Approximate Bayesian Inference

E.g. neural net regression:

Posterior $p(w | D)$

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

Prediction: fitting one mode might be ok.

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

Setting hyper-parameters might work:

$\log p(D)$

true $\log p(D)$

from Laplace approx.

If more modes
Laplace Approximation

\[ \log p(w, D) \]

Hessian, \( H_{ij} = \frac{\partial^2 \log p(w, D)}{\partial w_i \partial w_j} \)

Quadratic fit, tiny curvature

Posterior

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

Marginal Likelihood

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

Areas under both curves = 1

Quiz

Is \( p(D) \) approx.
A) Too big
B) Too small
C) \( \approx \) Correct
D) ???
Variational Methods

Another way to fit approx $p(w|D)$:

$$p(w|D) \approx q(w; \alpha)$$

For $q(w; \alpha) = N(w; m, \Sigma)$

Variational parameters: $\{m, \Sigma\}$

Have optimization problem

Fit $\alpha$, need cost function

Measure discrepancy between $p(w|D)$ and $q(w)$

Often Kullback-Leibler Divergence (KL-Divergence)

$$D_{KL}(r \parallel s) = \int r(z) \log \frac{r(z)}{s(z)} \, dz$$

$\geq 0$ (Gibbs' inequality)

It isn't a distance: $D_{KL}(r \parallel s) \neq D_{KL}(s \parallel r)$
Example Logistic Regression $N=1$

\[
p(w \mid D) \approx p(w)
\]

\[
\text{Laplace approx}
\]

\[w^*\]

mean $p(w \mid D)$

Minimize $D_{KL}(p(w \mid D) \parallel q)$

$\Rightarrow$ Match mean and variance of posterior.

$q(w) = N(w; m, V)$

$\Rightarrow$ posterior mean and covariance.

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

1) It's harder (check you know why)
2) Often not a good idea:

Example:

\[
p(w \mid D)
\]

Multi-Modal.

st. dev. of posterior $w$

mean $w$
Minimizing  \( \text{D}_{\text{KL}} \left( q \parallel p(w \mid D) \right) \)

\[
\text{D}_{\text{KL}} \left( q \parallel p(w \mid D) \right) = \int q(w \mid x) \log \frac{q(w \mid x)}{p(w \mid D)} \, dw
\]

\[
= -\int q(w \mid x) \log p(w \mid D) \, dw + \int q(w \mid x) \log q(w \mid x) \, dw
\]

**Good:** \( q(w \mid x) \) is big when \( p(w \mid D) \) is big

**Really bad:** \( q(w \mid x) \) is big when \( p(w \mid D) \) is tiny.

- Entropy of \( q \)
- \( -H[q(w \mid x)] \)

\[ \text{Entropy not Hessian} \]
Substitute in

\[ p(w|D) = \frac{p(w, D)}{p(D)} \]

\[
D_{KL}(q || p(w|D)) = \mathbb{E}_q[\log q] - \mathbb{E}_q[\log p(w, D)] \]

\[ J \] can evaluate? \\
\text{or estimate} \\
\text{log marginal likelihood.}

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

Gibbs' inequality \( D_{KL} \geq 0 \)

\[ J + \log p(D) \geq 0 \]

\[ \log p(D) \geq -J \]

\[ \Rightarrow \text{Lower bound on marginal likelihood.} \]

Model companion?
Bad case: