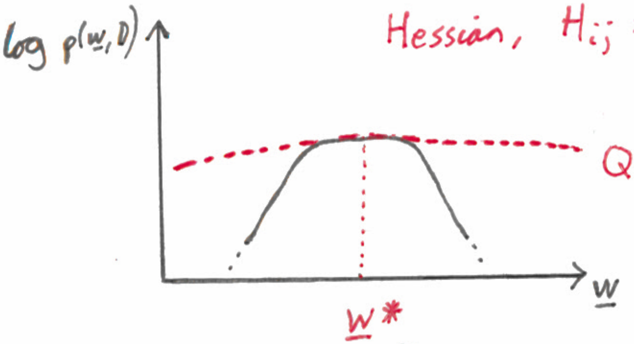
Laplace Approximation

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



Quadratic fit,
tiny curvature

Found with optimizer

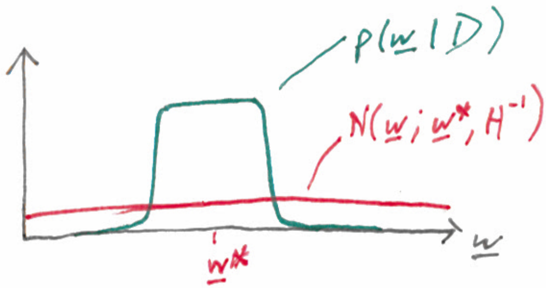
Posterior

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

Marginal Likelihood

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

Areas under both curves $= 1$



Quiz

Is $p(D)$ approx.

A) Too big

B) Too small

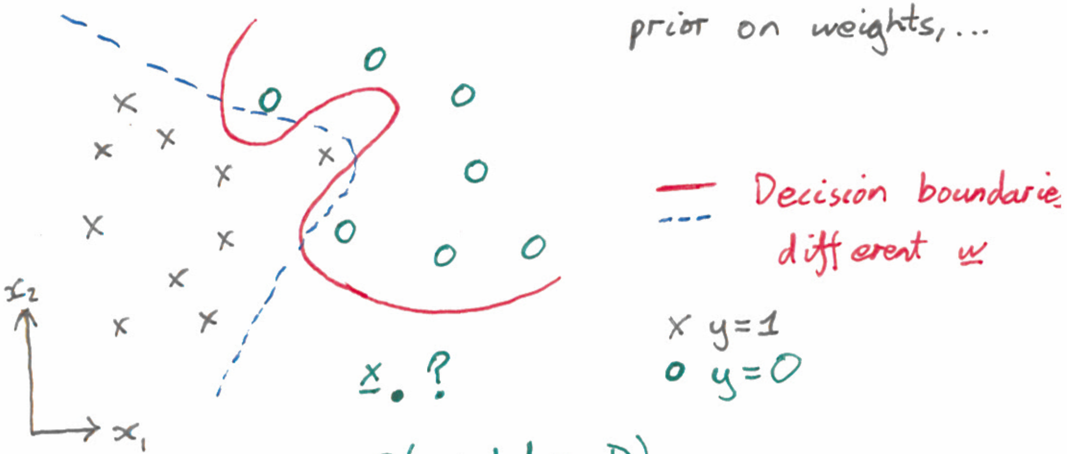
C) $\approx$ Correct

Z) ???

Reminders: why Bayesian stuff?

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

within model,
fixed basis functions,
prior on weights, ...



$$p(y=1 | \underline{x}, D)$$
$$= \int p(y=1 | \underline{x}, \underline{w})\, p(\underline{w} | D)\, d\underline{w}$$

Marginal Likelihood $p(D)$

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

$\Rightarrow$ Score for model assumptions,
How well did we predict training data?

Computing Predictions

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

$$= \int \sigma(\underline{w}^T\underline{x})\, \mathcal{N}(\underline{w}; \underline{w}^*, H^{-1})\, d\underline{w}$$

$$= \mathbb{E}_{\mathcal{N}(\underline{w}; \underline{w}^*, H^{-1})}\left[\sigma(\underline{w}^T\underline{x})\right]$$

(Could do Monte Carlo)

Average under an "activation" $\underline{w}^T\underline{x} = a$

$$= \mathbb{E}_{\underbrace{p(a)}_{\mathcal{N}(a;\, \underline{w}^{*T}\underline{x},\, \underline{x}^T H^{-1}\underline{x})}}\left[\sigma(a)\right]$$

$$= \int \sigma(a)\, \mathcal{N}(a; \underline{w}^{*T}\underline{x}, \underbrace{\underline{x}^T H^{-1}\underline{x}}_{\text{scalar}})\, da$$

Could solve numerically.

Murphy §8.4.4.2

$$\approx \sigma(\kappa\, \underline{w}^{*T}\underline{x})$$

$$\kappa = \frac{1}{\sqrt{1 + \frac{\pi}{8}\underline{x}^T H^{-1}\underline{x}}}$$

Variational Methods

Another way to fit an approx. to the posterior.

$$p(\underline{w} \mid D) \approx q(\underline{w}; \alpha)$$

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

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

Have optimization problem

Fit α, need cost function

measure discrepancy between $p(\underline{w} \mid D)$ and $q(\underline{w})$

A common way to compare distributions

Kullback-Leibler Divergence (KL)

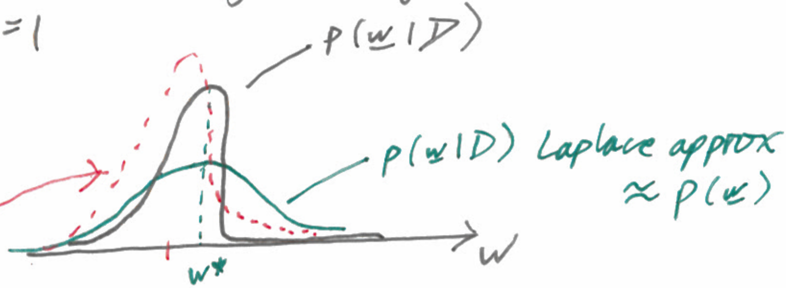
$$D_{KL}(p \,\|\, q) = \int p(\underline{z}) \log \frac{p(\underline{z})}{q(\underline{z})} \, d\underline{z}$$

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

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

Example Logistic Regression

$N = 1$



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

Match mean & variance of posterior.

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

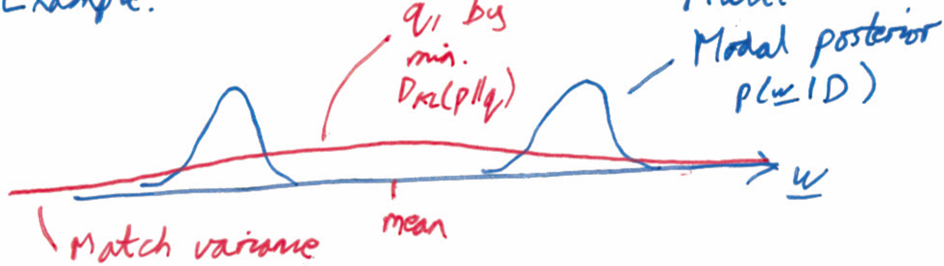
posterior mean & variance.

We don't usually minimize $D_{KL}(\underline{p \| q})$

1) We don't know how.

2) Often not a good idea

Example:



Minimizing $D_{KL}(q \| p)$

$$D_{KL}\left(q(\underline{w};\alpha) \,\|\, p(\underline{w} \mid D)\right)$$

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

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

It's good if $q(\underline{w};\alpha)$ is big when $p(\underline{w} \mid D)$ is big.

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

$-$ Entropy of q

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

Entropy not Hessian.

In 2nd example previous page get similar fit to Laplace approx.

In first example