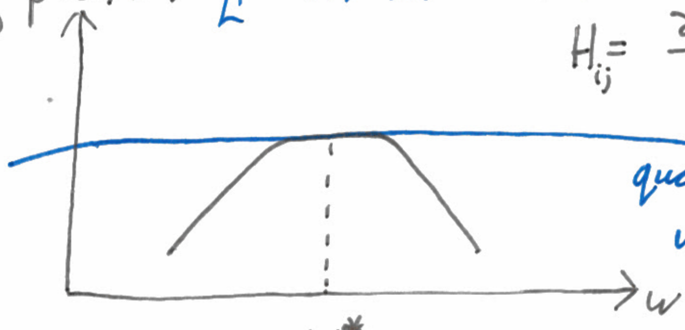
Laplace Approx

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

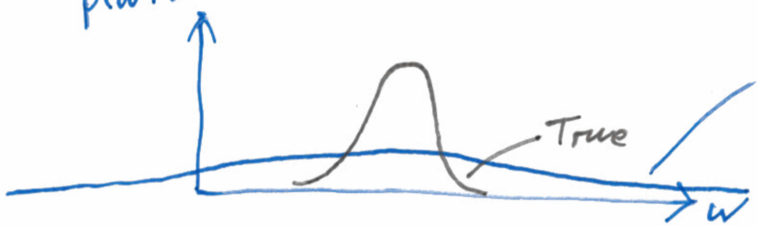
$$\log p(\underline{w}, D) = \log p(D|\underline{w})\,p(\underline{w})$$

Hessian

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



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



Areas under both curves are 1.

Approx: P(D) ? A) Too big; B) Too small; C) ≈ Right; Z)?

Computing Predictions

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

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

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

Average under an activation

$$a = \underline{w}^\top \underline{x}$$

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

mean a: $\underline{w}^{*\top} \underline{x}$

variance a: $\underline{x}^\top H^{-1} \underline{x}$] TODO check.

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

Could solve numerically.

Murphy §8.4.4.2:

$$p(y=1 \mid \underline{x}, D) \approx \sigma(\kappa\, \underline{w}^{*T} \underline{x})$$

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

Variational Methods

Another way to fit approx. to posterior

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

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

Variational Params: $\alpha = \{\underline{m}, V)$

Set up optimization problem.

$\Rightarrow$ Need cost f^n, measures discrepancy between $p(\underline{w} \mid D)$ and q, A common way to compare dists:

Kullback-Leibler Divergence

$\rightarrow$ KL Divergence

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

$$\geqslant 0$$ (Can show Gibbs' inequality.)

Isn't a distance:

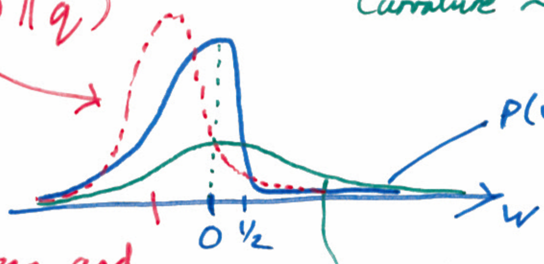
- Not symmetric $D_{KL}(p\|q) \neq D_{KL}(q\|p)$

<u>Logistic Regression eg.</u>

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

Mode at $\approx w=0$

Curvature $\approx$ matches prior



$p(w|D)$

Matches mean and variance of posterior

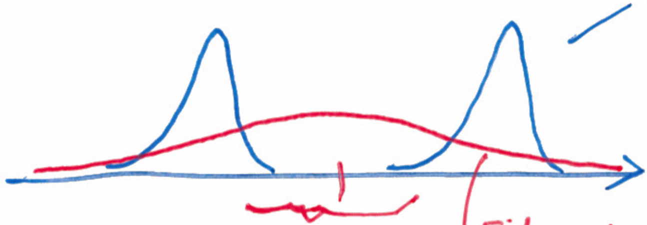
Laplace approx $p(w|D) \approx p(w)$

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

1) We don't know how.
2) Often Not a good idea

Example: match mean and variance

Multimodal $p(w|D)$



Fit minimizing $D_{KL}(p\|q)$

I would predict focussing on implausible parameters.

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

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

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

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

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

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

H ↑ Entropy

Encourages q to be spread out

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

From Bayes' Rule

$$= \underbrace{\mathbb{E}_q[\log q] - \mathbb{E}_q[\log p(D|w)] - \mathbb{E}_q[\log p(w)]}_{\text{Can evaluate, } J} + \underbrace{\cancel{\mathbb{E}_q[}\log p(D)\cancel{]}}_{\text{Don't know but it's a const.}}$$

$$D_{KL} \geqslant 0$$

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

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

$$\mathbb{E}_q[f(\underline{w})] = \int f(\underline{w})\, q(\underline{w};\alpha)\, d\underline{w}$$