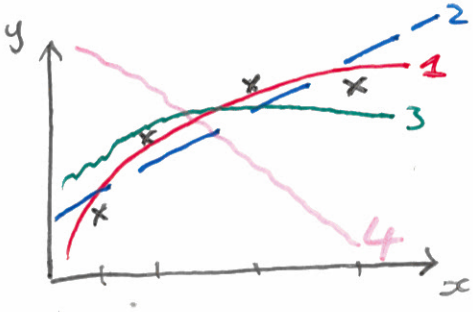
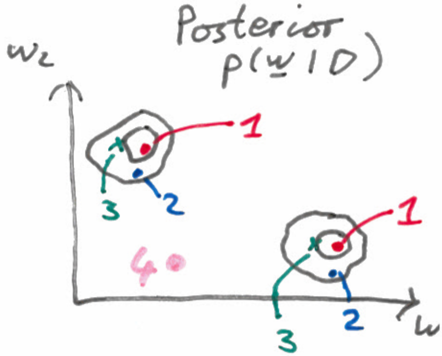
Variational Methods

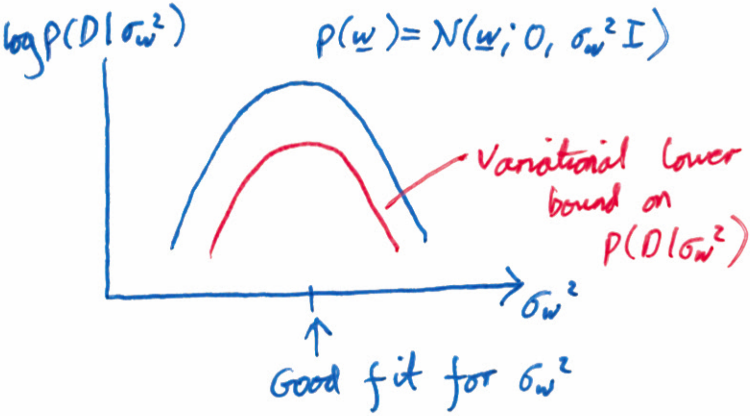




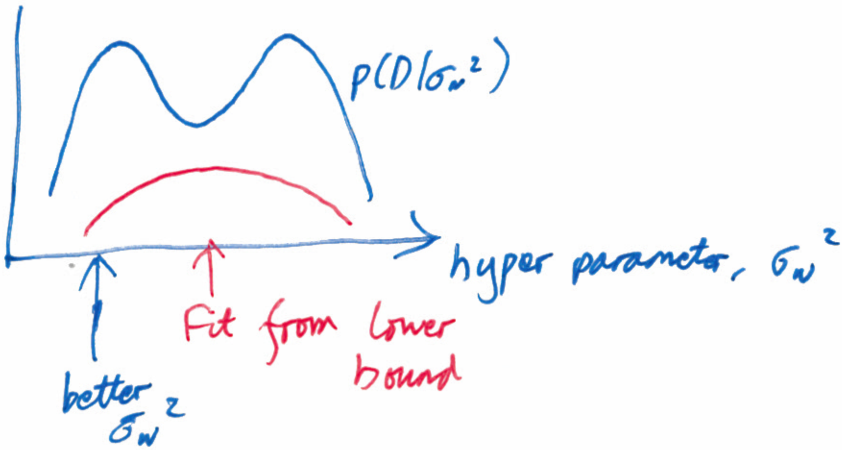
Approx. posterior with $q(\underline{w};\alpha) \underset{e.g.}{=} \mathcal{N}(\underline{w}; \underbrace{\underline{m}, V}_{\alpha})$

- For prediction 1 mode might be ok.

- We use q to approx. $p(D|M)$
  The number itself will be wrong by a large factor.



Bad case :



$$D_{KL}\left(q(\underline{w};\alpha) \,\|\, p(\underline{w}|D) \right) = \mathbb{E}_q\left[\log \frac{q(\underline{w};\alpha)}{p(\underline{w}|D)} \right]$$

$$= \int q(\underline{w};\alpha) \log \frac{q(\underline{w};\alpha)}{p(\underline{w}|D)} \, d\underline{w} \geqslant 0$$

$$= \underbrace{- \mathbb{E}_q\left[\log p(\underline{w}|D) \right]}_{\text{Minimize this term with } q(\underline{w};\alpha) = N(\underline{w};\, \underline{w}^*,\, \underline{0})} + \overbrace{\mathbb{E}_q\left[\log q(\underline{w};\alpha) \right]}^{-\text{Entropy of } q}$$

Substitute $p(\underline{w}|D) = \dfrac{p(D|\underline{w})\,p(\underline{w})}{p(D)}$

$$D_{KL} = \underbrace{\mathbb{E}_q[\log q(\underline{w};\alpha)] - \mathbb{E}_q[\log p(\underline{w})] - \mathbb{E}_q[\log p(D|w)]}_{= J(\alpha) = J(\underline{m}, V)} + \underbrace{\mathbb{E}_q[\log p(D)]}_{\log p(D)}$$

Fit $q(\underline{w}) = N(\underline{w}; m, V)$ by minimizing J

Also want to fit model hyperparameters.

We'd like to maximize $\log p(D|\text{hypers})$

$$\log p(D|\text{hypers}) \geqslant -J$$

(Because $D_{KL} \geqslant 0$)

→ Jointly minimize J

wrt $\underbrace{\{\underline{m}, V\}}_{\text{Variational params}}$ and $\underbrace{\text{model hypers}}_{\text{eg } \sigma_w^2}$

Optimizing $D_{KL}(q(w) \| p(\underline{w}|D))$

Gradient-based optimization

Particularly stochastic gradient descent. (S.G.D.)

Not on $\underline{w}$, weights of logistic regression or a n.n.

On beliefs about $\underline{w}$, $q(\underline{w}) = N(\underline{w}; \underbrace{\underline{m}, V}_{\text{Opt. these.}})$

Also optimize hyper-parameters.

Unconstrained optimization (Trick #1)

If we optimize σ_w^2 S.G.D. might make it -ve.

Also V has to be positive definite

Instead we optimize $\log \sigma_w$

Also transform V:

L is Cholesky decomp of V.

$V = LL^T$, L lower triangular matrix with +ve diagonal

We create another matrix

$$\tilde{L}_{ij} = \begin{cases} L_{ij} & i \neq j \\ \log L_{ii} & i = j \end{cases}$$

$$\text{Optimize } \hat{L} \xrightarrow[\text{diagonal}]{\exp} L \longrightarrow V = LL^T \to \text{est. cost}$$

backprop.

gradients

"Entropy Terms"

We can evaluate $\mathbb{E}_{\mathcal{N}(\underline{w};\, \underline{m}, V)}[\log \mathcal{N}(\underline{w};\, \mu, \Sigma)]$

For any $\underline{m}, V, \mu, \Sigma \ldots$

Likelihood Term

$$\mathbb{E}_q[\log p(D \mid \underline{w})]$$

$$= \mathbb{E}_q\left[\sum_{n=1}^{N} \log p(y^{(n)} \mid \underline{x}^{(n)}, \underline{w})\right]$$

Could do by numerical integration.

Stochastic estimation

$$\mathbb{E}_{N(\underline{w};\underline{m},V)}[f(\underline{w})]$$

$$= \mathbb{E}_{N(\underline{\nu};0,I)}[f(\underline{m}+L\underline{\nu})]$$

$$\left[\begin{array}{l}\text{Sample } \underline{w}, \quad \text{by } \underline{\nu} \sim N(0,I) \\ \qquad\qquad \underline{w} = \underline{m} + L\underline{\nu}\end{array}\right]$$

$$\approx \quad f(\underline{m}+L\underline{\nu}), \quad \underline{\nu} \sim N(0,I)$$

Monte Carlo estimate. Unbiased est.

$$\nabla_{\underline{m}}\, \mathbb{E}_{N(\underline{\nu};0,I)}[f(\underline{m}+L\underline{\nu})]$$

$$\approx \nabla_{\underline{m}}\, f(\underline{m}+L\underline{\nu}), \quad \underline{\nu} \sim N(0,I)$$

$$\nabla_{\tilde{L}}\, \mathbb{E}_{N(\nu;0,I)}[f(\underline{m}+L\nu)]$$

. . . chain rule only $\nabla_{\underline{w}} f$