Variational Methods

Approximate posterior with \( q(w; x) = N(w; m, V) \) e.g. \( \frac{w}{x} \)

For prediction, fitting one mode might be ok.

Minimize \( \text{D}_\text{KL}( q(w;x) \parallel p(w|D)) \)

\[
\text{D}_\text{KL} = \int q(w;x) \log \frac{q(w;x)}{p(w|D)} \, dw
\]

\[
= \mathbb{E}_q \left[ \log q(w;x) \right] - \mathbb{E}_q \left[ \log p(w|D) \right]
\]

- Entropy \( [q] \)

Could min. this term with \( q = N(w; w^*, 0) \)

⇒ Spread out distribution.
Substitute in
\[ p(w|D) = \frac{p(D|w)p(w)}{p(D)} \]

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

J, "can evaluate" Don't know, log Marginal Likelihood

Minimize \( D_{kl} \), by minimizing \( J \)

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

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

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

\( \Rightarrow \) Lower bound on marginal likelihood.

To fit model choices or hyperparameters

Jointly minimize \( J \) w.r.t. \( \theta, \psi, \beta \)

and w.r.t. model hyperparameters \( \omega^2 \)
Might work...

log p(D | \omega^2)

\sigma^2

Good fit to \sigma^2

lower bound, -J

Bad case:

True function

lower bound

out of bound.
Optimizing $J$

Gradient-based optimization.
Particularly stochastic gradient descent SGD

On $\alpha = \sum m_i V^3$ and hyper... eg $\omega^2$

Unconstrained optimization (Trick #1)

If we optimized $\omega^2$ with SGD we might make it -ve
Optimize $\log \omega$ instead

$V$ has to be positive definite, symmetric

$V = LL^T$, $L$ lower triangular
Diagonal is tve.

We create another matrix

$L_{i,j} = \begin{cases} L_{i,j} & i \neq j \\ \log L_{i,i} & i = j \end{cases}$

Optimize $\tilde{L}$ \(\exp\) diagonal $\rightarrow L \rightarrow V = LL^T \rightarrow$ est cost

SGD $\leftarrow$ backprop.
Evaluating the terms

"Entropy Terms" — we can compute...
For any $m, V, \sigma_w^2$...

Likelihood Term

\[ \mathbb{E}_q [ \log p(D | w) ] \]

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

At least for logistic regression
we can solve numerically.

Stochastic estimate — "Reparameterization trick"

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

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

Sample $w$, by $\varepsilon \sim N(0, I)$

\[ w = m + L \varepsilon \]
Monte Carlo estimate

\[ \approx \frac{1}{S} \sum_{s=1}^{S} f(m + L \varepsilon_s), \]

\( \varepsilon_s \sim N(0, I) \)

simplest approx. \( S = 1 \)

\[ \approx f(m + L \varepsilon), \quad \varepsilon \sim N(0, I) \]

Unbiased estimate.

\[ \nabla_m \mathbb{E}_{N(\varepsilon; 0, I)} \left[ f(m + L \varepsilon) \right] \]

\[ \approx \nabla_m f(m + L \varepsilon), \quad \varepsilon \sim N(0, I) \]

Unbiased

\[ \nabla_L \mathbb{E}_{N(\varepsilon; 0, I)} \left[ f(m + L \varepsilon) \right] \]

... \( \nabla_w f(w) \bigg|_{w=m+L\varepsilon} \)

As long as you can differentiate \( f(w) = \log p(y^{(n)}|x^{(n)}, w) \)

... apply chain rule.