SUPERVISED LEARNING

$\underline{x}, \underline{y} : (x_i, y_i), i = 1, .. N$ observations (input, output)

Assume $p(y|x) = p(y|x, \vartheta)$ parametric

MAXIMUM LIKELIHOOD:

$$\vartheta_{ML} = \operatorname{argmax} p(\underline{y} | \underline{x}, \vartheta)$$

LINEAR REGRESSION

Assume $p(y|x, \vartheta) = \mathcal{N}(y | f(x, w), \beta^{-1})$, i.e. $y = f(x, w) + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \beta^{-1})$ $\quad \vartheta = (w, \beta)$

and $f(x, w) = w_0 \phi_0(x) + .. + w_{M-1} \phi_{M-1}(x)$, $\quad \Phi = (\phi_0, .., \phi_{M-1})$ basis func. $\begin{cases} \text{monomials} \\ \text{Gaussian RBF} \\ \text{sigmoid} \\ \vdots \end{cases}$

As $\log p(y|x, \vartheta) \propto -E_D(w) = -\frac{1}{2} \sum_{n=1}^{N} (y_n - w^T \phi(x_n))^2$, and

$\nabla_w E_D(w) = \sum_{n=1}^{N} (y_n - w^T \phi(x_n)) \phi^T(x_n) = 0 \Rightarrow w_{ML} = (\Phi^T \Phi)^{-1} \Phi^T \underline{y}$, $\Phi_{ij} = \phi_j(x_i)$

REGULARIZATION

solve $\operatorname{argmin}_w E_D(w) + \lambda E_w(w)$

where $E_w(w) = \begin{cases} \frac{1}{2} \|w\|_2^2 & \text{(Ridge)} \\ \frac{1}{2} \|w\|_1 & \text{(Lasso)} \end{cases}$

- REGULARIZATION IS BIAS

- PRIOR BELIEFS ≡ BIASING

$$p(w \mid \alpha) = \mathcal{N}(w \mid 0, \alpha I)$$

force params to be small

$$\underbrace{p(w \mid \underline{x}, \underline{y}, \alpha, \beta)}_{\text{POSTERIOR}} = \frac{\overbrace{p(\underline{y} \mid \underline{x}, w, \alpha, \beta)}^{\text{LIKELIHOOD}} \overbrace{p(w \mid \alpha)}^{\text{PRIOR}}}{\underbrace{p(\underline{y} \mid \underline{x}, \alpha, \beta)}_{\text{MARGINAL LIKELIHOOD}}}$$

$$\log p(w \mid \underline{x}, \underline{y}, \alpha, \beta) = -\frac{\beta}{2} \sum_{j=1}^{N} \left(y_j - \underbrace{w^T \phi(x_j)}_{\text{quadratic in } w} \right)^2 - \alpha \underbrace{w^T w}_{\text{quadratic in } w} + \text{const}$$

$$p(w \mid \underline{x}, \underline{y}, \alpha, \beta) = \mathcal{N}(w \mid m_N, S_N)$$

$$m_N = \beta S_N \underline{\Phi}^T \underline{y}$$

$$S_N^{-1} = \alpha I + \beta \underline{\Phi}^T \underline{\Phi}$$

IF we use a general prior

$$p(w \mid m_0, S_0)$$

$$\downarrow$$

$$p(w \mid \underline{x}, \underline{y}, \alpha, \beta) = \mathcal{N}(w \mid m_N, S_N)$$

$$m_N = S_N \left[S_0^{-1} m_0 + \beta \underline{\Phi}^T \underline{y} \right]$$

$$S_N^{-1} = S_0^{-1} + \beta \underline{\Phi}^T \underline{\Phi}$$

LINEAR

BAYESIAN REGRESSION

$\downarrow$

GAUSSIAN PRIOR

$\downarrow$

GAUSSIAN POSTERIOR

$\updownarrow$

ANALYTICALLY COMPUTABLE!

PREDICTIVE DISTRIBUTION

TAKE A NEW x.

$$p(y \mid x, \underline{x}, \underline{y}, \alpha, \beta) = \int p(y \mid x, w, \alpha, \beta)\, p(w \mid \underline{x}, \underline{y}, \alpha, \beta)\, dw$$

$p(y \mid x, w, \alpha, \beta)$: GAUSSIAN, LINEAR ON w

$p(w \mid \underline{x}, \underline{y}, \alpha, \beta)$: GAUSSIAN

GAUSSIAN

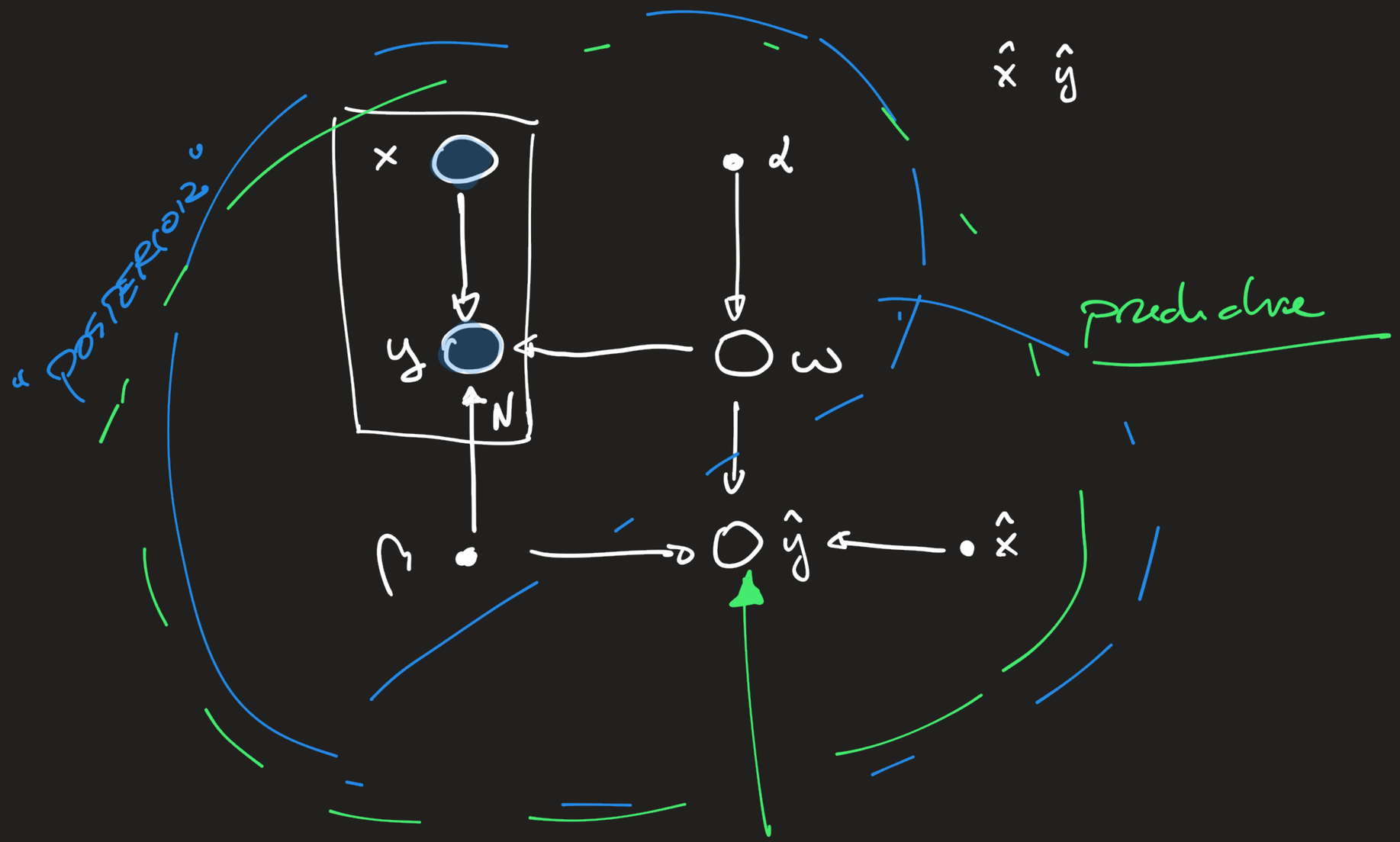
GAUSSIAN

$$= \mathcal{N}(y \mid m_N^T \phi(x), \sigma_N^2(x))$$

$$\sigma_N^2(x) = \frac{1}{\beta} + \phi(x)^T S_N \phi(x)$$

$$\sigma_{N+1}^2(x) \geq \sigma_N^2(x), \quad \sigma_N^2(x) \to \frac{1}{\beta}, \quad N \to \infty$$

PGM of BLR !



MARGINAL LIKELIHOOD

$$p(\underline{y} \mid \underline{x}, \alpha, \mathcal{M}) = \int p(\underline{y} \mid \underline{x}, \omega, \alpha, \mathcal{M})\, p(\omega \mid \alpha)\, d\omega$$

MARGINAL LIKELIHOOD

as the posterior $p(\omega \mid \underline{y}, \underline{x}, \alpha, \mathcal{M}) = \dfrac{p(\underline{y} \mid \underline{x}, \omega, \alpha, \mathcal{M})\, p(\omega \mid \alpha)}{p(\underline{y} \mid \underline{x}, \alpha, \mathcal{M})}$

The Bayesian way for α and β $\Rightarrow$ $p(\alpha, \mathcal{M})$ HYPERPRIOR

$$p(\alpha, \mathcal{M} \mid \underline{y}, \underline{x}) \propto p(\underline{y} \mid \underline{x}, \alpha, \mathcal{M})\, p(\alpha, \mathcal{M})$$

ⓐ UNINFORMATIVE PRIOR
‖
$p(\alpha, \mathcal{M})$ IS CONSTANT

ⓑ POSTERIOR IS SHARPLY PEAKED AROUND MAP $\Rightarrow$

$$p(\alpha, \mathcal{M} \mid \underline{y}, \underline{x}) \approx \delta_{MAP(\alpha, \mathcal{M})}$$

ⓐ + ⓑ $\Rightarrow$ we can FIX $\alpha, \mathcal{M}$ TO M.L. $\Rightarrow$ WE CAN MAXIMIZE THE MARGINAL LIKELIHOOD.

$$\log p(\underline{y} \mid \underline{x}, \alpha, \beta) = \frac{M}{2}\log\alpha + \frac{N}{2}\log\beta - E(m_N) - \frac{1}{2}\log|S_N^{-1}| - \frac{N}{2}\log 2\pi$$

$$E(m_N) = \frac{\beta}{2}\|y - \Phi m_N\|^2 + \frac{\alpha}{2} m_N^T m_N$$

$$\left[m_N = \beta S_N \Phi^T \underline{y} \;\; ; \;\; S_N^{-1} = \alpha I + \beta \Phi^T \Phi \right]$$

FIX. POINT METHOD ($\underline{\nabla \log p(\underline{y} \mid \underline{x}, \alpha, \beta) = 0}$ and derive F.P. equations)

$$\begin{cases} \alpha = g_\alpha(\alpha, \beta) \leftarrow \\ \beta = g_\beta(\alpha, \beta) \leftarrow \end{cases}$$

ALGORITHM : Fix α_0, β_0, then compute

$$\alpha_{n+1} = g_\alpha(\alpha_n, \beta_n)$$
$$\beta_{n+1} = g_\beta(\alpha_n, \beta_n)$$

iterate until convergence $\|\alpha_{n+1} - \alpha_n\| + \|\beta_{n+1} - \beta_n\| < \varepsilon$

$$\log p(\underline{y} \mid \underline{x}, \alpha, \beta) = \frac{M}{2}\log\alpha + \frac{N}{2}\log\beta - E(m_N) - \frac{1}{2}\log|S_N^{-1}| - \frac{N}{2}\log 2\pi$$

$$E(m_N) = \frac{\beta}{2}\|\underline{y} - \underline{\Phi}\cdot m_N\|^2 + \frac{\alpha}{2} m_N^T m_N$$

$$S_N^{-1} = \alpha I + \beta \underline{\Phi}^T \underline{\Phi}$$

$\beta \underline{\Phi}^T \underline{\Phi}$ and compute eigenvalues $\lambda_i > 0$ do not depend on α

$$|S_N^{-1}| = \prod_{i=0}^{M-1} (\alpha + \lambda_i)$$

$$\frac{\partial}{\partial\alpha}\log|S_N^{-1}| = \frac{\partial}{\partial\alpha}\sum \log(\alpha+\lambda_i) = \sum_{i=0}^{M-1}\frac{1}{\alpha+\lambda_i}$$

$$\frac{\partial}{\partial\beta}\lambda_i = \frac{\lambda_i}{\beta}$$

$$\Rightarrow \quad \alpha = \frac{\gamma}{m_N^T m_N} = g_\alpha(\alpha,\beta) \quad \text{, where } \gamma = \sum_{i=0}^{M-1}\frac{\lambda_i}{\alpha+\lambda_i}$$

$$\frac{1}{\beta} = \frac{1}{N-\gamma}\sum_{n=1}^{N}\left(y_n - m_N^T \phi(x_n)\right)^2 = \frac{1}{g_\beta(\alpha,\beta)}$$

$$\gamma = \sum_{i=0}^{M-1} \frac{\lambda_i}{\alpha + \lambda_i}$$

EFFECTIVE # OF PARAMETERS

λ_i = eigenvalues of $\Phi^T \Phi$

λ_i gives info on w_{ML}^i

SMALL λ_i => LARGE UNCERTAINTY ON w_{ML}^i ($\lambda_i \ll \alpha \Rightarrow \frac{\lambda_i}{\lambda_i + \alpha} \approx 0$)

LARGE λ_i => SMALL UNCERTAINTY ($\lambda_i \gg \alpha \Rightarrow \frac{\lambda_i}{\lambda_i + \alpha} \approx 1$)

for $N \gg M$ $\gamma \approx M$

(MARGINAL LIKELIHOOD = EVIDENCE)

BAYESIAN MODEL COMPARISON

M_1 , M_2 ARE DIFFERENT MODELS

WHICH ONE IS THE BEST TO EXPLAIN DATA $D = \{x, y\}$

$p(M_j)$ PRIOR ON THE MODELS

$p(M_j|D)$ = POSTERIOR

$$= \frac{p(D|M_j)\cdot p(M_j)}{\sum_j p(D|M_j)\, p(M_j)}$$

MARGINAL LIKELIHOOD

1) MODEL AVERAGING.

$$\underbrace{p(y \mid x, D)}_{\text{predictive}} = \sum_{0} \underbrace{p(y \mid x, D, M_j)}_{} \cdot p(M_j \mid D)$$

2) CHOOSE THE BEST MODEL

$$\frac{p(D \mid M_1)}{p(D \mid M_2)} \doteq \text{BAYES FACTOR}$$

choose M_j with largest Bayes Factor

because $\int p(D \mid M_2) \log \frac{p(D \mid M_1)}{p(D \mid M_2)} \, dD > 0$ if M_1 is the true model

EQUIVALENT KERNEL

$$p(y|x, \underline{x}, \underline{y}, \alpha, \beta) = \mathcal{N}\left(y \mid \underbrace{m_N^T \phi(x)}, \sigma_N^2(x)\right)$$

$$y(x, m_N) = m_N^T \phi(x) = \beta \phi(x)^T S_N \underline{\Phi}^T \underline{y} = \sum_{n=1}^{N} \beta \phi^T(x) S_N \phi(x_n) y_n$$

$$= \sum_{n=1}^{N} k(x, x_n) y_n$$

$$k(x, x') = \beta \phi^T(x) S_N \phi(x') \quad \text{on}$$

equivalent kernel

$$\text{cov}[y(x), y(x')] = \phi^T(x) S_N \phi(x') = \underline{\beta^{-1} k(x, x')}$$

$$k(x, x') = \psi^T(x) \psi(x') \qquad \psi(x) = \beta^{1/2} S_N^{1/2} \phi(x) \quad \text{on}$$