A (very) brief introduction into how to learn hyperparameters

So far in our coverage of the Bayesian approach to neural networks, the *hyperparameters* α and β were assumed to be known and fixed.

- But this is not a good assumption because...
- ...α corresponds to the width of the prior and β to the noise variance.
- So we really want to learn these from the data as well.
- How can this be done?

We now take a look at one of several ways of addressing this problem.

The Bayesian approach to neural networks

Earlier we looked at the Bayesian approach to *neural networks* using the following notation. We have:

- A neural network computing a function $f(\mathbf{w}; \mathbf{x})$.
- A training sequence $\mathbf{s} = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m))$, split into

$$\mathbf{y} = \begin{pmatrix} y_1 & y_2 & \cdots & y_m \end{pmatrix}$$

  and

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_m \end{pmatrix}$$

The *prior distribution* $p(\mathbf{w})$ is now on the weight vectors and Bayes' theorem tells us that

$$p(\mathbf{w}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{w})p(\mathbf{w})}{p(\mathbf{y})}$$

In addition we have a *Gaussian prior* and a likelihood assuming *Gaussian noise*.

The Bayesian approach to neural networks

The prior and likelihood depend on α and β respectively so we now make this clear and write

$$p(\mathbf{w}|\mathbf{y}, \alpha, \beta) = \frac{p(\mathbf{y}|\mathbf{w}, \beta)p(\mathbf{w}|\alpha)}{p(\mathbf{y}|\alpha, \beta)}$$

(Don't worry about recalling the *actual expressions* for the prior and likelihood just yet, they appear in a few slides time.)

In the earlier slides we found that the Bayes classifier should in fact compute

$$p(Y|\mathbf{y}, \mathbf{x}, \alpha, \beta) = \int_{\mathbb{R}^W} p(y|\mathbf{w}, \mathbf{x}, \beta)p(\mathbf{w}|\mathbf{y}, \alpha, \beta)\, d\mathbf{w}$$

and we found an approximation to this integral. (Again, the necessary parts of the result are repeated later.)

Hierarchical Bayes and the evidence

Let's write down directly something that might be useful to know:

$$p(\alpha, \beta|\mathbf{y}) = \frac{p(\mathbf{y}|\alpha, \beta)p(\alpha, \beta)}{p(\mathbf{y})}$$

If we know $p(\alpha, \beta|\mathbf{y})$ then a straightforward approach is to *use the values for* α *and* β *that maximise it*.

Here is a standard trick: *assume that the prior* $p(\alpha, \beta)$ *is flat*, so that we can just maximise

$$p(\mathbf{y}|\alpha, \beta)$$

This is called *type II maximum likelihood* and is one common way of doing the job.

As usual there are other ways of handling α and β, some of which are regarded as more "correct".

Hierarchical Bayes and the evidence

The quantity

$$p(\mathbf{y}|\alpha, \beta)$$

is called the *evidence*.

When we re-wrote our earlier equation for the posterior density of the weights, making α and β explicit, we found

$$p(\mathbf{w}|\mathbf{y}, \alpha, \beta) = \frac{p(\mathbf{y}|\mathbf{w}, \alpha, \beta)p(\mathbf{w}|\alpha, \beta)}{p(\mathbf{y}|\alpha, \beta)}$$

So *the evidence is the denominator in this equation*.

This is the *common pattern* and leads to the idea of *hierarchical Bayes*: the *evidence for the hyperparameters* at one level is the *denominator in the relevant application of Bayes theorem*.

An expression for the evidence

We have already *derived everything necessary* to write an *explicit equation for the evidence* for the case of regression that we've been following.

First, as we know about a lot of expressions involving $\mathbf{w}$ we can introduce it by the standard trick of *marginalising*:

$$\begin{aligned} p(\mathbf{y}|\alpha, \beta) &= \int p(\mathbf{y}, \mathbf{w}|\alpha, \beta) d\mathbf{w} \\ &= \int p(\mathbf{y}|\mathbf{w}, \alpha, \beta) p(\mathbf{w}|\alpha, \beta) d\mathbf{w} \\ &= \int p(\mathbf{y}|\mathbf{w}, \beta) p(\mathbf{w}|\alpha) d\mathbf{w} \end{aligned}$$

where we've made the obvious independence simplifications.

The two densities in this integral *are just the likelihood and prior we've already studied*.

We've just conditioned on α and β, which previously were constants but are now being treated as random variables.

An expression for the evidence

Here are the actual expression for the prior and likelihood.

The prior is

$$p(\mathbf{w}|\alpha) = \frac{1}{Z_W(\alpha)} \exp\left(-\alpha E_W(\mathbf{w})\right)$$

where

$$Z_W(\alpha) = \left(\frac{2\pi}{\alpha}\right)^{W/2} \text{ and } E_W(\mathbf{w}) = \frac{1}{2}||\mathbf{w}||^2$$

and the likelihood is

$$p(\mathbf{y}|\mathbf{w}, \beta) = \frac{1}{Z_{\mathbf{y}}(\beta)} \exp\left(-\beta E_{\mathbf{y}}(\mathbf{w})\right)$$

where

$$Z_{\mathbf{y}}(\beta) = \left(\frac{2\pi}{\beta}\right)^{m/2} \text{ and } E_{\mathbf{y}}(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{m}(y_i - h(\mathbf{w}; \mathbf{x}_i))^2$$

Both of these equations have been copied directly from earlier slides: *there is nothing to add*.

An expression for the evidence

That gives us

$$p(\mathbf{y}|\alpha, \beta) = \left(\frac{2\pi}{\alpha}\right)^{-W/2} \left(\frac{2\pi}{\beta}\right)^{-m/2} \int \exp\left(-S(\mathbf{w})\right) d\mathbf{w}$$

where

$$S(\mathbf{w}) = \alpha E_W(\mathbf{w}) + \beta E_{\mathbf{y}}(\mathbf{w})$$

This is *exactly the integral we first derived an approximation for*.

Specifically

$$\int \exp\left(-S(\mathbf{w})\right) d\mathbf{w} \simeq (2\pi)^{W/2} |\mathbf{A}|^{-1/2} \exp(-S(\mathbf{w}_{\text{MAP}}))$$

where

$$\mathbf{A} = \alpha \mathbf{I} + \beta \nabla\nabla E_{\mathbf{y}}(\mathbf{w}_{\text{MAP}})$$

and $\mathbf{w}_{\text{MAP}}$ is the *maximum a posteriori solution*.

An expression for the evidence

Putting all that together we get an *expression for the logarithm of the evidence*:

$$\boxed{\begin{aligned}\log p(\mathbf{y}|\alpha,\beta) \simeq & \frac{W}{2}\log\alpha - \frac{m}{2}\log 2\pi + \frac{m}{2}\log\beta \\ & - \frac{1}{2}\log|\mathbf{A}| \\ & - \alpha E_W(\mathbf{w}_{\text{MAP}}) - \beta E_{\mathbf{y}}(\mathbf{w}_{\text{MAP}})\end{aligned}}$$

Again, we're using the fact that we want to *maximise the evidence* and this is equivalent to *maximising its logarithm* which turns a product into a more friendly sum.

Maximising the evidence

We want to maximise this, so let's differentiate it with respect to α and β.

For α

$$\frac{\partial \log p(\mathbf{y}|\alpha,\beta)}{\partial \alpha} = \frac{W}{2\alpha} - E_W(\mathbf{w}_{\text{MAP}}) - \frac{1}{2}\frac{\partial \log |\mathbf{A}|}{\partial \alpha}$$

How do we handle the final term? This is straightforward if we can compute the *eigenvalues* of $\mathbf{A}$.

Recall that the n eigenvalues λ_i and n eigenvectors $\mathbf{v}_i$ of an $n \times n$ matrix $\mathbf{M}$ are defined such that

$$\mathbf{M}\mathbf{v}_i = \lambda_i \mathbf{v}_i \text{ for } i = 1, \ldots, n$$

and the eigenvectors are orthonormal

$$\mathbf{v}_i^T \mathbf{v}_j = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise.} \end{cases}$$

One standard result is that *the determinant of a matrix is the product of its eigenvalues*.

$$|\mathbf{M}| = \prod_{i=1}^{n} \lambda_i$$

Maximising the evidence

We have

$$\mathbf{A} = \alpha\mathbf{I} + \beta\nabla\nabla E_{\mathbf{y}}(\mathbf{w}_{\mathrm{MAP}})$$

Say the eigenvalues of $\beta\nabla\nabla E_{\mathbf{y}}(\mathbf{w}_{\mathrm{MAP}})$ are λ_i. *(These can be computed using standard numerical algorithms.)*

Then the eigenvalues of $\mathbf{A}$ are $\alpha + \lambda_i$ and

$$\begin{aligned}
\frac{\partial \log|\mathbf{A}|}{\partial\alpha} &= \frac{\partial}{\partial\alpha}\left(\log\prod_{i=1}^{W}(\alpha+\lambda_i)\right) \\
&= \frac{\partial}{\partial\alpha}\left(\sum_{i=1}^{W}\log(\alpha+\lambda_i)\right) \\
&= \sum_{i=1}^{W}\frac{1}{\alpha+\lambda_i}\frac{\partial(\alpha+\lambda_i)}{\partial\alpha}
\end{aligned}$$

This remains tricky because *the eigenvalues might be functions of α.*

Maximising the evidence

To make further progress, assume *(sometimes correct, sometimes not!)* that the λ_i *do not* depend on α.

In that case

$$\frac{\partial \log|\mathbf{A}|}{\partial \alpha} = \sum_{i=1}^{W} \frac{1}{\alpha + \lambda_i}$$
$$= \mathrm{Trace}(\mathbf{A}^{-1})$$

because $\mathbf{M}^{-1}$ has eigenvalues $1/\lambda_i$ and the trace of a matrix is equal to the sum of its eigenvalues.

Finally, equating the derivative to zero gives:

$$\frac{W}{2\alpha} - E_W(\mathbf{w}_{\mathrm{MAP}}) - \frac{1}{2}\mathrm{Trace}(\mathbf{A}^{-1}) = 0$$

or

$$\alpha = \frac{1}{2E_W(\mathbf{w}_{\mathrm{MAP}})}\left(W - \sum_{i=1}^{W} \frac{\alpha}{\alpha + \lambda_i}\right)$$

which can be used to update the value for α.

Maximising the evidence

We can now repeat the process to obtain an update for β:

$$\frac{\partial \log p(\mathbf{y}|\alpha,\beta)}{\partial \beta} = \frac{m}{2\beta} - E_{\mathbf{y}}(\mathbf{w}_{\mathrm{MAP}}) - \frac{1}{2}\frac{\partial \log |\mathbf{A}|}{\partial \beta}$$

In this case

$$\begin{aligned}
\frac{\partial \log |\mathbf{A}|}{\partial \beta} &= \frac{\partial}{\partial \beta}\left(\sum_{i=1}^{W} \log(\alpha+\lambda_i)\right) \\
&= \sum_{i=1}^{W} \frac{1}{\alpha+\lambda_i}\frac{\partial}{\partial \beta}(\alpha+\lambda_i) \\
&= \sum_{i=1}^{W} \frac{1}{\alpha+\lambda_i}\frac{\partial \lambda_i}{\partial \beta}
\end{aligned}$$

and again we have a *potentially tricky derivative*.

Maximising the evidence

As the λ_i are the eigenvalues of $\beta \nabla\nabla E_{\mathbf{y}}(\mathbf{w}_{\mathrm{MAP}})$ we have

$$\frac{\partial \lambda_i}{\partial \beta} = \frac{\lambda_i}{\beta}$$

(can you see why?) so

$$\frac{\partial \log |\mathbf{A}|}{\partial \beta} = \frac{1}{\beta} \sum_{i=1}^{W} \frac{\lambda_i}{\alpha + \lambda_i}$$

Equating the derivative to zero gives

$$\beta = \frac{1}{2E_{\mathbf{y}}(\mathbf{w}_{\mathrm{MAP}})} \left(m - \sum_{i=1}^{W} \frac{\lambda_i}{\alpha + \lambda_i} \right)$$

which can be used to update the value for β.

Maximising the evidence

Here's why the derivative works.

Say

$$\mathbf{M} = \nabla\nabla E_{\mathbf{y}}(\mathbf{w}_{\text{MAP}})$$

so we're interested in $\partial\lambda_i/\partial\beta$ when the λ_i are the eigenvalues of $\beta\mathbf{M}$. Thus

$$(\beta\mathbf{M})\mathbf{v}_i = \lambda_i\mathbf{v}_i$$

and using the fact that the eigenvectors are orthonormal

$$\beta\mathbf{v}_i^T\mathbf{M}\mathbf{v}_i = \lambda_i\mathbf{v}_i^T\mathbf{v}_i = \lambda_i.$$

So

$$\mathbf{v}_i^T\mathbf{M}\mathbf{v}_i = \frac{\lambda_i}{\beta}$$

and

$$\frac{\partial\lambda_i}{\partial\beta} = \mathbf{v}_i^T\mathbf{M}\mathbf{v}_i = \frac{\lambda_i}{\beta}.$$

Summary:

Define

$$\theta_t = \sum_{i=1}^{W} \frac{\lambda_i}{\alpha_t + \lambda_i}$$

where the subscript denotes the fact that we're using the following equations to periodically update our estimates of α and β.

Collecting the two update equations together we have

$$\boxed{\alpha_{t+1} = \frac{\theta_t}{2E_W(\mathbf{w}_{MAP})}}$$

and

$$\boxed{\beta_{t+1} = \frac{m - \theta_t}{2E_{\mathbf{y}}(\mathbf{w}_{MAP})}}$$

Maximising the evidence

This suggests a *method for the overall learning process*:

1. Choose the initial values α_0 and β_0 at random.
2. Choose an initial weight vector $\mathbf{w}$ according to the prior.
3. Use a standard optimisation algorithm to iteratively estimate $\mathbf{w}_{\text{MAP}}$.
4. While the optimisation progresses, periodically use the equations above to re-estimate α and β.

Step 4 requires that we compute an eigendecomposition, which might well be time-consuming. If necessary we can make a simplification.

When $m >> W$ it is reasonable to expect that $\theta_t \simeq W$ an so we can use

$$\boxed{\alpha_{t+1} = \frac{W}{2E_W(\mathbf{w}_{\text{MAP}})}}$$

and

$$\boxed{\beta_{t+1} = \frac{m}{2E_{\mathbf{y}}(\mathbf{w}_{\text{MAP}})}}$$

An alternative: integrate the hyperparameters out

While choosing α and β by maximising the evidence leads to an effective algorithm, it might be argued that a more correct way to deal with these parameters would be to *integrate them out*.

$$p(\mathbf{w}|\mathbf{y}) = \int\int p(\mathbf{w}, \alpha, \beta|\mathbf{y}) d\alpha d\beta.$$

(Recall the *general equation for probabilistic inference* where we integrate out unobserved random variables.)

Re-arranging this we have

$$\begin{aligned}
\int\int p(\mathbf{w}, \alpha, \beta|\mathbf{y}) d\alpha d\beta &= \frac{1}{p(\mathbf{y})} \int\int p(\mathbf{y}|\mathbf{w}, \alpha, \beta) p(\mathbf{w}, \alpha, \beta) d\alpha d\beta \\
&= \frac{1}{p(\mathbf{y})} \int\int p(\mathbf{y}|\mathbf{w}, \alpha, \beta) p(\mathbf{w}|\alpha, \beta) p(\alpha, \beta) d\alpha d\beta \\
&= \frac{1}{p(\mathbf{y})} \int\int p(\mathbf{y}|\mathbf{w}, \beta) p(\mathbf{w}|\alpha) p(\alpha) p(\beta) d\alpha d\beta
\end{aligned}$$

where we're assuming α and β are independent.

An alternative: integrate the hyperparameters out

In order to continue we need to specify priors on α and β.

On this occasion we have a good reason to choose particular priors, as α and β are *scale parameters*.

In general, a scale parameter σ is one that appears in a density of the form

$$p(x|\sigma) = \frac{1}{\sigma} f\left(\frac{x}{\sigma}\right)$$

The standard deviation of a Gaussian density is an example.

What happens to this density if we *scale* x such that $x' = cx$?

Standard result number 1

We need to recall how to deal with *transformations of continuous random variables*.

Say we have a random variable x with *probability density* $p_x(x)$.

We then transform x to $y = f(x)$ where f is strictly increasing.

What is the probability density function of y*?* There is a standard method for computing this. (See NST maths, or the 1A Probability course.)

$$p_y(y) = \frac{p_x(f^{-1}(y))}{f'(f^{-1}(y))}$$

An alternative: integrate the hyperparameters out

Applying this when $x' = cx$ we have

$$f(x) = cx$$

$$f^{-1}(x') = \frac{x'}{c}$$

$$f'(x) = c$$

and so

$$p_{x'}(x') = \frac{1}{c\sigma} f\left(\frac{x'}{c\sigma}\right) = \frac{1}{\sigma'} f\left(\frac{x'}{\sigma'}\right)$$

Thus the transformation leaves the density essentially unchanged, and in particular we want the densities $p(\sigma)$ and $p(\sigma')$ to be identical.

It turns out that this forces the choice

$$p(\sigma) = \frac{c'}{\sigma}.$$

This is an *improper prior* and it is conventional to take $c' = 1$.

Standard result number 2

Returning to the integral of interest

$$\frac{1}{p(\mathbf{y})} \int\int p(\mathbf{y}|\mathbf{w}, \beta)p(\mathbf{w}|\alpha)p(\alpha)p(\beta)d\alpha d\beta$$

Taking the integral for α first we have

$$\begin{aligned}\int p(\mathbf{w}|\alpha)p(\alpha)d\alpha &= \int \frac{1}{\alpha Z_W(\alpha)} \exp(-\alpha E_W(\mathbf{w}))d\alpha \\ &= \int \frac{1}{\alpha}\left(\frac{\alpha}{2\pi}\right)^{W/2} \exp\left(-\frac{\alpha}{2}||\mathbf{w}||^2\right) d\alpha\end{aligned}$$

and to evaluate this we use the following *standard result*:

$$\int_0^\infty x^n \exp(-ax)dx = \frac{\Gamma(n+1)}{a^{n+1}}$$

where $n > -1$ and $a > 0$. So the integral becomes

$$(2\pi)^{-W/2}\frac{\Gamma(W/2)}{E_W(\mathbf{w})^{W/2}}$$

An alternative: integrate the hyperparameters out

Repeating the process for β and using the same standard result we have

$$\begin{aligned}\int p(\mathbf{y}|\mathbf{w},\beta)p(\beta)d\beta &= \int \frac{1}{\beta}\left(\frac{\beta}{2\pi}\right)^{m/2} \exp(-\beta E_{\mathbf{y}}(\mathbf{w}))d\beta \\ &= (2\pi)^{-m/2}\frac{\Gamma(m/2)}{E_{\mathbf{y}}(\mathbf{w})^{m/2}}\end{aligned}$$

Combining the two expression we obtain

$$\begin{aligned}-\log p(\mathbf{w}|\mathbf{y}) &= -\log\left(\frac{1}{p(\mathbf{y})}(2\pi)^{-W/2}\frac{\Gamma(W/2)}{E_W(\mathbf{w})^{W/2}}(2\pi)^{-m/2}\frac{\Gamma(m/2)}{E_{\mathbf{y}}(\mathbf{w})^{m/2}}\right) \\ &= \frac{W}{2}\log E_W(\mathbf{w}) + \frac{m}{2}\log E_{\mathbf{y}}(\mathbf{w}) + \text{constant}\end{aligned}$$

and *we want to minimise this* so we need

$$\boxed{\frac{W}{2}\frac{1}{E_W(\mathbf{w})}\frac{\partial E_W(\mathbf{w})}{\partial \mathbf{w}} + \frac{m}{2}\frac{1}{E_{\mathbf{y}}(\mathbf{w})}\frac{\partial E_{\mathbf{y}}(\mathbf{w})}{\partial \mathbf{w}} = 0}$$

An alternative: integrate the hyperparameters out

The *actual value for the evidence* is

$$-\log p(\mathbf{w}|\mathbf{y}) = -\log\left(\frac{1}{p(\mathbf{y})}\frac{1}{Z_{\mathbf{y}}(\alpha,\beta)}\exp(-(\alpha E_W(\mathbf{w}) + \beta E_{\mathbf{y}}(\mathbf{w})))\right)$$
$$= \alpha E_W(\mathbf{w}) + \beta E_{\mathbf{y}}(\mathbf{w}) + \text{constant}$$

and *we want to minimise this* so we need

$$\boxed{\alpha\frac{\partial E_W(\mathbf{w})}{\partial \mathbf{w}} + \beta\frac{\partial E_{\mathbf{y}}(\mathbf{w})}{\partial \mathbf{w}} = 0}$$

This should make us *VERY VERY HAPPY* because if we equate the two boxed equations we get

$$\alpha = \frac{W}{2E_W(\mathbf{w})}$$

and

$$\beta = \frac{m}{2E_{\mathbf{y}}(\mathbf{w})}$$

and so the result for *integrating out the hyperparameters* agrees with the result for *optimising the evidence*.