Proba. sup. learning & MLE $\Leftrightarrow$ ERM

ERM: $R(\theta)$

RRM: $R(\theta) = \text{PredLoss}(\theta) + \underbrace{\text{Regul}(\theta)}_{\lambda \|\theta\|^2}$

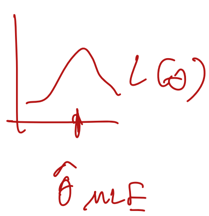
Prob. interp. as Bayesian learning

MLE vs. (max)-Bayesian estim.

z, θ

data, param

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \overbrace{p(z|\theta)}^{Lik(\theta)}$$



Bayesian estim.

prior distrib $p(\theta|\lambda)$ ← fixed hyperparam.

← prior knowledge

$\lambda \to \theta \to z$

θ is RV!

Joint distrib: $p(z, \theta|\lambda) = p(z|\theta, \cancel{\lambda})\, p(\theta|\lambda)$

Posterior distrib. of θ

$$p(\theta|z, \lambda) = \frac{p(z|\theta)\, p(\theta|\lambda)}{p(z|\lambda)}$$

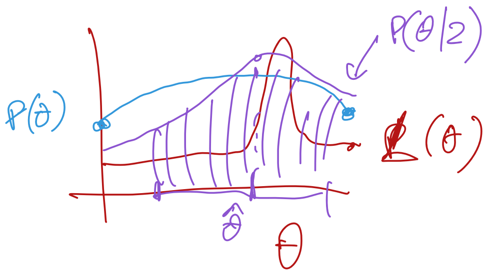
$$P(\theta \mid z, \lambda) = \frac{1}{P(z|\lambda)} \, \underline{P(z|\theta)} \; \underline{P(\theta|\lambda)}$$

$$\propto P(z|\theta) \, P(\theta|\lambda)$$ "Unnorm. Posterior of θ"

<u>MAP</u> = Maximum a posteriori

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}} \; P(z|\theta) \, P(\theta|\lambda)$$

if $P(\theta|\lambda) = \text{const.} \Rightarrow \hat{\theta}_{MAP} = \hat{\theta}_{MLE}$



$$\mathcal{L}(\theta) = \underline{P(z|\theta)}$$

Prob. posterior for θ: More stuff

- MAP]
- Mean
- 95% CI

L2-Reg. LinReg as Generative Model

λ, σ^2: fixed $\qquad \vec{\theta}\ \vec{y}$: R.V. $\qquad \vec{X}$: fixed

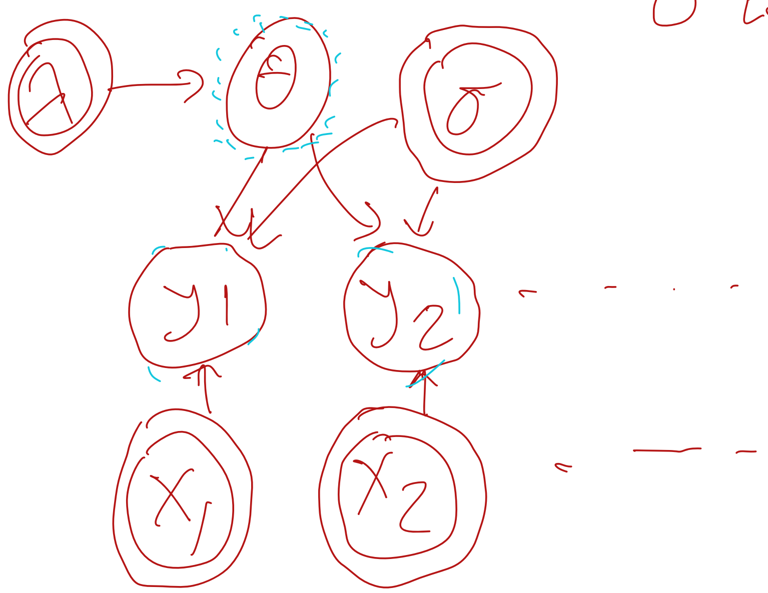
① For $j = 1..D$: $P(\theta_j | \lambda) = N(0, \frac{1}{\lambda}) \iff \theta_j \sim N(0, \frac{1}{\lambda})$

② For $i = 1..N$: $P(y_i | x_i, \theta, \sigma^2) = N(\theta^T x_i, \sigma^2)$

Joint Prob. Distrib

$$P(\theta, Y | X, \sigma^2, \lambda)$$
$$= P(Y | X, \theta, \sigma^2, \cancel{\lambda})\ P(\theta | \lambda, \cancel{X}, \cancel{\sigma^2})$$

Dir. Graphical Model



◎ Observed/Fixed/Conditioned on

○ Latent

$$\hat{\theta}_{MAP} = \arg\max_{\theta} \underbrace{P(\theta \mid Y, X, \sigma^2, \lambda)}$$

$$P(Y \mid X, \sigma^2, \theta) \quad P(\theta \mid \lambda)$$

$$\log P(Y \mid X, \sigma^2, \theta) + \boxed{\log P(\theta \mid \lambda)} \qquad N\left(0, \sigma^2 = \frac{1}{\lambda}\right)$$

$\boxed{\text{Prior}}$

$$\log P(\theta \mid \lambda) = \sum_{j=1}^{D} \log\left[\frac{1}{\sqrt{2\pi \frac{1}{\lambda}}} \exp\left[\frac{-1}{2\left(\frac{1}{\lambda}\right)} \theta_j^2\right]\right]$$

$$= \sum_j \left[\underbrace{\left(-\log\sqrt{2\pi}\right) + \log\sqrt{\lambda}}_{\text{constant !!}} - \frac{\lambda}{2}\theta_j^2\right]$$

$$= C + \frac{\lambda}{2}\sum_j \theta_j^2$$

$\underline{\text{Like}}$ $P(Y|X,\theta,\sigma^2) = \prod_i \underbrace{p(y_i|x_i,\theta,\sigma^2)}_{N(\theta^T x_i,\ \sigma^2)}$

$\log(Y|X,\theta,\sigma^2) =$

$\sum_i \log\left[\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{-1}{2\sigma^2}(y_i-\theta^T x_i)^2\right)\right]$

$= \sum_i \underbrace{\log\left[\frac{1}{\sqrt{2\pi\sigma^2}}\right]}_{\text{const wrt }\theta} - \frac{1}{2\sigma^2}(y_i-\theta^T x_i)^2$

$= -\frac{1}{2\sigma^2}\sum_i (y_i-\theta^T x_i)^2 \quad + \text{const}$

$J(\theta) =$

$\log P(Y|X,\theta,\sigma^2) + \log P(\theta|\lambda)$

$= -\frac{1}{2\sigma^2}\sum_i (y_i-\theta^T x_i)^2 - \frac{\lambda}{2}\sum_j \theta_j^2 \quad + \text{const}$

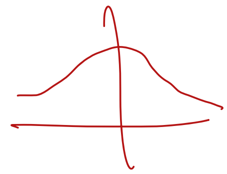
$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}}\ J(\theta)$

Why normal?

- Common
- convenient. L2 norm is CVX

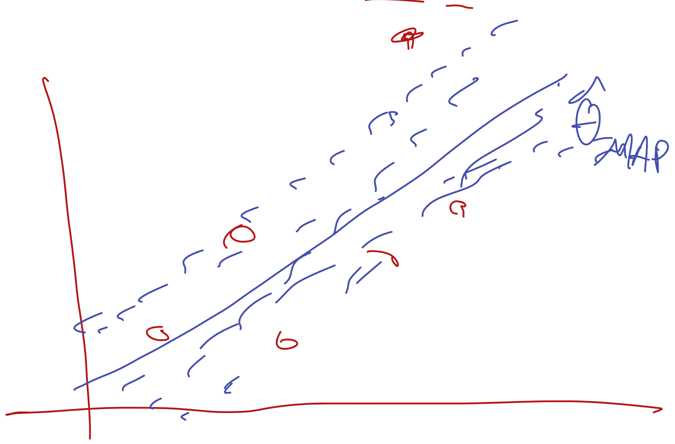
Why param. prob. models?

↳ Makes assumps clear



or: be assump.-free ??

$$y_i \sim N\left(\sum_j \alpha_j K(x_j, x_i), \ \sigma^2\right)$$



$P(\theta \mid Y, X \ldots)$