ORIE 4742 - Info Theory and Bayesian ML

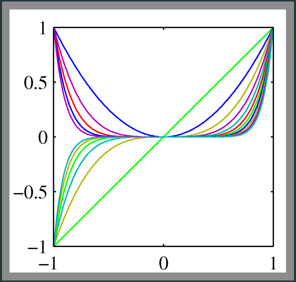
Chapter 8: Bayesian Regression

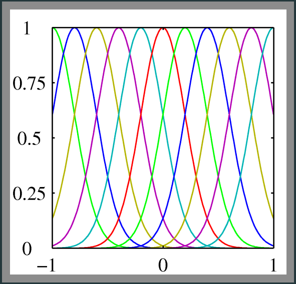
March 29, 2021

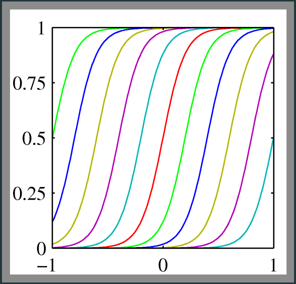
Sid Banerjee, ORIE, Cornell

what is linear regression?

basis functions







regression: the frequentist view

Bayesian linear regression

normal-normal model for unknown μ

- 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 μ, 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 $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $m_D = \frac{n\tau \cdot \overline{x} + \tau_\mu \cdot m_\mu}{n\tau + \tau_\mu}$ and $\tau_D = n\tau + \tau_\mu$

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

- posterior predictive distribution:

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

Bayesian linear regression

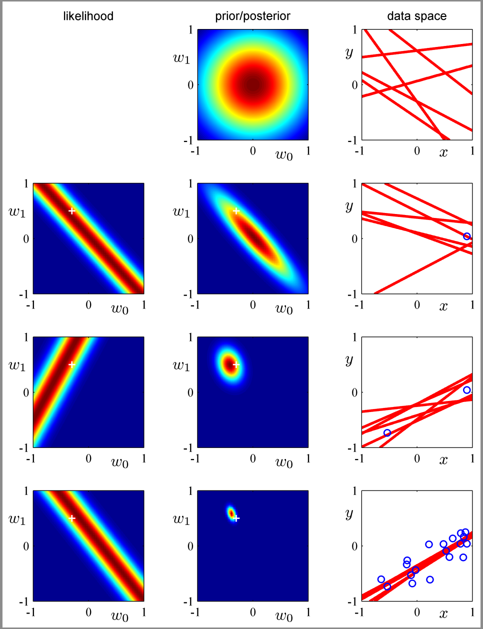
- data $D = \{(t_1, x_1), (t_2, x_2), \ldots, (t_N, X_N)\} \in \mathbb{R}^n$
- model $\mathcal{M}$: $t_i = \sum_{j=0}^{M-1} W_j \phi(x_i) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \beta^{-1})$

Bayesian linear regression model

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

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

Bayesian linear regression: example



ground truth: $f(x) = 0.1x - 0.3$

Bayesian linear regression

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

Bayesian linear regression model

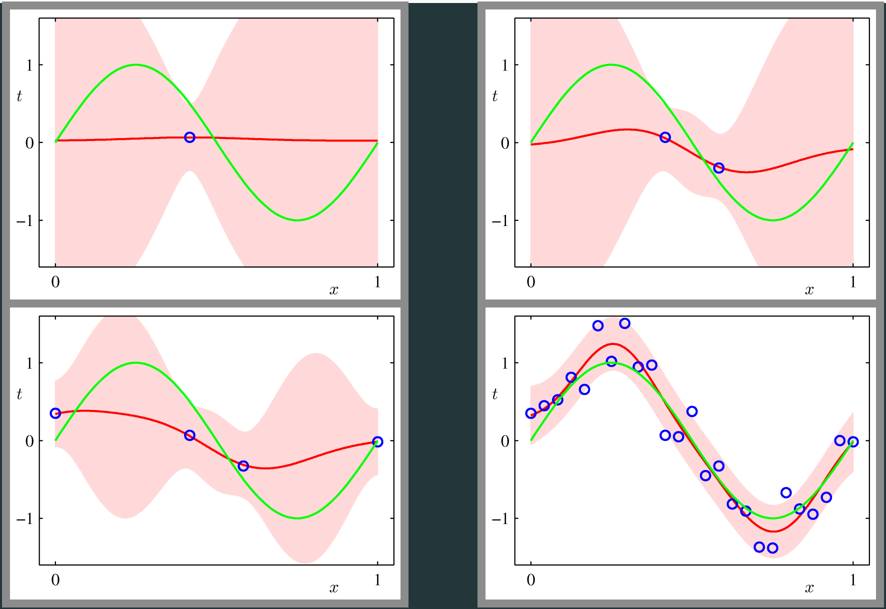
- likelihood: $p(D|W) \propto \exp\left(-\beta \sum_{i=1}^{N} (x_i - W^\intercal \phi(x_i))^2/2\right)$
- prior: $W \sim \mathcal{N}(0, \alpha^{-1} I)$
- posterior: let $m_D = T_D^{-1} \beta \Phi^\intercal t$ and $T_D = \beta \Phi^\intercal \Phi + \alpha I$

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

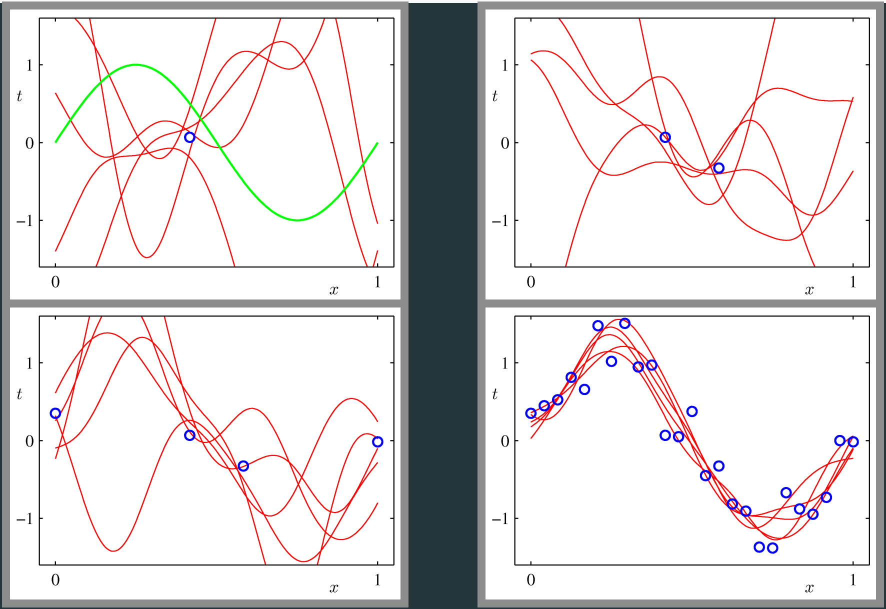
- posterior predictive distribution:

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

Bayesian linear regression: posterior prediction



Bayesian linear regression: posterior sampling



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 $\epsilon \sim \mathcal{N}(0, \beta^{-1})$

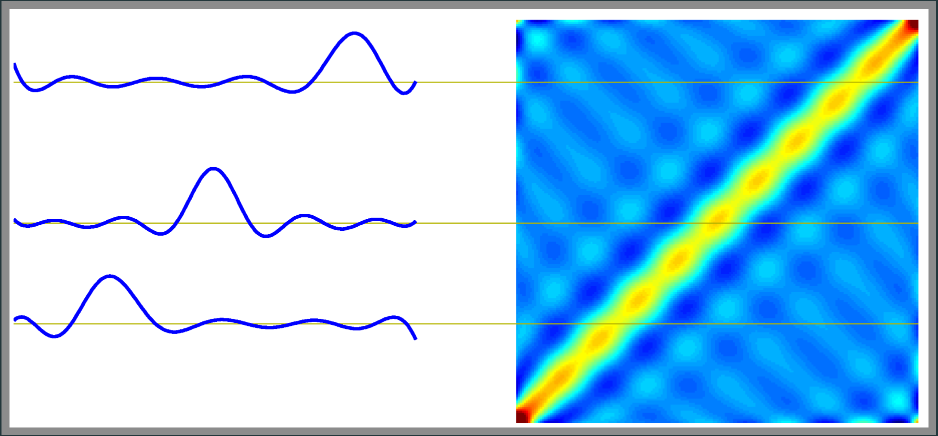
- prior: $W \sim \mathcal{N}(0, \alpha^{-1} I)$
- posterior: let $m_D = T_D^{-1} \beta \Phi^\intercal t$ and $T_D = \beta \Phi^\intercal \Phi + \alpha I$, then

$$t(x|D) = m_D^\intercal \phi(x) + \epsilon_D$$

where $\epsilon_D \sim \mathcal{N}(0, \beta^{-1} + \Phi^\intercal T_D^{-1} \Phi^\intercal)$

alternately, $y(x|D) = \sum_{n=1}^{N} k(x, x_n) t_n$, where $k(x, y) = \beta \phi(x)^T S_D \phi(y)$

the equivalent kernel: example



equivalent kernels