$$L \sim \sum_{i} I_i(1-||W_i||^2)^2 \quad + \quad \sum_{i\neq j} I_j(W_j \cdot W_i)^2$$

Feature benefit is the value a model attains from representing a feature. In a real neural network, this would be analagous to the potential of a feature to improve predictions if represented accurately.

Interference betwen x_i and x_j occurs when two features are embedded non-orthogonally and, as a result, affect each other's predictions. This prevents superposition in linear models.

The Saxe results reveal that there are fundamentally two competing forces which control learning dynamics in the considered model. Firstly, the model can attain a better loss by representing more features (we've labeled this "feature benefit"). But it also gets a worse loss if it represents more than it can fit orthogonally due to "interference" between features.[11] In fact, this makes it never worthwhile for the linear model to represent more features than it has dimensions.[12]

Can we achieve a similar kind of understanding for the ReLU output model? Concretely, we'd like to understand $L = \int_x ||I(x - \text{ReLU}(W^TWx+b))||^2 d\mathbf{p}(x)$ where x is distributed such that $x_i = 0$ with probability S.

The integral over x decomposes into a term for each sparsity pattern according to the binomial expansion of $((1-S)+S)^n$. We can group terms of the sparsity together, rewriting the loss as $L = (1-S)^n L_n + \ldots + \ (1-S)S^{n-1}L_1 + S^n L_0$, with each L_k corresponding to the loss when the input is a k-sparse vector. Note that as $S \to 1$, L_1 and L_0 dominate. The L_0 term, corresponding to the loss on a zero vector, is just a penalty on positive biases, $\sum_i \text{ReLU}(b_i)^2$. So the interesting term is L_1, the loss on 1-sparse vectors:

$$L_1 = \sum_{i} \int_{0\le x_i\le 1} I_i(x_i - \text{ReLU}(||W_i||^2 x_i + b_i))^2 \quad + \quad \sum_{i\neq j} \int_{0\le x_i\le 1} I_j \text{ReLU}(W_j \cdot W_i x_i + b_j)^2$$

If we focus on the case $x_i = 1$, we get something which looks even more analagous to the linear case:

$$= \sum_{i} I_i(1 - \text{ReLU}(||W_i||^2 + b_i))^2 \quad + \quad \sum_{i\neq j} I_j \text{ReLU}(W_j \cdot W_i + b_j)^2$$

Feature benefit is similar to before. Note that ReLU never makes things worse, and that the bias can help when the model doesn't represent a feature by taking on the expected value.

Interference is similar to before but ReLU means that negative interference, or interference where a negative bias pushes it below zero, is "free" in the 1-sparse case.

This new equation is vaguely similar to the famous Thomson problem in chemistry. In particular, if we assume uniform importance and that there are a fixed number of features with $||W_i|| = 1$ and the rest have $||W_i|| = 0$, and that $b_i = 0$, then the feature benefit term is constant and the interference term becomes a generalized Thomson problem – we're just packing points on the surface of the sphere with a slightly unusual energy function. (We'll see this can be a productive analogy when we resume our empirical investigation in the following sections!)

Another interesting property is that ReLU makes negative interference free in the 1-sparse case. This explains why the solutions we've seen prefer to only have negative interference when possible. Further, using a negative bias can convert small positive interferences into essentially being negative interferences.