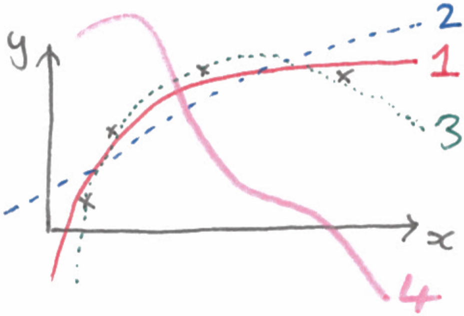
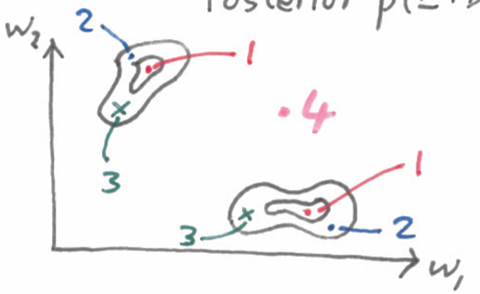
Approximate Bayesian Inference

E.g. neural net regression:



Posterior $p(\underline{w} | D)$

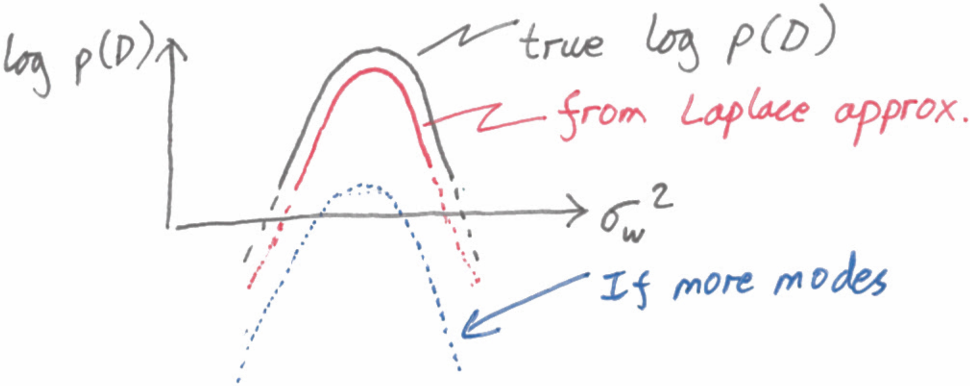


Laplace: $p(\underline{w} | D)$ "$\approx$" $\mathcal{N}(\underline{w}; \underline{w}^*, H^{-1})$

Prediction: fitting one mode might be ok.

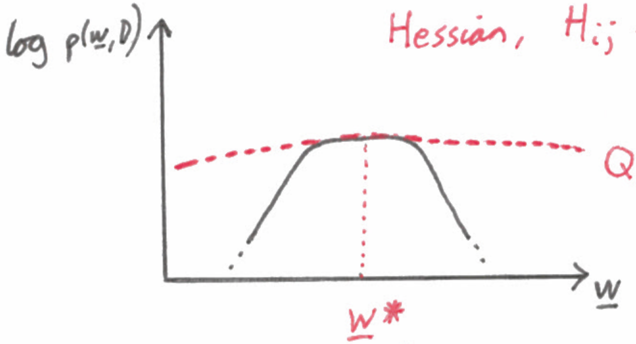
If 2 Gaussian modes, $p(D)$ half true answer

Setting hyper-parameters might work:



Laplace Approximation

Hessian, $H_{ij} = \dfrac{\partial^2 -\log p(\underline{w}, D)}{\partial w_i \, \partial w_j}$



Quadratic fit, tiny curvature

Found with optimizer

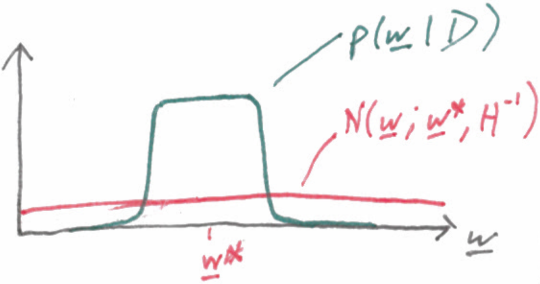
Posterior

$$p(\underline{w} \mid D) \approx \mathcal{N}(\underline{w}; \underline{w}^*, H^{-1})$$

Marginal Likelihood

$$p(D) = \frac{p(\underline{w}^*, D)}{p(\underline{w}^* \mid D)} \approx \frac{p(\underline{w}^*, D)}{\mathcal{N}(\underline{w}^*; \underline{w}^*, H^{-1})}$$

Areas under both curves = 1



Quiz

Is p(D) approx.

A) Too big

B) Too small

C) ≈ Correct

Z) ???

Variational Methods

Another way to fit approx $p(\underline{w}|D)$:

$$p(\underline{w}|D) \simeq q(\underline{w};\alpha)$$

For us $q(\underline{w};\alpha) = N(\underline{w};\underline{m},V)$

Variational parameters $= \{m, V\}$

Have optimization problem

Fit α, need cost function

Measure discrepancy between $p(\underline{w}|D)$ and $q(\underline{w})$

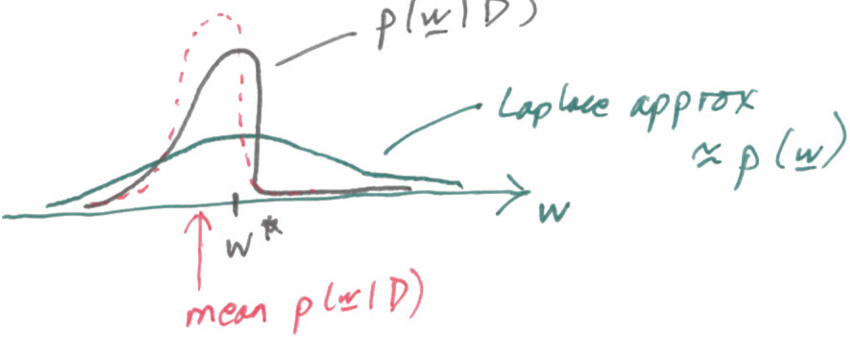
Often Kullback-Leibler Divergence (KL-Divergence)

$$D_{KL}(r\|s) = \int r(\underline{z}) \log \frac{r(\underline{z})}{s(\underline{z})} \, d\underline{z}$$

$$\geq 0 \quad \text{(Gibbs' inequality)}$$

It isn't a distance: $D_{KL}(r\|s) \neq D_{KL}(s\|r)$

Example Logistic Regression N=1



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

$\Rightarrow$ Match mean and variance of posterior.

$q(\underline{w}) = N(\underline{w}; \underline{m}, V)$

$\uparrow\ \uparrow$ posterior mean and covariance.

Don't normally min $D_{KL}(p \| q)$

1) It's harder (check you know why)

2) Often not a good idea:

Example:

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

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

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

$$= \underbrace{-\int q(\underline{w};\alpha) \log p(\underline{w}|D)\, d\underline{w}}_{} + \underbrace{\int q(\underline{w};\alpha) \log q(\underline{w})\, d\underline{w}}_{}$$

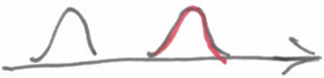
good: $q(\underline{w};\alpha)$ is big when $p(\underline{w}|D)$ is big

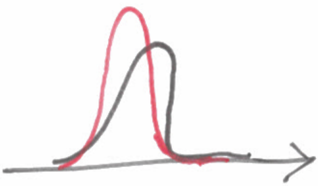
Really bad: $q(\underline{w};\alpha)$ is big when $p(\underline{w}|D)$ is tiny.

$-$Entropy of q

$-H[q(\underline{w};\alpha)]$

Entropy not Hessian.





Substitute in

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

$$D_{KL}(q \| p(\underline{w}|D)) = \underbrace{\mathbb{E}_q[\log q] - \mathbb{E}_q[\log p(\underline{w}, D)]}_{J} + \cancel{\mathbb{E}_q[}\underbrace{\log p(D)}_{\text{log marginal likelihood.}}\cancel{]}$$

can evaluate?

or estimate

Minimize $D_{KL}(q \| p)$ by minimize J.

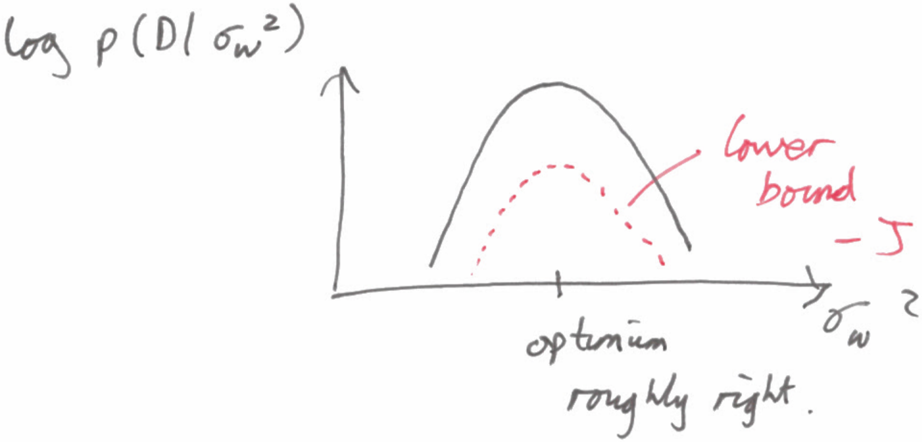
Gibbs' inequality $D_{KL} \geq 0$

$$J + \log p(D) \geq 0$$

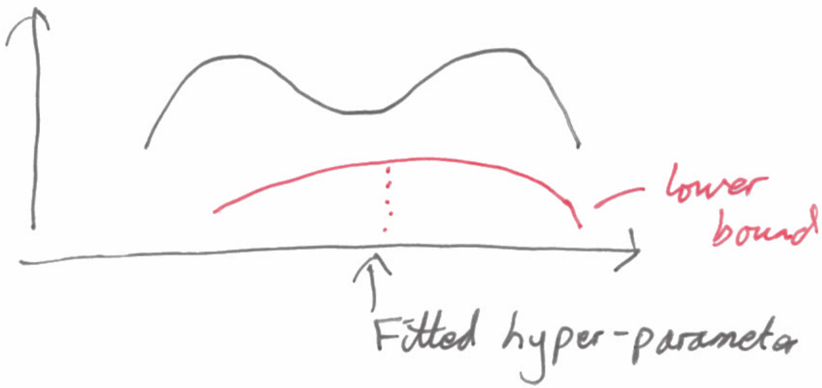
$$\log p(D) \geq -J$$

$\Rightarrow$ Lower bound on marginal likelihood.

Model comparison?


$\log p(D|\sigma_w^2)$
lower bound
$-J$
σ_w^2
optimum
roughly right.

Bad case:


lower bound
Fitted hyper-parameter