

To see that the weight scaling rule is exact, we can simplify $\tilde{P}_{\text{ensemble}}$:

$$\tilde{P}_{\text{ensemble}}(y = y \mid \mathbf{v}) = \sqrt[2^n]{\prod_{\mathbf{d} \in \{0,1\}^n} P(y = y \mid \mathbf{v}; \mathbf{d})} \quad (7.60)$$

$$= \sqrt[2^n]{\prod_{\mathbf{d} \in \{0,1\}^n} \text{softmax}(\mathbf{W}^\top (\mathbf{d} \odot \mathbf{v}) + \mathbf{b})_y} \quad (7.61)$$

$$= \sqrt[2^n]{\prod_{\mathbf{d} \in \{0,1\}^n} \frac{\exp(\mathbf{W}_{y,:}^\top (\mathbf{d} \odot \mathbf{v}) + b_y)}{\sum_{y'} \exp(\mathbf{W}_{y',:}^\top (\mathbf{d} \odot \mathbf{v}) + b_{y'})}} \quad (7.62)$$

$$= \frac{\sqrt[2^n]{\prod_{\mathbf{d} \in \{0,1\}^n} \exp(\mathbf{W}_{y,:}^\top (\mathbf{d} \odot \mathbf{v}) + b_y)}}{\sqrt[2^n]{\prod_{\mathbf{d} \in \{0,1\}^n} \sum_{y'} \exp(\mathbf{W}_{y',:}^\top (\mathbf{d} \odot \mathbf{v}) + b_{y'})}} \quad (7.63)$$

Because \tilde{P} will be normalized, we can safely ignore multiplication by factors that are constant with respect to y :

$$\tilde{P}_{\text{ensemble}}(y = y \mid \mathbf{v}) \propto \sqrt[2^n]{\prod_{\mathbf{d} \in \{0,1\}^n} \exp(\mathbf{W}_{y,:}^\top (\mathbf{d} \odot \mathbf{v}) + b_y)} \quad (7.64)$$

$$= \exp\left(\frac{1}{2^n} \sum_{\mathbf{d} \in \{0,1\}^n} \mathbf{W}_{y,:}^\top (\mathbf{d} \odot \mathbf{v}) + b_y\right) \quad (7.65)$$

$$= \exp\left(\frac{1}{2} \mathbf{W}_{y,:}^\top \mathbf{v} + b_y\right). \quad (7.66)$$

Substituting this back into equation 7.58 we obtain a softmax classifier with weights $\frac{1}{2}\mathbf{W}$.

The weight scaling rule is also exact in other settings, including regression networks with conditionally normal outputs, and deep networks that have hidden layers without nonlinearities. However, the weight scaling rule is only an approximation for deep models that have nonlinearities. Though the approximation has not been theoretically characterized, it often works well, empirically. [Goodfellow et al. \(2013a\)](#) found experimentally that the weight scaling approximation can work better (in terms of classification accuracy) than Monte Carlo approximations to the ensemble predictor. This held true even when the Monte Carlo approximation was allowed to sample up to 1,000 sub-networks. [Gal and Ghahramani \(2015\)](#) found that some models obtain better classification accuracy using twenty samples and