Neural Nets: Initialisation Strategies

David Barber

Initialisation

- Particularly for saturating transfer functions that have regions of zero gradient, sensible initialisation of the weights is essential.
- Consider a simple regression net with error

$$\sum_n \left(y^n - f\left(\mathbf{w}^\mathsf{T}\mathbf{x}^n\right)\right)^2$$

- Firstly, it makes sense to scale the error so that it remains roughly order 1, even as the number of datapoints increases. For example

$$E(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N} \left(y^n - f\left(\mathbf{w}^\mathsf{T}\mathbf{x}^n\right)\right)^2$$

- Gradient is

$$E(\mathbf{w}) = -\frac{2}{N}\sum_{n=1}^{N} \left(y^n - f\left(\mathbf{w}^\mathsf{T}\mathbf{x}^n\right)\right) f'\left(\mathbf{w}^\mathsf{T}\mathbf{x}^n\right)\mathbf{x}^n$$

- For $f(x) = 1/(1+e^{-x})$, for large $|x|$, this function saturates to 1 or 0 and the gradient becomes zero. We need therefore to ensure we don't immediately get trapped in these zero gradient regions.

Initialisation

- Let's assume that we have scaled the inputs x_i so that they have zero mean and unit variance

$$x_i^n \to \frac{x_i^n - \mu_i}{\sigma_i}$$

where

$$\mu_i = \sum_{n=1}^{N} x_i^n, \qquad \sigma_i^2 = \sum_{n=1}^{N} (x_i^n - \mu_i)^2$$

After this rescaling,

$$\frac{1}{N}\sum_{n=1}^{N} x_i^n = 0, \qquad \frac{1}{N}\sum_{n=1}^{N} (x_i^n)^2 = 1$$

Initialisation

- Define the activation (input of the transfer function) as

$$z_n \equiv \mathbf{w}^{\mathsf{T}} \mathbf{x}^n$$

Let's assume that we sample each w_i identically and independently from a zero mean distribution $\mathbb{E}\left(w_i\right) = 0$.

$$\mathbb{E}\left(z_n\right) = \sum_i \mathbb{E}\left(w_i\right) x_i^n = 0$$

$$\begin{aligned} \mathbb{E}\left(z_n^2\right) &= \sum_{i \neq j} \mathbb{E}\left(w_i\right) \mathbb{E}\left(w_j\right) x_i^n x_j^n + \sum_i \mathbb{E}\left(w_i^2\right) \left(x_i^n\right)^2 \\ &= \mathbb{E}\left(w^2\right) \sum_{i=1}^{D} \left(x_i^n\right)^2 \end{aligned}$$

- Hence if we sample the w_i from a distribution with zero mean and variance $1/D$, then the activation will be zero mean with variance

$$\mathbb{E}\left(z_n^2\right) = \frac{1}{D} \sum_{i=1}^{D} \left(x_i^n\right)^2$$

Initialisation

- Since each x_i^n has values roughly in the range -2 to 2 (2 standard deviations from the zero mean), then the activation will have values roughly in the range -2 to 2 as well.
- A reasonable initialisation then of the weights of a network with transfer function $f(x)$ that has a non-saturating region when x is between -2 and 2, is: Draw each w_i from a distribution with zero mean and unit variance and then rescale. For example:

$$w_i \sim \mathcal{N}\left(w_i|0,1\right), \qquad w_i \to \frac{w_i}{\sqrt{D}}$$

$$w_i \sim \mathsf{sign}\left(\mathcal{N}\left(w_i|0,1\right)\right), \qquad w_i \to \frac{w_i}{\sqrt{D}}$$

$$w_i \sim \mathcal{U}\left(w_i|-\sqrt{3},\sqrt{3}\right), \qquad w_i \to \frac{w_i}{\sqrt{D}}$$

- For nets with multiple hidden layers, the previous initialisation strategy is also commonly used. or weights in layer l, we set (for example)

$$w_i^l \sim \mathcal{N}\left(w_i^l|0,1\right), \qquad w_i^l \to \frac{w_i^l}{\sqrt{D_{l-1}}}$$

where D_l is the number of units in layer l.

Decorrelating Inputs

For an $\mathbf{x} \to y$ net the Hessian is

$$\mathbf{H} = -\frac{2}{N} \sum_n \underbrace{\left[\left(y^n - f(\mathbf{w}^\mathsf{T} \mathbf{x}_n \right) f'' \left(\mathbf{w}^\mathsf{T} \mathbf{x}_n \right) - \left(f'(\mathbf{w}^\mathsf{T} \mathbf{x}_n) \right)^2 \right]}_{\equiv \gamma_n} \mathbf{x}_n \mathbf{x}_n^\mathsf{T}$$

As a very crude approximation, we can assume that all the γ_n are roughly equal to a value γ. In that case, the approximate Hessian is

$$\mathbf{H} \approx -\frac{2\gamma}{N} \sum_n \mathbf{x}_n \mathbf{x}_n^\mathsf{T}$$

If we perform SVD of the matrix $\mathbf{X} = \left[\mathbf{x}^1, \ldots, \mathbf{x}^N \right]$ then

$$\sum_n \mathbf{x}_n \mathbf{x}_n^\mathsf{T} = \mathbf{X}\mathbf{X}^\mathsf{T} = \mathbf{U}\mathbf{S}\mathbf{V}^\mathsf{T}\mathbf{V}\mathbf{S}^\mathsf{T}\mathbf{U}^\mathsf{T} = \mathbf{U}\mathbf{S}^\mathbf{2}\mathbf{U}^\mathsf{T}$$

Hence applying the 'whitening' transformation

$$\mathbf{x}^n \to \mathbf{S}^{-1}\mathbf{U}^\mathsf{T}\mathbf{x}^n$$

will make the approximate Hessian diagonal and can be a useful initial preprocessing step to make optimisation easier.

A more complex discussion

- Let's consider what happens for the second hidden layer. This will have inputs

$$f(z_{nj}), \qquad z_{nj} \equiv \mathbf{w}_j^\mathsf{T}\mathbf{x}^n$$

  where j is the index of the neuron in the first layer.
- And the activation to a unit in the second layer will be

$$\tilde{z}_n \equiv \sum_j u_j f(z_{nj})$$

- If we assume that $z_{nj} \sim \mathcal{N}(z_{nj}|0,1)$ and the u_j are independently sampled from a zero mean distribution then the activation of the second layer has zero mean and variance

$$\mathbb{E}\left(u^2\right) \sum_{j=1}^{D_1} \left(f(z_{nj})\right)^2$$

- If we therefore draw each u_j^l from a distribution with zero mean and variance

$$\frac{1}{D_{l-1}\mathbb{E}\left(f(z)^2\right)_{\mathcal{N}(z|0,1)}}$$

  the activation of each unit j in layer l will be roughly zero mean with unit variance distributed.