Bayesian linear regression

N data points, M basis fns

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

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: $W \sim \mathcal{N}(0, \underline{\alpha}^{-1} I)$ prior precision hyperparam
- posterior: let $m_D = \underbrace{T_D^{-1} \beta \Phi^\intercal t}_{(\Phi^T\Phi + \frac{\alpha}{\beta} I)^{-1} \Phi^T t}$ and $T_D = \underbrace{\beta \Phi^\intercal \Phi + \alpha I}_{\text{'precisions add'}}$

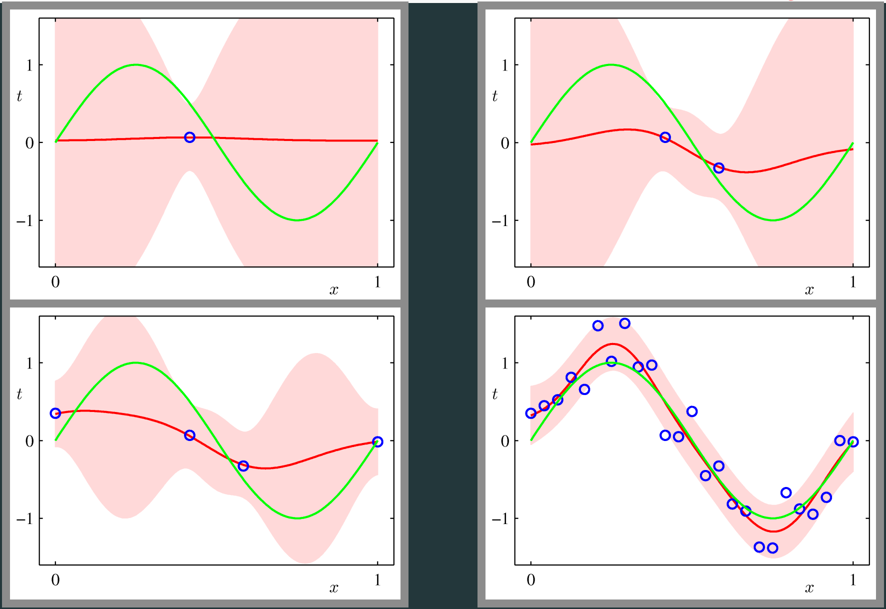
$$p(W|D) \sim \mathcal{N}\left(m_D, T_D^{-1}\right)$$

- posterior predictive distribution: i.e., what is $p(t|D)$ for new x

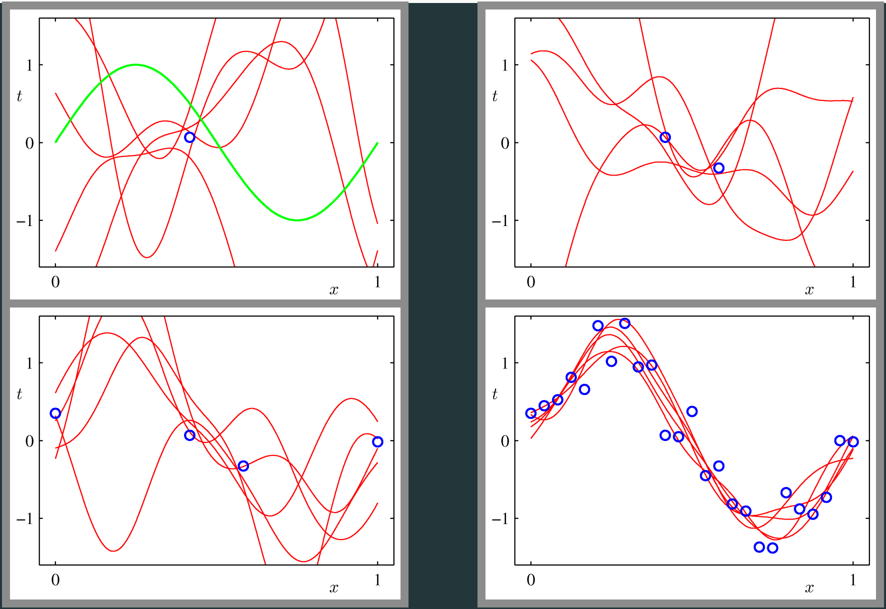
$$p(t|D) \sim \mathcal{N}\left(m_D^\intercal \phi(x), \underbrace{\beta^{-1} + \phi(x)^\intercal T_D^{-1} \phi(x)}_{\text{variances add up, depends on } x}\right)$$

Bayesian linear regression: posterior prediction

Gaussian basis fns
true model - $y(x) = \sin(2\pi x)$



Bayesian linear regression: posterior sampling



Last class - Ch 3 of Bishop (Section 3.3)

Today - Model selection (Sec 3.4), GP (Ch 6)

- Have uploaded Jupyter notebooks for Bayesian regression, GPs

---

- $t(x) \sim N\left(m_D^T \phi(x),\ \beta^{-1} + \phi(x)^T T_D^{-1} \phi(x)\right)$

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

$Z \sim N(0, T_D)$

$$\Rightarrow t(x) = m_D^T \phi(x) + \underbrace{Z^T \phi(x) + \varepsilon}_{N\left(0,\ \beta^{-1} + \phi(x)^T T_D^{-1} \phi(x)\right)}$$

Aside (Model Selection)

- Bayesian ML : $p(\theta \mid D, M) = \dfrac{p(\theta \mid M)\, \mathcal{L}(\theta \mid D, M)}{\underbrace{p(D \mid M)}_{\text{marginal likelihood}}}$

- Usually - prior $\equiv p(\theta \mid M) = \dfrac{1}{Z_{prior}} f(\theta)$

  $\left(\text{Eg. Beta-Bernoulli } \quad p(\theta \mid M) = \dfrac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{Z_\beta} \right)$

  posterior $\equiv p(\theta \mid D, M) = \dfrac{1}{Z_{post}} f(\theta)\, \mathcal{L}(\theta \mid D)$

$\Rightarrow \quad p(D \mid M) = \dfrac{Z_{prior}}{Z_{posterior}}$ (some formula based on conjugate prior family)

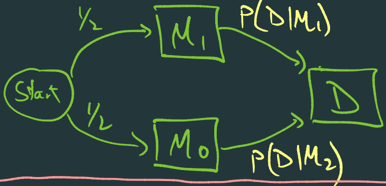
- Suppose we want to 'compare' models M_1, M_2

Eg - Given $X_i \in \{0,1\}$, are $X_i \sim \mathrm{Ber}(1/2)$ or not?

- Idea - $M_0 \equiv X_i \sim \mathrm{Ber}(\theta)$, $\theta = 1/2$

$M_1 \equiv X_i \sim \mathrm{Ber}(\theta)$, $\theta \sim \mathrm{Beta}(1,1)$

Which of these 'explains' D better



If we have a flat prior on models

Then 'most likely model given data'

$\equiv \arg\max_i \{P(D|M_i)\}$

(ie, maximum marginal likelihood)

- $P(D|M) \equiv$ 'evidence' of model M

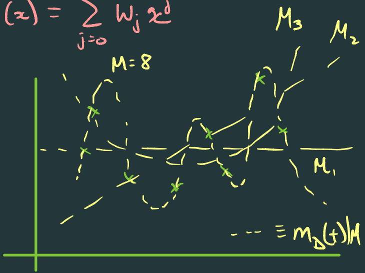
- For 2 models, $\dfrac{P(D|M_1)}{P(D|M_0)} \equiv$ 'Bayes factor'

Why is this a good idea? 'Bayesian Occam's Razor'

'Bayesian ML methods automatically choose correct model complexity'

Eg - Polynomial regression - $t(x) = y(x) + \varepsilon$, $y(x) = \sum_{j=0}^{M-1} W_j x^j$

$M_1 = \{M=0\}$, $M_2 = \{M=1\}$, ..., $M_k \equiv \{M = k-1\}$



Fact - If $M = N$, then $\exists$ a polynomial which goes through every data point

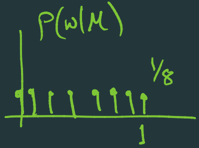
(Lagrange poly) - $y(x) = \sum_{i=1}^{n} t_i \dfrac{\prod_{j \neq i} (x - x_j)}{\prod_{j \neq i} (x_i - x_j)}$

Q: Does it make sense to have $M \geq N$?

- Suppose we know $w_j \in \{0, 1/k, 2/k, \dots, 1\}$, $k = 7$

$$y(x) = \sum_{j=0}^{M-1} w_j x^j$$

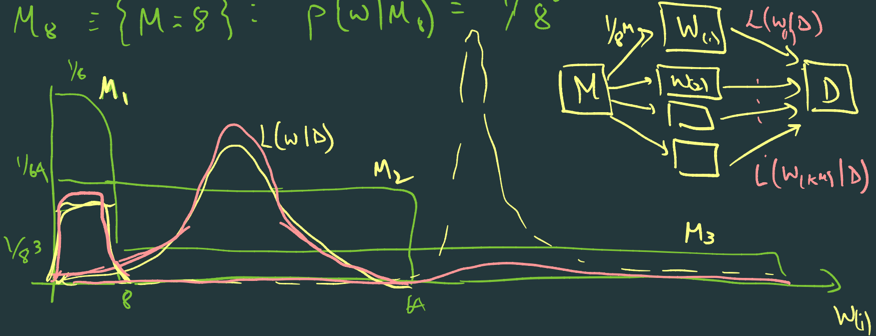
- $M_1 \equiv \{M = 1\}$



$M_2 \equiv \{M = 2\}$: $p((w_1, w_2)|M) = 1/64$

⋮

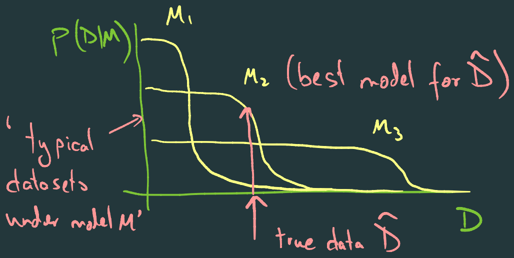
$M_8 \equiv \{M = 8\}$: $p(w|M_8) = 1/8^8$



- Another view of Bayesian Occam's Razor

  - Probabilistic model ≡ 'distribution over datasets'

$$M \equiv p(D|M) = \text{'prob of seeing } D \text{ under } M\text{'}$$



Bayesian model selection chooses the 'simplest model' (ie, smallest typical set of D) such that true data $\hat{D}$ lies in the typical set

'Empirical Bayes / Evidence approximation' heuristic

- Suppose model has hyperparams
  (Eg- polynomial regression - M, α, β)
- <u>Idea</u> - select M, α, β s.t they maximize

$$p(D \mid M, \alpha, \beta)$$

  - For Bayesian regression - can optimize over α, β (given M) is closed form - $\beta^*(M,D), \alpha^*(M,D)$