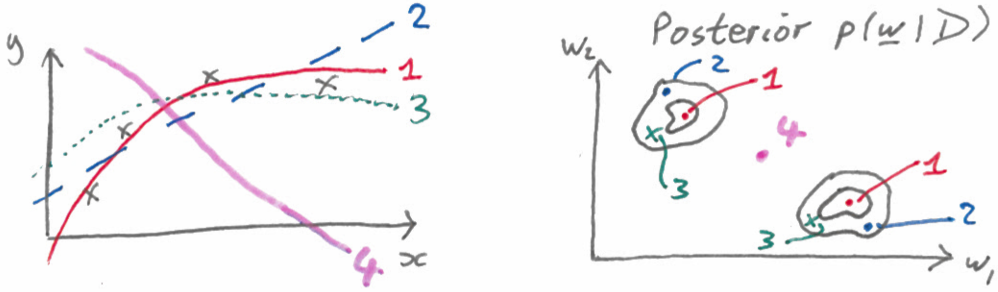
Variational Methods



Approximate posterior with $q(\underline{w};\alpha) \underset{\text{e.g.}}{=} N(\underline{w}; \underbrace{\underline{m}, V}_{\alpha})$

For prediction, fitting one mode <u>might</u> be ok.

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

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

$$= \underbrace{\mathbb{E}_q\left[\log q(\underline{w};\alpha)\right]}_{-\text{Entropy}[q]} - \underbrace{\mathbb{E}_q\left[\log p(\underline{w} \mid D)\right]}_{\text{Could min. this term with } q = N(\underline{w};\underline{w}^*,\underline{\underline{0}})}$$

$\Rightarrow$ Spread out distribution.

Substitute in

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

$$D_{KL} = \underbrace{\mathbb{E}_q[\log q] - \mathbb{E}_q[\log p(D \mid \underline{w})] - \mathbb{E}_q[\log p(\underline{w})]}_{J,\ \text{"can evaluate"}} + \underbrace{\cancel{\mathbb{E}_q}[\log p(D)\cancel{]}}_{\text{Don't know, log Marginal Likelihood}}$$

Minimize D_{KL}, by minimizing J

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

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

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

$\Rightarrow$ Lower bound on marginal likelihood.

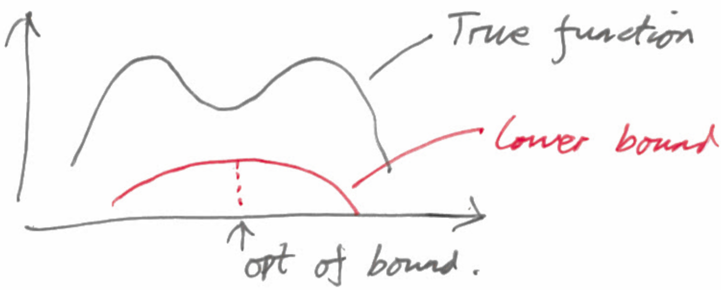
<u>To fit model choices or hyperparameters</u>

Jointly minimize J wrt $\{\underline{m}, V\}$

and wrt model hyperparameters σ_w^2

Might work...

Bad case:



Optimizing J

Gradient-based optimization.

Particulary stochastic gradient descent SGD

On $\alpha = \{\underline{m}, V\}$ and hypers... eg σ_w^2

Unconstrained optimization (Trick # I)

If we optimized σ_w^2 with SGD

we might make it $-$ve

Optimize $\log \sigma_w$ instead

V has to be positive definite, symmetric

$V = LL^T$, L lower triangular

Diagonal is $+$ve.

We create another matrix

$$\tilde{L}_{ij} = \begin{cases} L_{ij} & i \neq j \\ \log L_{ii} & i = j \end{cases}$$

Optimize $\tilde{L} \xrightarrow[\text{exp diagonal}]{} L \rightarrow V = LL^T \rightarrow$ est cost

SGD $\leftarrow$ $\longleftarrow$ backprop.

Evaluating the terms

"Entropy Terms" – we can compute...

For any $\underline{m}, V, \sigma_w^2$...

Likelihood Term

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

$$= \mathbb{E}_q\Big[\underbrace{\textstyle\sum_{n=1}^{N}}\log p(y^{(n)}|\underline{x}^{(n)}, \underline{w})\Big]$$

At least for logistic regression can solve numerically.

Stochastic estimate ("Reparameterization trick")

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

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

Sample $\underline{w}$, by $\underline{\nu} \sim N(0, \mathbb{I})$

$$\underline{w} = \underline{m} + L\underline{\nu}$$

Monte Carlo estimate

$$\approx \frac{1}{S} \sum_{s=1}^{S} f(\underline{m} + L\underline{\nu}^{(s)}),$$

$$\underline{\nu}^{(s)} \sim \mathcal{N}(0, \mathbb{I})$$

laziest
~~Simplest~~ approx $S=1$

$$\approx f(\underline{m} + L\underline{\nu}), \quad \underline{\nu} \sim \mathcal{N}(0, \mathbb{I})$$

Unbiased estimate.

$$\nabla_{\underline{m}} \mathbb{E}_{\mathcal{N}(\underline{\nu}; 0, \mathbb{I})} [f(\underline{m} + L\underline{\nu})]$$

$$\approx \nabla_{\underline{m}} f(\underline{m} + L\underline{\nu}), \quad \underline{\nu} \sim \mathcal{N}(\underline{\nu}; 0, \mathbb{I})$$

Unbiased

$$\nabla_{L} \mathbb{E}_{\mathcal{N}(\underline{\nu}; 0, \mathbb{I})} [f(\underline{m} + L\underline{\nu})]$$

$$\ldots \quad \nabla_{\underline{w}} f(\underline{w}) \Big|_{\underline{w} = \underline{m} + L\underline{\nu}} \underline{\nu}^{\top}$$

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

... apply chain rule.