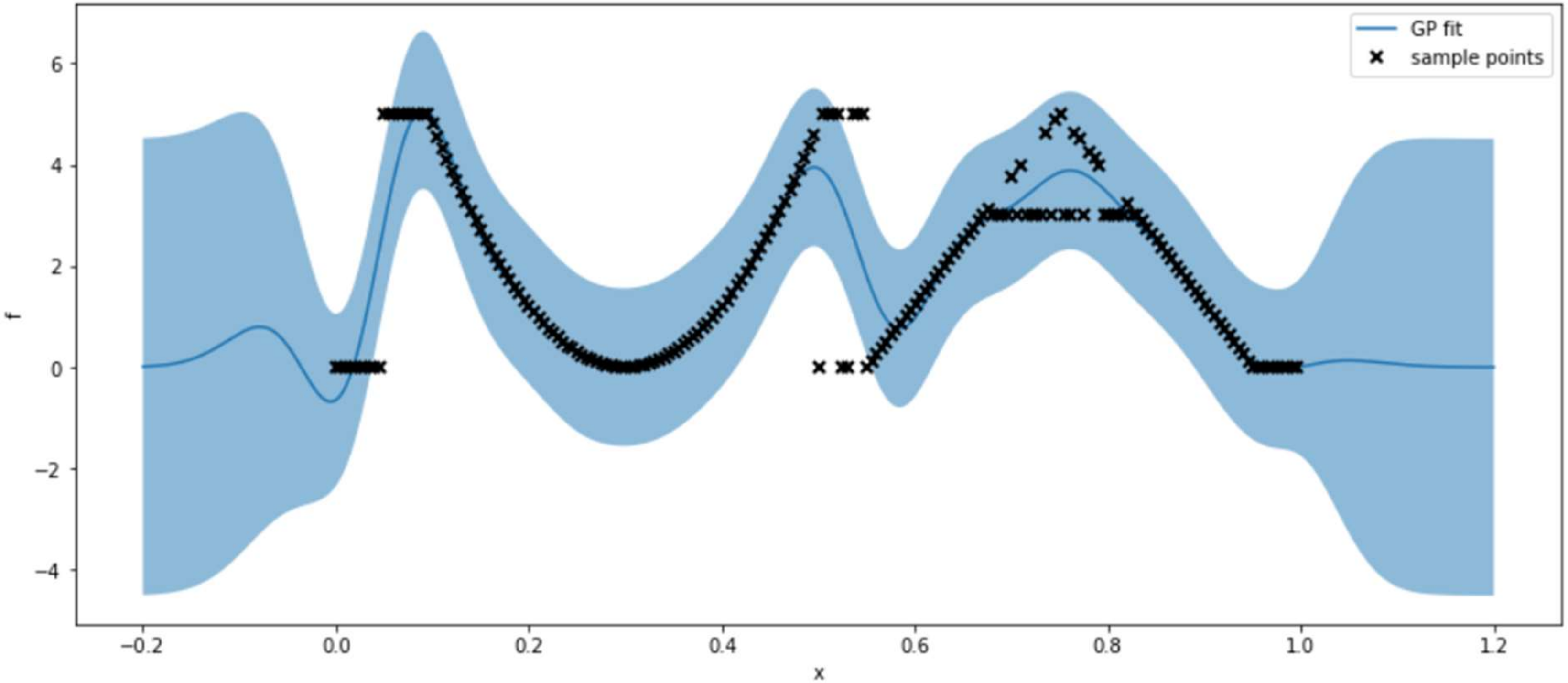


Figure 5.2: Two equivalent views of deep GPs as neural networks. *Top:* A neural network whose every other layer is a weighted sum of an infinite number of fixed hidden units, whose weights are initially unknown. *Bottom:* A neural network with a finite number of hidden units, each with a different unknown non-parametric activation function. The activation functions are visualized by draws from 2-dimensional GPs, although their input dimension will actually be the same as the output dimension of the previous layer.

David Kristjanson Duvenaud, "Automatic Model Construction with Gaussian Processes", https://www.cs.toronto.edu/~duvenaud/thesis.pdf

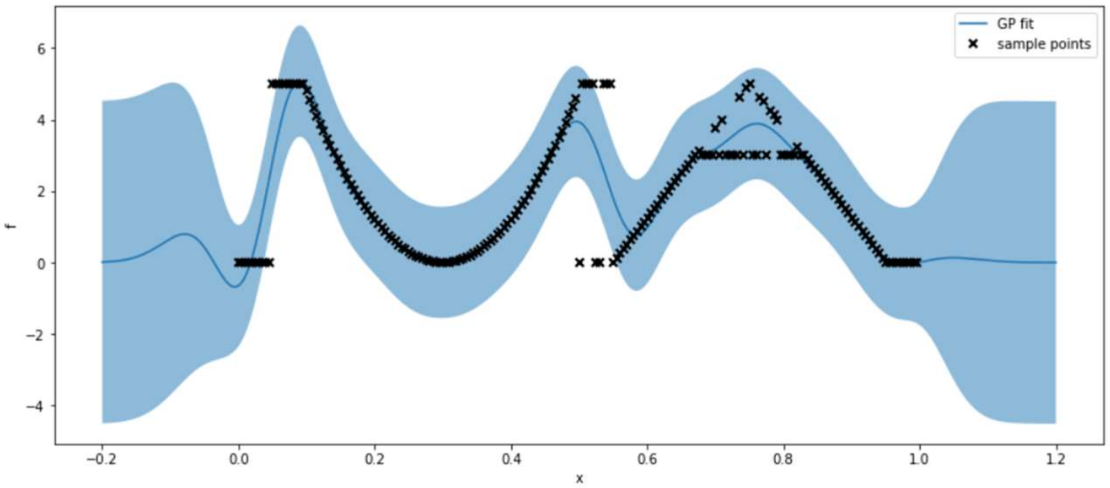
Deep GP

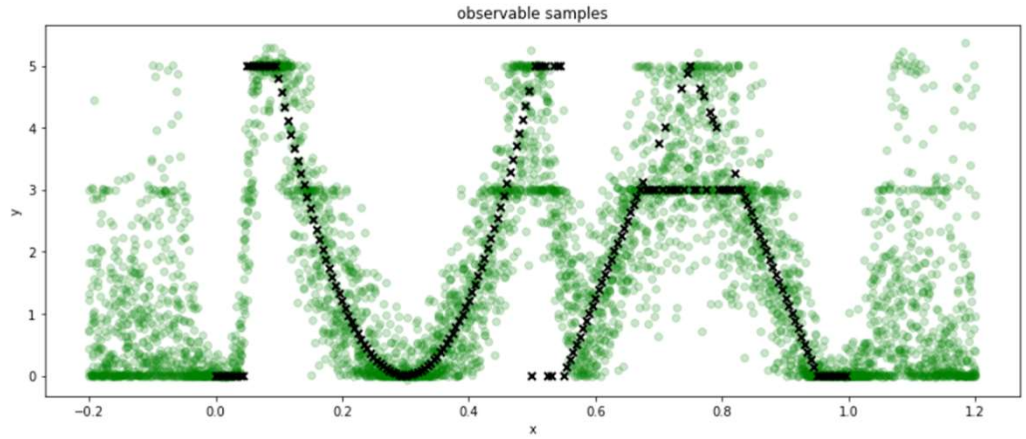
$y = f(x) = f_1(f_2(x)), f_1 \sim \mathcal{GP}$ and $f_2 \sim \mathcal{GP}, f: \mathbf{x} \xrightarrow[f_2]{} \mathbf{z} \xrightarrow[f_1]{} y$

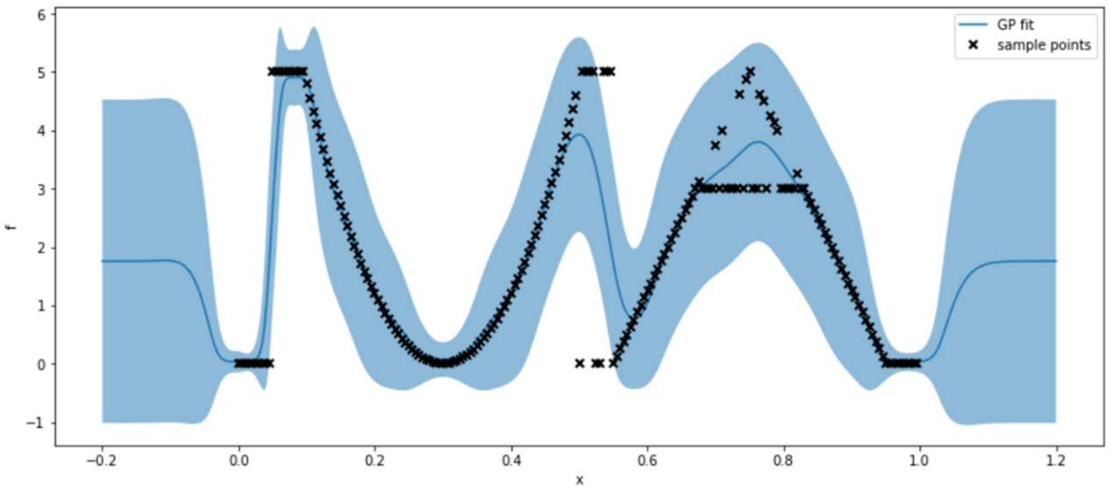


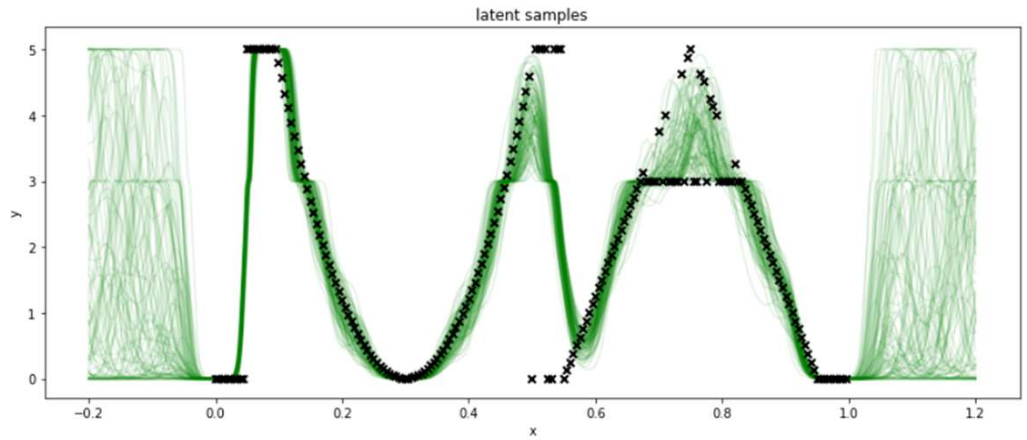
Deep GP

$y = f(x) = f_1(f_2(x)), f_1 \sim \mathcal{GP}$ and $f_2 \sim \mathcal{GP}, f: \mathbf{x} \underset{f_2}{\rightarrow} \mathbf{z} \underset{f_1}{\rightarrow} y$



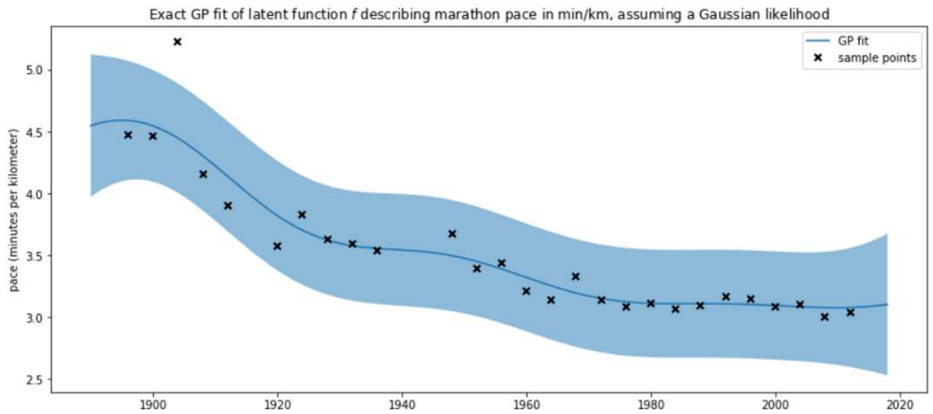


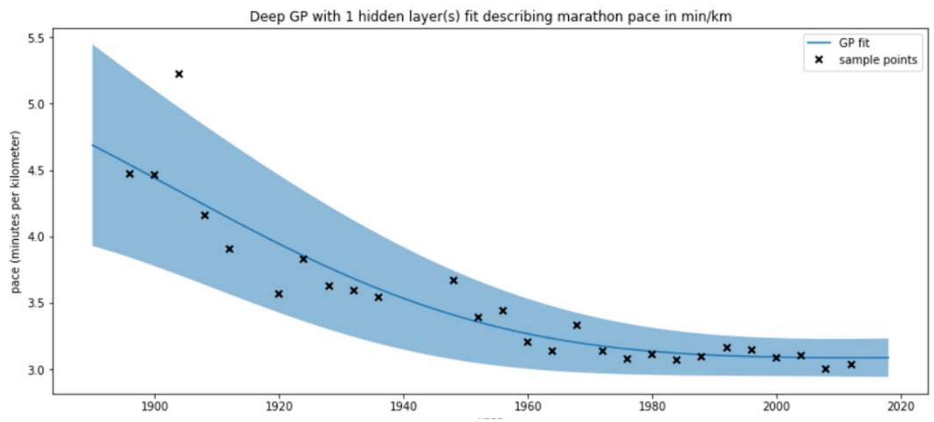


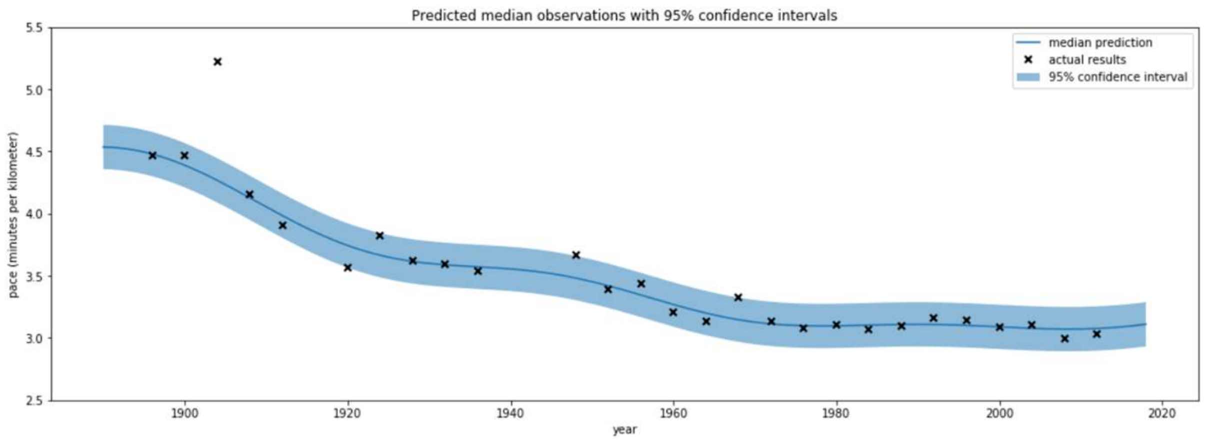


Deep GP

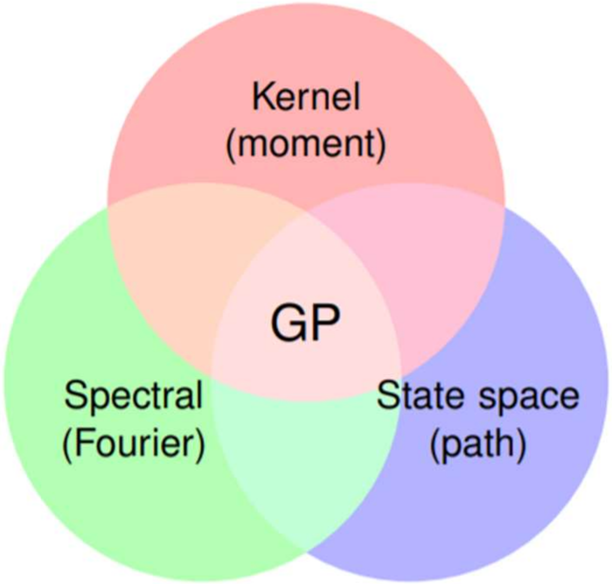
$$y = f(x) = f_1(f_2(x)), f_1 \sim \mathcal{GP} \text{ and } f_2 \sim \mathcal{GP}, f: \mathbf{x} \underset{f_2}{\rightarrow} \mathbf{z} \underset{f_1}{\rightarrow} y$$

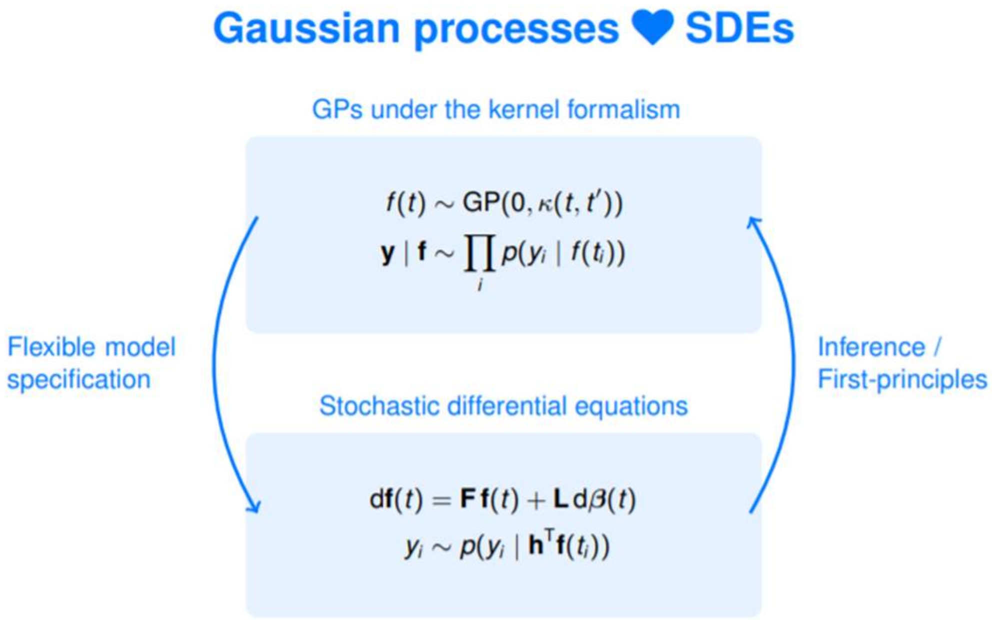






Gaussian Processes and Stochastic Differential Equations





(Exact) inference of the latent functions, can be done in O(n) time and memory complexity by Kalman filtering.

Arno Solin, Aalto University, http://gpss.cc/gpss19/slides/Solin2019.pdf

Software

Write your own

Gpy, GpyOpt, DeepGP by University of Sheffield

GPML by Carl Edward Rasmussen and Hannes Nickisch

C++, Python, Matlab, Octave, R, Tensorflow, sklearn, …

DiceKriging, George, Block GP, Gpmat, mlegp, JMP, GPfit, laGP, DACE, …

Many more: http://www.gaussianprocess.org/#code

References 1

- Thang D Bui, Josiah Yan, and Richard E Turner. A unifying framework for gaussian process pseudo-point approximations using power expectation propagation. Journal of Machine Learning Research, 18:3649–3720, 2017.
- Zhenwen Dai, Andreas Damianou, James Hensman, and Neil D. Lawrence. Gaussian process models with parallelization and gpu acceleration. In NIPS workshop Software Engineering for Machine Learning, 2014.
- Yarin Gal, Mark van der Wilk, and Carl Edward Rasmussen. Distributed variational inference in sparse gaussian process regression and latent variable models. In Advances in Neural Information Processing Systems 27, pages 3257–3265, 2014.
- Joaquin Quiñonero-Candela and Carl Edward Rasmussen. A unifying view of sparse approximate gaussian process regression. Journal of Machine Learning Research, 6:1939—-1959, 2005.
- Edward Snelson and Zoubin Ghahramani. Sparse gaussian processes using pseudo-inputs. In Advances in Neural Information Processing Systems, pages 1257–1264. 2006.
- Michalis Titsias. Variational learning of inducing variables in sparse gaussian processes. In Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics, pages 567–574, 2009.
- Christopher K. I. Williams and Matthias Seeger. Using the Nyström method to speed up kernel machines. In Advances in Neural Information Processing Systems, pages 682–688. 2001.

References 2

- Neal 1994. Bayesian Learning for Neural Networks. PhD thesis
- Dayan et al. 1995.The Helmholtz machine. Neural Computation, 1995.
- Rahimi and Recht 2007. Random Features for Large-Scale Kernel Machines. NeurIPS 2007
- L´azaro-Gredilla et al. 2010. Sparse spectrum Gaussian process regression. JMLR 2010
- Bui et al. 2016. Deep Gaussian Processes for Regression using Approximate Expectation Propagation. ICML 2016
- Li and Gal 2016. Dropout inference in Bayesian neural networks with alpha-divergences. ICML 2017
- Kendall and Gal 2017. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? NeurIPS 2017
- Cutajar et al. 2017. Random Feature Expansions for Deep Gaussian Processes. ICML 2017
- Matthews et al. 2018. Gaussian Process Behaviour in Wide Deep Neural Networks. ICLR 2018
- Lee et al. 2018. Deep Neural Networks as Gaussian Processes. ICLR 2018
- Ma et al. 2019. Variational Implicit Processes. ICML 2019
- Tanno et al. 2019. Uncertainty Quantification in Deep Learning for Safer Neuroimage Enhancement. arXiv:1907.13418
- Foong et al. 2019. Pathologies of Factorised Gaussian and MC Dropout Posteriors in Bayesian Neural Networks. arXiv:1909.00719

References 3

- Hartikainen, J. and Sarkkä, S. (2010). Kalman filtering and smoothing solutions to temporal Gaussian process regression models. Proceedings of IEEE International Workshop on Machine Learning for Signal Processing (MLSP).
- Sarkkä, S., Solin, A., and Hartikainen, J. (2013). ¨ Spatio-temporal learning via infinite-dimensional Bayesian filtering and smoothing. IEEE Signal Processing Magazine, 30(4):51–61.
- Sarkkä, S. (2013). Bayesian Filtering and Smoothing. Cambridge University Press. Cambridge, UK.
- Sarkkä, S., and Solin, A. (2019). Applied Stochastic Differential Equations. Cambridge University Press. Cambridge, UK.
- Solin, A. (2016). Stochastic Differential Equation Methods for Spatio-Temporal Gaussian Process Regression. Doctoral dissertation, Aalto University.
- Durrande, N., Adam, V., Bordeaux, L., Eleftheriadis, E., Hensman, J. (2019). Banded matrix operators for Gaussian Markov models in the automatic differentiation era. International Conference on Artificial Intelligence and Statistics (AISTATS). PMLR 89:2780–2789.
- Nickisch, H., Solin, A., and Grigorievskiy, A. (2018). State space Gaussian processes with non-Gaussian likelihood. International Conference on Machine Learning (ICML). PMLR 80:3789–3798.
- Solin, A., Hensman, J., and Turner, R.E. (2018). Infinite-horizon Gaussian processes. Advances in Neural Information Processing Systems (NeurIPS), pages 3490–3499.
- Hou, Y., Kannala, J. and Solin, A. (2019). Multi-view stereo by temporal nonparametric fusion. International Conference on Computer Vision (ICCV).

More

PCA

GP-LVM

Multiple outputs

BNN

SVM

Unsupervised Learning

Invariances

Op3Mech
University of Antwerp

11-12-2019