# ORIE 4742 - Info Theory and Bayesian ML

Chapter 9: Gaussian Processes (Ch 6, Sec 2,4 of Bishop)

April 1, 2021

Sid Banerjee, ORIE, Cornell

## normal-normal model (Gaussian rv with unknown $\mu$)

- data $D = \{X_1, X_2, \ldots, X_n\} \in \mathbb{R}^n$
- model $\mathcal{M}$: $X_i$ i.i.d. from $\mathcal{N}(\mu, \tau)$, with unknown $\mu$, known $\tau = 1/\sigma^2$

### normal-normal model

- likelihood: $p(D|\mu) \propto \exp\left(-\tau \sum_{i=1}^n (x_i - \mu)^2/2\right)$
- prior: $\mu \sim \mathcal{N}(M_\mu, 1/\tau_\mu) \propto \exp\left(-\tau_\mu(\mu - m_\mu)^2/2\right)$
- posterior: let $\overline{x} = \frac{\sum_{i=1}^n x_i}{n}$,$\tau_D = n\tau + \tau_\mu$ and $m_D = \underbrace{\tau_D^{-1}(n\tau \cdot \overline{x} + \tau_\mu \cdot m_\mu)}_{\text{shrinkage estimator for MLE}}$

$$p(\mu|D) \sim \mathcal{N}\left(m_D, \tau_D^{-1}\right)$$

$\Rightarrow \mu_{|D} = m_D + (\tau_D)^{-1/2} Z_1$, $Z_1 \sim N(0,1)$, $Z_2 \sim N(0,1)$, $\perp\!\!\!\perp$

- posterior predictive distribution:

$$p(x|D) \sim \mathcal{N}\left(m_D, \tau^{-1} + \tau_D^{-1}\right)$$

$X = m_D + (\tau_D)^{-1/2} Z_1 + (\tau)^{-1/2} Z_2$

# Bayesian linear regression

$\circledast\, W^T = m_D^T + Z_1^T A_D$, $W^T\phi(x) = m_D^T\phi(x) + \underbrace{Z_1^T T_D^{-1/2}\phi(x)}_{=\phi(x)^T T_D^{-1/2} Z_1}$

- data $D = \{(t_1, x_1), (t_2, x_2), \dots, (t_N, X_N)\} \in \mathbb{R}^n$
- model $\mathcal{M}$: $t_i = \sum_{j=0}^{M-1} W_j \phi(x_i) + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, \beta^{-1})$

## Bayesian linear regression model

Recall

$t = \begin{pmatrix} t_1 \\ t_2 \\ t_N \end{pmatrix}$

$\Phi \equiv N \times M$ design matrix

$\Phi_{ij} = \phi_j(x_i)$

- likelihood: $p(D|W) \propto \exp\left(-\beta \sum_{i=1}^N (x_i - W^\intercal \phi(x_i))^2/2\right)$
- prior: $W \sim \mathcal{N}(0, \alpha^{-1} I)$ $\Rightarrow W = \alpha^{-1/2} Z_0$, $Z_0 \sim N(0, I)$
- posterior:

$$m_D = T_D^{-1} \beta \Phi^\intercal t,\ T_D = \beta \Phi^\intercal \Phi + \alpha I \Rightarrow p(W|D) \sim \mathcal{N}\left(m_D, T_D^{-1}\right)$$

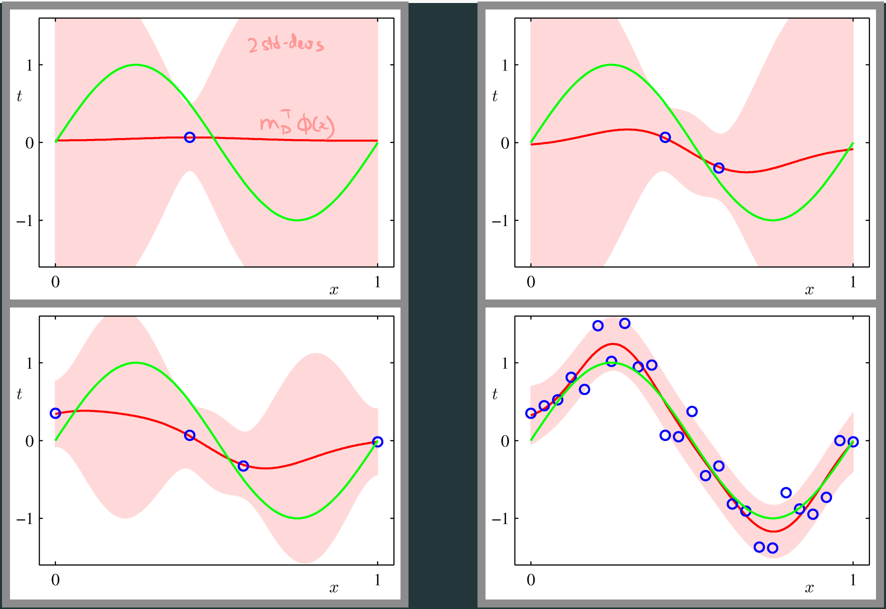
$A_D = T_D^{-1/2}$

$$\underset{M\times 1}{W} = \underset{M\times 1}{m_D} + \underset{M\times M}{A_D}\underset{M\times 1}{Z_1}, \quad \text{where} \quad A_D A_D^T = T_D^{-1}, Z_1 \sim \mathcal{N}(0, I)$$

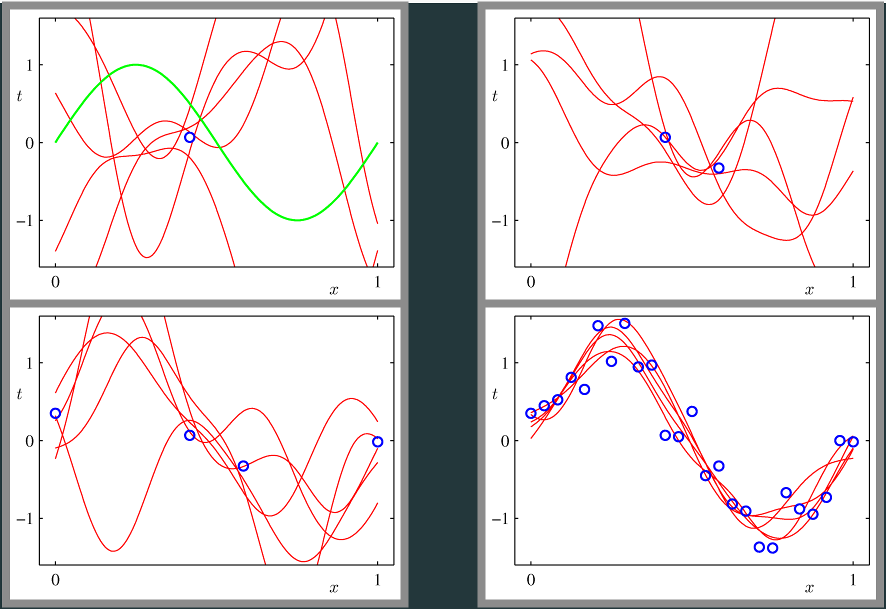
- posterior prediction: $p(t|D) \sim \mathcal{N}\left(m_D^\intercal \phi(x), \beta^{-1} + \phi(x)^\intercal T_D^{-1} \phi(x)\right)$

$$t(x|D) \overset{\circledast}{=} W^T \phi(x) + \beta^{-1/2} \underbrace{Z_2}_{N(0,1)} = m_D^T \phi(x) + \phi(x)^T T_D^{-1/2} \underbrace{Z_1}_{N(0,I)} + \beta^{-1/2} \underbrace{Z_2}_{N(0,1)}$$

# Bayesian linear regression: posterior prediction



# Bayesian linear regression: posterior sampling



## another way to write $y(x|D)$

$\left( y(x) = \sum_{j=0}^{M-1} w_j \phi_j(x) = W^T \phi(x) \right)$

$$t(x|D) = \underbrace{m_D^T \phi(x)}_{y(x|D)} + \phi(x)^T T_D^{-1/2} Z_1 + \beta^{-1/2} Z_2 , \quad Z_1 \sim N(0, I_M), \quad Z_2 \sim N(0,1)$$

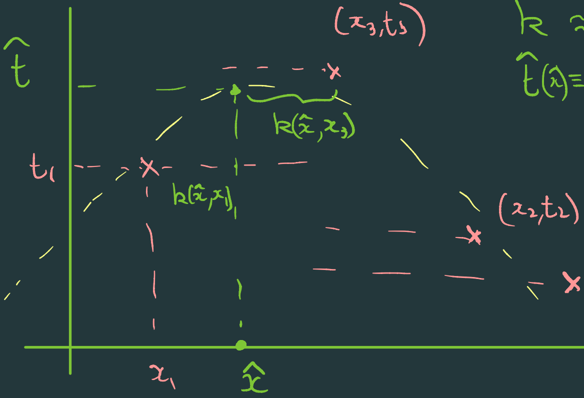
$$\phi(x) = \begin{pmatrix} \phi_1(x) \\ \phi_2(x) \\ \vdots \\ \phi_M(x) \end{pmatrix} \qquad \Phi^T = \begin{pmatrix} \phi_1(x_1) & \phi_1(x_2) \cdots \phi_1(x_N) \\ \phi_2(x_1) & \vdots \\ \vdots & \\ \phi_M(x_1) & \phi_M(x_2) \cdots \end{pmatrix}$$

- $$y(x|D) = \left( \beta T_D^{-1} \Phi^T t \right)^T \phi(x)$$

$$= \beta \left( \underset{1\times M}{\phi(x)^T} \underset{M\times M}{T_D^{-1}} \underset{M\times N}{\Phi^T} \right) \underset{N\times 1}{t}$$

$$= \underbrace{\sum_{i=1}^{N} \underbrace{k(x, x_i)}_{\text{'kernel'}} t_i}_{\text{'looks like a weighted sum of data'}}$$

where $k(x, x_i) = \left( \phi_1(x)\ \phi_2(x) \cdots \phi_M(x) \right) T_D^{-1} \begin{pmatrix} \phi_1(x_i) \\ \phi_2(x_i) \\ \vdots \\ \phi_M(x_i) \end{pmatrix}$




$k \approx$ 'distance fn'

$\hat{t}(\hat{x}) \equiv$ Sum of observed $t_i$s weighted by 'inverse distance' $k(\hat{x}, x_i)$

# the 'equivalent' kernel

- data $D = \{(t_1, x_1), (t_2, x_2), \ldots, (t_N, X_N)\} \in \mathbb{R}^n$
- model $\mathcal{M}$: $t_i = \sum_{j=0}^{M-1} W_j \phi(x_i) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \beta^{-1})$

- prior: $W \sim \mathcal{N}(0, \alpha^{-1} I)$
- posterior: let $m_D = \beta T_D^{-1} \Phi^\intercal t$ and $T_D = \beta \Phi^\intercal \Phi + \alpha I$, then

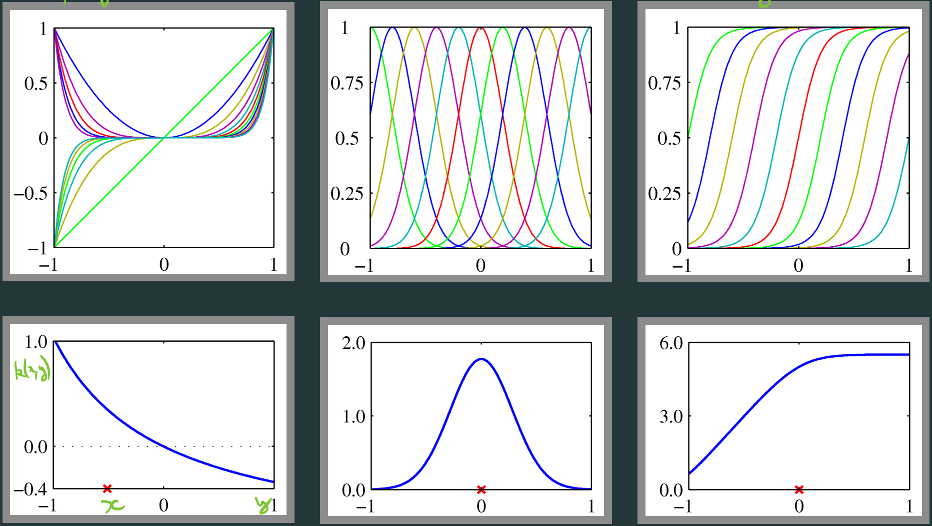
$$t(x|D) = W^T \phi(x) + \beta^{-1/2} Z_2 = m_D^T \phi(x) + \phi(x)^T T_D^{-1/2} Z_1 + \beta^{-1/2} Z_2$$

$$\mathrm{Var}(W|D) = \phi(x)^T T_D^{-1} \phi(x)$$

alternately, $y(x|D) = \underset{\text{sum over all data pts}}{\sum_{n=1}^N} \underbrace{k(x, x_n)}_{\text{wt of data pt } x_n \text{ as a fn of query pt } x} \overset{n^{th}\text{ observation}}{t_n}$, where $\underbrace{k(x, y) = \beta \phi(x)^T T_D^{-1} \phi(y)}_{\text{equivalent kernel for linear regression with basis fns } (\phi_1, \phi_2 \ldots \phi_M)}$

# basis functions and equivalent kernels

Polynomial Gaussian Sigmoid




For poly basis

$$\phi(x) = (1\ x\ x^2\ x^3)^T, \quad k(x,y) = \phi(x)^T\phi(y) = 1 + xy + x^2y^2 + x^3y^3$$

## what are kernel methods?

- generalized 'nearest-neighbor' methods
- given data $D = \{(x_1, t_1), \dots, (x_n, t_n)\}$, the resulting model is

$$y(x|D) = \sum_{i=n}^{N} k(x, x_n) t_n + \underbrace{\epsilon_D}_{\mathcal{N}(0,\ \text{Covariance matrix as a fn of } x)}$$

### properties of kernels

function $k(x, y)$ is a kernel of basis $\phi(x)$ if $k_\phi(x, y) = \phi(x)^\intercal \phi(y)$

$\phi(x)^T A^{-1} \phi(y)$

this is true if $k$ is

- symmetric $k(x, y) = k(y, x)$ (i.e., if $K$ is st $K_{xy} = k(x,y)$, then $K = K^T$)
- positive-definite $K = \{k(x_i, x_j)\} \succeq 0$ for all $\{x_i\}_{i=1}^{n}, n \in \mathbb{N}$

ie, $a^T K a \geq 0$ for any $a \in \mathbb{R}^n$

some special classes of kernels

- stationary kernel: $k(x, y) = \psi(x - y)$
- homogenous kernel: $k(x, y) = \psi(||x - y||)$ $\leftarrow$ 'inverse distance fn'

## Gaussian process

distribution over functions $G(x)$ such that: (sample pts $(x_1, x_2, \ldots x_n)$)

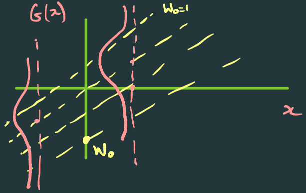
- any finite collection $(G(x_1), G(x_2), \ldots, G(x_n))$ is jointly Gaussian
- specified by mean $m(x) = \mathbb{E}[G(x)]$ and covariance $k(x, y) = \mathbb{E}[(G(x) - m(x))(G(y) - m(y))]$ (where $k$ is a kernel)

example: $y(x) = w^\intercal \phi(x)$, with $w \sim \mathcal{N}(0, \alpha^{-1} I)$

Eg 1 - $G(x) = \underbrace{W_0}_{\text{same for all } x} + x \ , \quad W_0 \sim N(0,1)$

$\mathbb{E}[G(x)] = x,$

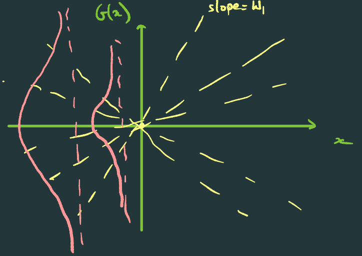
$\mathbb{E}[(G(x) - x)(G(y) - y)] = \mathbb{E}[W_0^2] = 1 = k(x,y)$



Eg 2 - $G(x) = W_1 x$

$m(x) = \mathbb{E}[G(x)] = 0$

$k(x,y) = \mathbb{E}[(W_1 x)(W_1 y)] = xy$



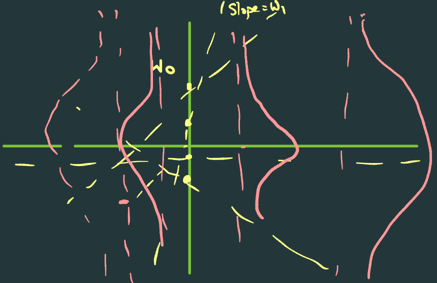

---

Eg - $G(x) = W_0 + W_1 x$

$m(x) = \mathbb{E}[W_0 + W_1 x] = 0$

$k(x,y) = \mathbb{E}[(W_0 + W_1 x)(W_0 + W_1 y)]$

$= 1 + xy$




---

- Ways of generating new kernels - Given kernels $k_1$, $k_2$, fn $\Phi$, the foll. are kernels

  * $c_1 k_1 + c_2 k_2$ * $e^{c k_1}$ * $k_2(\Phi(x), \Phi(y))$

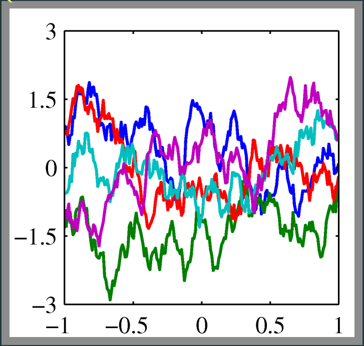
# Gaussian process examples

distribution over functions $G(x)$ with jointly Gaussian samples, mean $m(x) = \mathbb{E}[G(x)]$, covariance $k(x,y) = \mathbb{E}[(G(x) - m(x))(G(y) - m(y))]$

examples: $k(x,y) = exp(-\theta|x-y|)$, $k(x,y) = exp(-\theta(x-y)^2)$

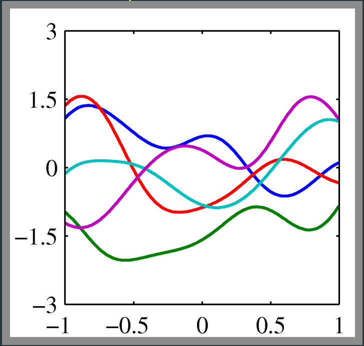
(Ornstein-Uhlenbeck) - OU kernel

radial basis fn (RBF) kernel




(related to Brownian motion)

- stationary, homogeneous




Lipschitz continuous fns