A EXPONENTIAL FAMILIES

This section reviews the standard results on exponential families in the literature [16, 21]. A 1-dimensional standard exponential family [2, page 2] at its core can be represented by a reference measure \((\mathbf{R}, \mathcal{B}, \rho_p)\), where the set of outcomes (whether possible or not: see the definition of \(\Omega_p\), below) is given by \(\mathbf{R}\), the real numbers; the set of measurable sets is given by \(\mathcal{B}\), the Borel algebra of \(\mathbf{R}\); and \(\rho_p : \mathcal{B} \to [0, \infty] \) is a positive measure (often not a probability measure). Define \(A_p : \mathbf{R} \to [-\infty, +\infty]\) to be a normalization function where:

\[
A_p(\theta) = \ln \left( \int \exp(\theta \omega) d\rho_p(\omega) \right).
\]  
(14)

If \(A_p\) is finite for all \(\theta \in \mathbf{R}\), then \(p\) is defined to be full [2]. If \(p\) is full, then for each \(\theta \in \mathbf{R}\), there is a distribution in the family of the form:

\[
\forall B \in \mathcal{B}, \Pr[x \in B|\theta] = \int_B \exp(\theta \omega - A_p(\theta))\rho_p(d\omega),
\]  
(15)

Full exponential families are also “regular” [2, page 2]. Define \(\Omega_p\) to be the support of \(\rho_p\), the minimal closed set \(\mathcal{W} \subseteq \mathbf{R}\) such that \(\rho_p(\mathcal{W} - \mathcal{W}) = 0\), i.e. the possible outcomes. If \(p\) is full and \(|\Omega_p| > 1\), it is “minimal” [2, page 2]. A full, minimal, 1-dimensional standard exponential family \(p\) has a strictly convex \(A_p\) [2, Theorem 1.13] that is infinitely differentiable everywhere [2, Theorem 2.2].

The likelihood of \(\omega\) given \(\theta\) is \(\exp(\theta \omega - A_p(\theta))\). The negative log likelihood of \(\omega\) given \(\theta\) is \(\ell_p(\omega, \theta) = A_p(\theta) - \theta \omega\). Notice that if \(A_p\) is strictly convex, then \(\ell_p\) is strictly convex in its second argument. The mean \(\mu_p : \mathbf{R} \to \mathbf{R}\) is:

\[
\mu_p(\theta) = \int \exp(\theta \omega - A_p(\theta))\rho_p(d\omega).
\]  
(16)

It is useful to define \(\lambda_p : \mathbf{R} \to \mathbf{R}\) to be:

\[
\lambda_p(\theta) = \int \exp(\theta \omega)\rho_p(d\omega).
\]  
(17)

Now given these standard definitions, we can prove Lemma 1.

**Lemma 1** The Bernoulli family, Gaussian family (fixed variance), and Poisson family are full, minimal, 1-dimensional standard exponential families. For a full, minimal, 1-dimensional standard exponential family \(p\):

1. for every \(\theta \in \mathbf{R}, \omega \in \Omega_p\), \(\ell_p(\omega, \theta)\) is differentiable with respect to \(\theta\), and \(\frac{\partial \ell_p(\omega, \theta)}{\partial \theta} = \mu_p(\theta) - \omega\); and
2. \(\ell_p\) is strictly convex in its second argument: for every \(\theta, \theta' \in \mathbf{R}\), if \(\theta \neq \theta'\), then for all \(\lambda \in (0, 1)\) we have \(\ell_p(\omega, \lambda \theta + (1 - \lambda)\theta') < \lambda \ell_p(\omega, \theta) + (1 - \lambda)\ell_p(\omega, \theta')\).

**Proof:** As we stated before, \(A_p\) is strictly convex if \(p\) is a full, minimal, 1-dimensional standard exponential family, and this implies that \(\ell_p\) is strictly convex in its second argument. If \(p\) is a standard exponential family that is full and minimal, then \(\lambda_p\) is infinitely differentiable everywhere [2, Theorem 2.2], and by [2, page 34]:

\[
\lambda_p'(\theta) = \int \omega \exp(\theta \omega)\rho_p(d\omega)
\]  
(18)

We then normalize:

\[
\frac{\lambda_p'(\theta)}{\lambda_p(\theta)} = \frac{\int \omega \exp(\theta \omega)\rho_p(d\omega)}{\lambda_p(\theta)} = \frac{\int \omega \exp(\theta \omega)\rho_p(d\omega)}{\exp(A_p(\theta))}
\]  
(19)

\[
= \frac{\int \omega \exp(\theta \omega - A_p(\theta))\rho_p(d\omega)}{\exp(A_p(\theta))}
\]  
(20)

\[
= \mu_p(\theta).
\]  
(21)

So, for \(\ell_p(\omega, \theta)\):

\[
\frac{\partial \ell_p(\omega, \theta)}{\partial \theta} = A_p'(\theta) - \omega = \lambda_p'(\theta) - \omega = \mu_p(\theta) - \omega.
\]  
(23)

Equation 25 is [16, Equation 3].

We now just have to show that \(|\Omega_p| > 1\) and \(A_p\) is finite everywhere for the families mentioned. It is natural to think of the definition of \(\rho_p\) in terms of the possible outcomes, but \(\Omega_p\) is defined in terms of \(\rho_p\). So, instead we define \(\mathcal{W}\) as a set (that will turn out to be the possible outcomes), define \(\rho_p\) in terms of \(\mathcal{W}\), and then show \(\Omega_p = \mathcal{W}\). Note that if \(\mathcal{W}\) is finite, to prove \(\Omega_p = \mathcal{W}\), we need only prove that for all \(\omega \in \mathcal{W}\), \(\rho_p(\omega) > 0\), and \(\rho_p(\mathcal{W} - \mathcal{W}) = 0\). If \(\mathcal{W} = \mathbf{R}\), then we need to show any finite, nonempty, open interval has positive measure to prove \(\Omega_p = \mathcal{W}\) (this is because \(\mathbf{R} - \Omega_p\) is open, and if there is some \(\omega \in \mathbf{R} - \Omega_p\), then there must be a neighborhood \(N\) of \(\omega\) (a finite, nonempty, open interval) in \(\mathbf{R} - \Omega_p\) where \(\rho_p(N) = 0\).
1. For the Bernoulli family of distributions, $W = \{0, 1\}$ and $\rho_p(B) = |B \cap \mathcal{W}|$. Thus, $\rho_p(\{1\}) = 1$, and $\rho_p(\mathcal{R} - W) = |(\mathcal{R} - W) \cap \mathcal{W}| = 0$, so $\Omega_p = \mathcal{W}$. Also, $A_p(\theta) = -\ln(1 + \exp(\theta))$, $Pr[|\theta| = \exp(\theta)]$, $\ell_p(\omega, \theta) = -\theta \omega + \ln(1 + \exp(\theta))$, and $\mu_p(\theta) = \frac{1}{1 + \exp(-\theta)}$. Since $A_p(\theta)$ is finite everywhere, the Bernoulli family is full, and since $|\Omega_p| > 1$, the Bernoulli family is minimal.

2. For the Gaussian family of distributions with fixed mean $\sigma^2 > 1$ we have $\mathcal{W} = \mathbb{R}$, but we also need to define a particular $\rho_p$. Specifically, define some $h(\omega) = \frac{\sigma^2}{2} \exp(-\frac{\omega^2}{2})$. For any Borel measurable set, define $\rho_p(B) = \int_B h(\omega)d\omega$, where the right hand side is the standard Lebesgue integral. Note $\rho_p$ itself is a Gaussian with mean zero and variance one. Since $h(\omega) > 0$ and $h(\omega)$ is continuous, on any finite, closed interval $[a, b]$ it has a minimum $m > 0$, and therefore on any finite, nonempty, open interval $(a, b)$, $\rho_p((a, b)) \geq (b - a)m$, so $\Omega_p = \mathcal{W}$. Thus, $A_p(\theta) = \frac{\sigma^2}{2} \int_B \exp(-\frac{\omega^2}{2})d\omega$, $\mu_p(\theta) = \theta$, and the likelihood is $\exp(\theta \omega - \frac{\sigma^2}{2})$, and $\ell_p(\omega, \theta) = \frac{\sigma^2}{2} - \theta^2 = \frac{1}{4}\theta^2 - \frac{\sigma^2}{2}$. Notice that this loss is off by $\frac{\sigma^2}{2}$ from squared loss, and this term is independent of $\theta$. Finally, $|\Omega_p| > 1$, $A_p$ is finite everywhere, implying $p$ is a full, minimal, standard, 1-dimensional family.

3. For the Poisson family of distributions $\mathcal{W} = \{0, 1, 2, \ldots\}$. Define $\rho_p(B) = \sum_{\omega \in B \cap \mathcal{W}} \frac{1}{\omega!}$, so $A_p(\theta) = \exp(\theta)$. Moreover, for any non-negative integer $\omega$, $\rho_p(\{\omega\}) = \frac{1}{\omega!}$, and $\rho_p(\mathcal{R} - \mathcal{W})$ is the sum over an empty set, and therefore zero. Finally, $Pr[\omega|\theta] = \frac{1}{\omega!}\exp(\theta \omega - \exp(\theta))$, and the likelihood is $\exp(\theta \omega - \exp(\theta))$, $\mu_p(\theta) = \exp(\theta)$ and $\ell_p(\omega, \theta) = \exp(\theta) - \theta \omega$. Again, $|\Omega_p| > 1$, and $A_p$ is finite everywhere.

B RELATION TO TRADITIONAL LAYERED NEURAL NETWORKS

A more conventional way to represent a network involves writing layers, by interleaving fixed activation functions with learned affine functions. We imagine a sequence of integers $n_0, \ldots, n_k$, representing the number of nodes in each layer, with 0 being the input layer (with $X = \mathbb{R}^{n_0}$), and $k$ being the output layer (with $n_k = 1$). We choose an activation function $a_i : \mathbb{R} \to \mathbb{R}$ for each layer $\{0, \ldots, k - 1\}$, and define $A^i : \mathbb{R}^{n_i} \to \mathbb{R}^{n_i}$ such that $(A^i(v))_j = a_i(v_j)$ for all $v \in \mathbb{R}^{n_i}$, for all $j \in \{1, \ldots, n_i\}$.

Then, in-between these layers, we have a matrix $W^i \in \mathbb{R}^{n_i \times n_{i-1}}$ and a vector $w^i \in \mathbb{R}^{n_i}$. We define $h^0 : \mathbb{R}^{n_0} \to \mathbb{R}^{n_0}$ to be the identity. We can then define $h^i : \mathbb{R}^{n_i} \to \mathbb{R}^{n_i}$ for $i \in \{1, \ldots k\}$ recursively such that $h^i(x) = W^i A^{i-1}(h^{i-1}(h^i(x)) + w^i)$. The model in this form is $M_{W, w}(x) = h^k(x)$. As we will show, these can be easily modeled in our graph representation. However, we will see that learning affine (as opposed to linear) functions show that the conventional concept of “layer” is not as crisp and neat as one would like.

B.1 REPRESENTATION AS A GRAPH

First of all, this kind of model can be represented in the graph network that we describe in Section 5. Most oddities of the representation come from the bias features.

1. For each $i \in \{1, \ldots, k\}$, define $V_i = \{v_{i,1}, \ldots, v_{i,n_i}\}$, and let $v_i$ be a special vertex (which will be a special input node that will always equal 1).
2. Define $V = \{v_c\} \cup \bigcup_{i=0}^k V_i$.
3. Define $I = V_0 \cup \{v_c\}$.
4. Define $o^c = v_{k,1}$.
5. For $i \in \{0, \ldots, k - 1\}$, for all $j \in \{1, \ldots, n_i\}$, let $a_{v_{i,j}} = a^c$.
6. Define $a_{o^c}$ and $a_{v_c}$ to be the identity.
7. For all $x \in X$, for all $j \in 1, \ldots, n_0$, let $g_{v_{0,j}}(x) = x_j$.
8. For all $x \in X$, $g_{v_{0,j}}(x) = 1$.
9. The edges associated with $W^i$ are $E_i = V_{i-1} \times V_i$.
10. The edges associated with $w^i$ are $E_i^c = \{v_c\} \times V_i$.
11. Define $E = \bigcup_{i=1}^k E_i \cup E_i^c$.

$D = (V, E, I, \{g_i\}_{i \in T}, o^c, \{a_{v_e}\}_{v \in V})$ is a neural network, equivalent to the layered form we described in the previous section.

In the next section, we will show how the edges in $E_i^c$ (associated with the bias parameters) play a unique role in our work.

B.2 AN ORDERED CUT IN A LAYERED NEURAL NETWORK

We introduced ordered cuts and cut sets in Section 6. The most obvious cut would be $B_i = \{v_c\} \cup \bigcup_{j=0}^{i-1} V_j$ and

---

5Note the distinction here between the conventional density defined with respect to the Lebesgue measure, and this likelihood defined with respect to $\rho$. However, since we are tuning $\theta$ and $\omega$ is given, this is simply a constant in $\ell_p$.

6As before, there is a slight distinction between the conventional probability mass and the likelihood as defined here.
one parameter, and the generated (holographic) feature is al-

to ask, what does a generated feature look like? How does it relate to a conventional activation feature? Let us break this down into features generated from weight matrices $W$ (or $E_1 \ldots E_k$), and features generated from bias vectors $w$ (or $E'_1 \ldots E'_k$). In this section, for simplicity we assume that partial derivatives exist where necessary: see Appendix II for a deeper discussion of differentiabil-

ity in deep networks.

We consider the bias features first for some layer $i \in \{1 \ldots k\}$. For any $j \geq i$, we can recursively define $h^{j,i} : \mathbb{R}^{n_i} \rightarrow \mathbb{R}^{n_j}$ such that $h^{j,i}(v) = v$ and for all $j \in \{i+1 \ldots k\}$, $h^{j,i}(v) = W^j A^{i-1}(h^{j-1,i}(v)) + w^j$. This is an accumulation of the transforms after the input of layer $i$, and for all $x \in X$:

$$M_{W,w}(x) = h^{k,i}(h^i(x))$$

$$M_{W,w}(x) = h^{k,i}(W^i A^{i-1}(h^{i-1}(x)) + w^i).$$

For some $q \in \{1 \ldots n_i\}$, if we define $f_q^i : X \rightarrow \mathbb{R}$ to be the feature generated from $(v_c, v_{i,q})$, then we can write:

$$f_q^i(x) = \frac{\partial M_{W,w}(x)}{\partial w^i_q} = \frac{\partial h^{k,i}(v)}{\partial v_q} \bigg|_{v=W^i A^{i-1}(h^{i-1}(x)) + w^i}.$$ (29)

Notice that this is the partial derivative of the prediction with respect to the input of node $v_{i,q}$.

Next, we consider features generated from a weight matrix. Consider some $i \in \{1 \ldots k\}$, some $p \in \{1 \ldots n_{i-1}\}$ and some $q \in \{1 \ldots n_i\}$. Then $(v_{i-1,p}, v_{i,q}) \in E_i$ is the edge related to parameter $W_{i,q,p}$. Define $F^i_{q,p} : X \rightarrow \mathbb{R}$ to be the feature generated from $(v_{i-1,p}, v_{i,q})$. Using $h^{k,i}$ again:

$$F^i_{q,p}(x) = \frac{\partial M_{W,w}(x)}{\partial W^i_{q,p}} = \frac{\partial h^{k,i}(v)}{\partial v_q} \bigg|_{v=W^i A^{i-1}(h^{i-1}(x)) + w^i} A_{i-1}^{-1}(h^{i-1}(x))$$

$$= f_{q,p}^i(x) A_{i-1}^{-1}(h^{i-1}(x)).$$ (32)

So, the generated feature from $(v_{i-1,p}, v_{i,q})$ is the activation feature of $v_{i-1,p}$ times the generated feature of $(v_c, v_{i,q})$.

In summary, the generated features of conventional layers have a nice form, and one can take a conventional layered network and translate it into a graph. However,
from a mathematical perspective, it is much easier to reason about graphs, because of the clean concept of cuts and cut sets. Moreover, the bias features work in highly counterintuitive ways, and presenting a theory without them tells an incomplete story.

C CONVEXITY

In this section, we will state some known results about convexity.

Fact 27 Given a supervised learning problem \( \mathcal{P} = (p, X, \{x_1, \ldots, x_m\}, \{y_1, \ldots, y_m\}) \), given two models \( M : X \to \mathbb{R} \) and \( M' : X \to \mathbb{R} \) that are equal on the training data, if \( M \) is calibrated on a feature \( f : X \to \mathbb{R} \), then \( M' \) is calibrated on \( f \).

One can think about the model as mapping a matrix of inputs to a vector of predictions, which is then composed with a loss function that maps a vector of predictions to a single loss. Next, we show when this second mapping will be (strictly) convex.

Lemma 28 Given convex sets \( C_1, \ldots, C_m \subseteq \mathbb{R} \), for each \( i \in \{1, \ldots, m\} \) a function \( L_i : C_i \to \mathbb{R} \), then if \( C = \times_{i=1}^m C_i \), and there is a function \( L : C \to \mathbb{R} \) such that for all \( x \in C \):

\[
L(x) = \sum_{i=1}^m L_i(x_i),
\]

then:

1. if for all \( i \), \( L_i \) is convex, then \( L \) is convex.
2. if for all \( i \), \( L_i \) is strictly convex, then \( L \) is strictly convex.

Proof: Consider \( x, y \in C \), and \( \lambda \in [0, 1] \). Without loss of generality, assume \( x \neq y \), and \( \lambda \in (0, 1) \). By convexity, for all \( i \), \( \lambda L_i(x_i) + (1 - \lambda) L_i(y_i) \geq \lambda L_i(x_i) + (1 - \lambda) y_i \), so:

\[
L(\lambda x + (1 - \lambda)y) = \sum_{i=1}^m L_i(\lambda x_i + (1 - \lambda)y_i) \leq \sum_{i=1}^m (\lambda L_i(x_i) + (1 - \lambda) L_i(y_i)) \leq \lambda \sum_{i=1}^m L_i(x_i) + (1 - \lambda) \sum_{i=1}^m L(y_i) \leq \lambda L(x) + (1 - \lambda)L(y).
\]

D PROOFS OF CALIBRATION ON GENERATED FEATURES

Next we show that if the partial derivative of the loss with respect to a parameter is zero, then the feature generated by that parameter is calibrated.

Theorem 7 For a problem \( \mathcal{P} \) and model family \( \mathcal{M} = \{M_w\}_{w \in \mathbb{R}^d} \), given a \( w \in \mathbb{R}^S \) and \( s \in S \) such that \( f_s \) is the feature generated from parameter \( s \) of model \( M_w \): if \( \{s\} \) is total on the training data given \( M_w \) and \( \frac{\partial L_P(M_w)}{\partial w_s} = 0 \), then \( M_w \) is calibrated with respect to \( f_s \).

Proof: Define \( p, m, X, x_1, \ldots, x_m, y_1, \ldots, y_m \) such that \( \mathcal{P} = (p, X, \{x_1, \ldots, x_m\}, \{y_1, \ldots, y_m\}) \). We begin with the partial derivative of \( L_P \). Note that from Lemma 11 \( L_p \) is partially differentiable with respect to its second argument, and since for any \( i \in \{1, \ldots, m\} \) \( \{s\} \) is total for \( x_i \) given \( M_w \), then \( M_w(x_i) \) is partially
differentiable with respect to \( w_s \). Therefore:
\[
0 = \frac{\partial L_P(M_w)}{\partial w_s} = \frac{\partial}{\partial w_s} m \ell_p(y_i, M_w(x_i)) = \sum_{i=1}^{m} \frac{\partial \ell_p(y_i, M_w(x_i))}{\partial w_s} = \sum_{i=1}^{m} \frac{\partial \ell_p(y_i, \hat{y}_i)}{\partial \hat{y}_i} \left| \frac{\partial M_w(x_i)}{\partial \hat{y}_i} \right| = \sum_{i=1}^{m} \left( \mu_p(M_w(x_i)) - y_i \right) \frac{\partial M_w(x_i)}{\partial w_s}.
\]
\[
0 = \sum_{i=1}^{m} \left( \mu_p(M_w(x_i)) - y_i \right) \frac{\partial M_w(x_i)}{\partial w_s}.
\]

The last step is because of Lemma 1. For any \( i \in \{1 \ldots m\} \), since \( M_w(x_i) \) is partially differentiable with respect to \( w_s \), we can use \( f_s(x_i) = \frac{\partial M_w(x_i)}{\partial w_s} \) to get:
\[
0 = \sum_{i=1}^{m} \left( \mu_p(M_w(x_i)) - y_i \right) f_s(x_i).
\]
\[
\sum_{i=1}^{m} y_i f_s(x_i) = \sum_{i=1}^{m} \mu_p(M_w(x_i)) f_s(x_i).
\]

\[\text{Lemma 9} \] For finite \( S \), given a set of features \( \{f_s\}_{s \in S} \), the family of linear models \( \mathcal{L}(\{f_s\}_{s \in S}) \), and \( N_w \in \mathcal{L} \); the set \( S \) is total for all \( x \in X \) given \( N_w \) and the feature generated from parameter \( s \) by model \( N_w \) is \( f_s \).

\[\text{Proof:} \] First note that, by definition, a linear model is a linear function of \( w \), and therefore a differentiable function of \( w \) regardless of the input or \( w \). Thus, for any \( x \in X \), for any \( w \), the set \( S \) is total. Notice that, for any \( x \in X \):
\[
\frac{\partial M_w(x)}{\partial w_s} = \frac{\partial}{\partial w_s} \sum_{t \in S} w_t f_t(x) = \sum_{t \in S} \frac{\partial}{\partial w_s} (w_t f_t(x)) = f_s(x).
\]

Another known key result about linear models is that any two linear models that minimize loss will produce the same predictions on the training data.

\[\text{Theorem 29} \] Given a supervised learning problem \( \mathcal{P} = (p, X, \{x_1 \ldots x_m\}, \{y_1 \ldots y_m\}) \), a set of features \( f_1 \ldots f_n : X \to \mathbb{R} \), and a family of linear models \( \mathcal{L}(\{f_s\}_{s \in S}) \), if two models \( M, M' \in \mathcal{L} \) are optimal in \( \mathcal{L} \) for \( \mathcal{P} \), then they are equal on the training data.

\[\text{Proof:} \] Define \( L^* : \mathbb{R}^m \to \mathbb{R} \) such that for all \( \hat{y} \in \mathbb{R}^m \):
\[
L^*(\hat{y}) = \sum_{i=1}^{m} \ell_p(y_i, \hat{y}_i).
\]

By Lemma 1 each \( \ell_p \) is strictly convex in its second argument. By Lemma 28, \( L^* \) is strictly convex. We define \( P : \mathbb{R}^n \to \mathbb{R}^m \), such that for all \( v \in \mathbb{R}^n \), for all \( i \in \{1 \ldots m\} \), \( P_i(v) = M_v(x_i) \). Therefore, \( L^*(P(v)) = L_P(M_v) \). We need to prove \( P(w) = P(w') \).

Consider \( w'' = \frac{w + w'}{2} \). Since the models are linear, we know that for all \( x \in X \), \( M_{w''}(x) = \frac{M_w(x) + M_{w'}(x)}{2} \). Thus, \( P(w'') = \frac{P(w) + P(w')}{2} \). Assume for the sake of contradiction, \( P(w) \neq P(w') \). Since \( L^* \) is strictly convex:
\[
L^*(P(w'')) < \frac{L^*(P(w)) + L^*(P(w'))}{2};
\]
\[
\text{since } L^*(P(v)) = L_P(M_v); \text{ we get:}
\]
\[
L_P(M_{w''}) < \frac{L_P(M_w) + L_P(M_{w'})}{2},
\]
\[
L_P(M_{w''}) < L_P(M_w).
\]

This contradicts the original hypothesis that both are minima. Thus, \( P(w) = P(w') \), which is the same as saying, for all \( i \in \{1 \ldots m\} \), \( M_w(x_i) = M_{w'}(x_i) \).

\[\text{Lemma 10 (c.f. [16 Equation 10])} \] Given a problem \( \mathcal{P} \), a finite set \( S \) and a set of features \( \{f_s\}_{s \in S} \); a model \( N \) in the family of linear models \( \mathcal{L}(\{f_s\}_{s \in S}) \) is optimal if and only if \( N \) is calibrated with respect to \( f_s \) for all \( s \in S \).

\[\text{Proof:} \] Define \( p, m, X, x_1 \ldots x_m, y_1 \ldots y_m \) such that \( \mathcal{P} = (p, X, \{x_1 \ldots x_m\}, \{y_1 \ldots y_m\}) \). Define \( \{N_w\}_{w \in \mathbb{R}^S} = \mathcal{L}(\{f_s\}_{s \in S}) \) where \( N_w(x) = \sum_{s \in S} w_s f_s(x) \). Define \( w^* \in \mathbb{R}^S \) such that \( N_{w^*} = N \). If \( N \) (and therefore \( N_{w^*} \)) is calibrated, for all \( s \in S \):
\[
\sum_{i=1}^{m} \mu_p(N_{w^*}(x_i)) f_s(x_i) = \sum_{i=1}^{m} y_i f_s(x_i)
\]
\[
\sum_{i=1}^{m} \mu_p(N_{w^*}(x_i)) f_s(x_i) = \sum_{i=1}^{m} y_i f_s(x_i).
\]

\[\text{Formally, for all } i \in \{1 \ldots m\}, \text{ we could define } L_i : \mathbb{R} \to \mathbb{R} \text{ such that } L_i(y_i) = \ell_p(y_i, \hat{y}_i).\]
By Lemma 11:

\[ 0 = \sum_{i=1}^{m} \frac{\partial L_P(y_i, \hat{y}_i)}{\partial \hat{y}_i} \bigg|_{\hat{y}_i = N_{w^*}(x_i)} f_s(x_i) \quad (60) \]

\[ 0 = \sum_{i=1}^{m} \frac{\partial L_P(y_i, N_{w^*}(x_i))}{\partial w_s^*} \quad (61) \]

\[ 0 = \frac{\partial L_P(N_{w^*})}{\partial w_s^*}. \quad (62) \]

If we define \( L^*: \mathbb{R}^n \to \mathbb{R} \) as \( L^*(w) = L_P(N_w) \), it is convex, then \( N_{w^*} \) (and therefore \( N \)) is optimal.

To prove the converse, suppose that there is some \( s \in S \) that is not calibrated. Then \( \frac{\partial L_P(N_{w^*})}{\partial w_s^*} \neq 0 \), implying that there is a model with lower loss.

**Lemma 11** Given a problem \( \mathcal{P} \), a finite set \( S \), a subset \( S' \subseteq S \), and a set of features \( \{f_s\}_{s \in S} \): if a model \( N \) is optimal in the family of linear models \( \mathcal{L} \) and \( N \) is calibrated with respect to \( f_s \) for all \( s \in S - S' \), then \( N \) is also an optimal model in \( \mathcal{L} \).

**Proof:** Define \( p, m, X, x_1 \ldots x_m, y_1 \ldots y_m \) such that \( \mathcal{P} = \{(p, X, \{x_1 \ldots x_m\}, \{y_1 \ldots y_m\}\} \). Note that any model in \( \mathcal{L} \) (and specifically \( N \)) is in \( \mathcal{L} \). Since \( N \) is optimal in \( \mathcal{L} \), by Lemma 10 it is calibrated with respect to \( \{f_s\}_{s \in S'} \). Moreover, by assumption it is calibrated with respect to \( \{f_s\}_{s \in S - S'} \), and therefore it is calibrated with respect to \( \{f_s\}_{s \in S} \), and again by Lemma 10 it is optimal with respect to \( \mathcal{L} \).

**Lemma 13** Given a problem \( \mathcal{P} \), a finite set \( S \), a subset \( S' \subseteq S \), a model family \( \{M_w\}_{w \in \mathbb{R}^S} \), and \( w \in \mathbb{R}^S \) such that \( \frac{\partial L_P(M_w)}{\partial w} = 0 \) for all \( s \in S' \): if \( \mathcal{L}' = \mathcal{L} \) is the family of linear models associated with \( S' \) given \( M_w \), \( S' \) is total on the training data given \( M_w \), and \( M_w \) is equal on the training data to some \( N' \subseteq \mathcal{L}' \), then \( N' \) is optimal in \( \mathcal{L}' \).

**Proof:** Define \( p, m, X, x_1 \ldots x_m, y_1 \ldots y_m \) such that \( \mathcal{P} = \{(p, X, \{x_1 \ldots x_m\}, \{y_1 \ldots y_m\}\} \). For all \( s \in S' \), define \( f_s : X \to \mathbb{R} \) to be the feature generated by \( s \) given \( M_w \). By Theorem 7, \( M_w \) is calibrated with respect to \( f_s \). Moreover, the family of linear models associated with \( S' \) given \( M_w \) is \( \mathcal{L}' = \mathcal{L} \). Since \( N' \) and \( M_w \) are equal on the training data, by Fact 7 \( N' \) is calibrated with respect to all \( \{f_s\}_{s \in S'} \). Thus, by Lemma 10 \( N' \) is optimal in \( \mathcal{L}' \).

**Lemma 14** Given a problem \( \mathcal{P} \), a finite set \( S \), a subset \( S' \subseteq S \), a model family \( \{M_w\}_{w \in \mathbb{R}^S} \), and \( w \in \mathbb{R}^S \) such that \( \frac{\partial L_P(M_w)}{\partial w} = 0 \) for all \( s \in S \): if \( \mathcal{L}' \) is the family of linear models associated with \( S' \) given \( M_w \), \( S' \) is total, \( M_w \) is equal on the training data to some \( N' \subseteq \mathcal{L}' \), and \( \mathcal{L}' \) is the family of linear models associated with \( S \), then any optimal \( N'' \subseteq \mathcal{L}' \) equals \( M_w \) on the training data.

**Proof:** For all \( s \in S \), define \( f_s : X \to \mathbb{R} \) to be the feature generated by \( s \) given \( M_w \). By Theorem 7, \( M_w \) is calibrated with respect to \( f_s \), and by Fact 7 \( N'' \) is calibrated with respect to \( f_s \). By Lemma 13 \( N'' \) is optimal in \( \mathcal{L}' \). By Lemma 11 \( N'' \) is optimal in \( \mathcal{L}' \). Thus, by Theorem 29, any optimal \( N'' \subseteq \mathcal{L} \) is equal on the training data to \( N' \), and transitive to \( M_w \).

**E Euler’s Homogeneous Function Theorem**

This appendix can be skipped, as Lemma 17 is a well-known result. However, this section provides a deeper discussion of total differentiability and partial differentiability that can be helpful in understanding the rest of the paper.

Although a variant of Lemma 17 is in [11], we prove the specific variant here.

**Lemma 17 (Euler’s Homogeneous Function Theorem)**

**Degree 1 Case** If a homogeneous function \( f : \mathbb{R}^p \to \mathbb{R}^q \) is differentiable at \( x \in \mathbb{R}^p \), then for all \( k \in \{1 \ldots q\} \),

\[ f_k(x) = \sum_{j=1}^{p} \frac{\partial f_k(x)}{\partial x_j} x_j. \quad (9) \]

**Proof:** If \( x = 0 \), then for any \( k \in \{1 \ldots q\} \):

\[ \sum_{j=1}^{p} \frac{\partial f_k(x)}{\partial x_j} x_j = \sum_{j=1}^{p} \frac{\partial f_k(0)}{\partial x_j} x_j = 0 \times f_k(x) = 0 \times f_k(0x) = f_k(0x). \]

For the remainder, assume \( x \neq 0 \). Consider a particular \( x \) where \( f(x) \) is differentiable, and the derivative is a matrix \( G \in \mathbb{R}^{q \times p} \) where:

\[ \lim_{h \to 0} \frac{\|f(x + h) - f(x) - Gh\|}{\|h\|} = 0. \quad (68) \]

Specifically, we can write:

\[ \lim_{\epsilon \to 0} \frac{\|f(x + \epsilon x) - f(x) - G\epsilon x\|}{\|\epsilon x\|} = 0. \quad (69) \]
In the limit, $\epsilon > -1$, so by the homogeneous property:

$$\lim_{\epsilon \to 0} \frac{\| (1 + \epsilon) f(x) - f(x) - G \epsilon x \|}{\| \epsilon x \|} = 0$$  \hspace{1cm} (70)
$$\lim_{\epsilon \to 0} \frac{\| f(x) - G \epsilon x \|}{\| \epsilon x \|} = 0$$  \hspace{1cm} (71)
$$\lim_{\epsilon \to 0} \frac{\| f(x) - G \epsilon x \|}{\| x \|} = 0$$  \hspace{1cm} (72)
$$\| f(x) - G \epsilon x \| = 0$$  \hspace{1cm} (73)
$$\| f(x) - G x \| = 0$$  \hspace{1cm} (74)
$$f(x) = G x.$$  \hspace{1cm} (75)

Since $G_{k,j} = \frac{\partial f_k(x)}{\partial x_j}$, the result follows.

One might wonder, is it possible to prove this for functions that are only partially (and not totally) differentiable? Yes and no: for one or two dimensions partially differentiability is sufficient for Euler’s function to hold, but for three or more dimensions, there is no such guarantee, and we can demonstrate this with a counterexample.

**Lemma 30** If a homogeneous function $f : \mathbb{R}^p \to \mathbb{R}^q$ is partially differentiable at a point $x \in \mathbb{R}^p$, and $p \leq 2$, then for all $k \in \{1 \ldots q\}$:

$$f_k(x) = \sum_{j=1}^{p} \frac{\partial f_k(x)}{\partial x_j} x_j.$$  \hspace{1cm} (76)

**Proof:** Since partial differentiability implies differentiability when $p = 1$, the proof for $p = 1$ follows directly from Lemma 17. So, we assume $p = 2$. Consider a specific point $(x, y)$ where $f$ is partially differentiable, i.e. $\frac{\partial f(x,y)}{\partial x}$ and $\frac{\partial f(x,y)}{\partial y}$ exist. If $x = 0$ or $y = 0$, then it is basically analogous to the one dimensional case, as the function along the axis can be considered a 1 dimensional homogeneous function that is differentiable at that point.

Next, assume $x > 0$ and $y > 0$, i.e. a point in the positive quadrant as in Figure 6 near the end of the proof, we will show how to reduce a problem in a different quadrant to one in the positive quadrant.

We can define $g : (0, x) \to (0, \infty)$ and $h : (0, x) \to (0, \infty)$ such that $g(\epsilon) = \frac{y}{x - \epsilon}$ and $h(\epsilon) = \frac{y + g(\epsilon)}{y}$.

Then, for any $\epsilon \in (0, x)$, observe that $h(\epsilon)(x - \epsilon, y) = (x, y + g(\epsilon))$, so $\lim_{\epsilon \to 0^+} h(\epsilon) = f(x, y + g(\epsilon))$. Since

In Figure 6: For homogeneous functions with a 2-dimensional domain, the partial derivatives are inextricably linked. Note that the definition of $g(\epsilon)$ can be derived from $\frac{g(\epsilon) + y}{x} = \frac{g(\epsilon)}{y}$. The function values at $(x - \epsilon, y)$ and $(x, y + g(\epsilon))$ are connected through the homogeneous property.

$$\lim_{\epsilon \to 0^+} g(\epsilon) = 0,$$ we know that:

$$\frac{\partial f(x, y)}{\partial y} = \lim_{\epsilon \to 0^+} \frac{f(x, y + g(\epsilon)) - f(x, y)}{\frac{g(\epsilon)}{y}}$$  \hspace{1cm} (77)
$$\frac{\partial f(x, y)}{\partial y} = \lim_{\epsilon \to 0^+} \frac{f(x - \epsilon, y)h(\epsilon) - f(x, y)}{\frac{g(\epsilon)}{y}}$$  \hspace{1cm} (78)
$$\frac{\partial f(x, y)}{\partial y} = \lim_{\epsilon \to 0^+} \frac{f(x - \epsilon, y)h(\epsilon) - f(x, y)h(\epsilon)}{\frac{g(\epsilon)}{y}} + \lim_{\epsilon \to 0^+} \frac{f(x, y)h(\epsilon) - f(x, y)}{\frac{g(\epsilon)}{y}}$$  \hspace{1cm} (79)
$$\frac{\partial f(x, y)}{\partial y} = \lim_{\epsilon \to 0^+} \frac{f(x - \epsilon, y) - f(x, y)}{\epsilon} \lim_{\epsilon \to 0^+} \frac{h(\epsilon)\epsilon}{g(\epsilon)}$$  \hspace{1cm} (80)

First, observe that $\lim_{\epsilon \to 0^+} f(x - \epsilon, y) - f(x, y) = \frac{\partial f(x, y)}{\partial x}$.

Notice that for $\epsilon \in (0, x)$:

$$\frac{h(\epsilon) - 1}{g(\epsilon)} = \frac{\frac{y + g(\epsilon)}{y} - 1}{g(\epsilon)} = \frac{1}{y}.$$  \hspace{1cm} (81)
Also, since \( \lim_{\epsilon \to 0^+} h(\epsilon) = 1 \), for \( \epsilon \in (0, x) \):

\[
\lim_{\epsilon \to 0^+} \frac{h(\epsilon)}{g(\epsilon)} = \lim_{\epsilon \to 0^+} \frac{\epsilon}{g(\epsilon)} = \frac{\epsilon(x - \epsilon)}{ye} = \frac{x}{y}.
\]

(83)

So, from Equation 80

\[
\frac{\partial f(x, y)}{\partial y} = f(x, y),
\]

(85)

which is Euler’s function.

Now if \( x < 0 \) or \( y < 0 \), we could flip the function on one axis or on both without affecting homogeneity, partial differentiability, or Euler’s function. Specifically, note that if we defined \( f^* : \mathbb{R}^2 \to \mathbb{R} \) such that for all \( x', y' \in \mathbb{R} \), \( f^*(x', y') = f(-x', y') \), then

\[
\frac{\partial f^*(-x, y)}{\partial x} = -\frac{\partial f(x, y)}{\partial x}, \quad \frac{\partial f^*(-x, y)}{\partial y} = \frac{\partial f(x, y)}{\partial y},
\]

and if \( f^*(-x, y) = \frac{\partial f^*(-x, y)}{\partial x}(-x) + \frac{\partial f^*(-x, y)}{\partial x}y \) then:

\[
f(x, y) = f^*(-x, y).
\]

(87)

\[
= \frac{\partial f^*(-x, y)}{\partial x}(-x) + \frac{\partial f^*(-x, y)}{\partial x}y
\]

(88)

\[
= \frac{\partial f^*(-x, y)}{\partial x} x + \frac{\partial f^*(-x, y)}{\partial x}y.
\]

(89)

Similarly for flipping on the \( x \) axis.

However, if we add a third dimension, things get incredibly complex. Consider the following function:

\[
f(x, y, z) = \begin{cases} 
7z - x & \text{if } x + y - 2z \geq 0 \text{ and } x - y < 0 \\
7z - y & \text{if } x - y \geq 0 \text{ and } x + y - 2z > 0 \\
x + 5z & \text{if } x + y - 2z \leq 0 \text{ and } x - y > 0 \\
x + 5z & \text{if } x - y \leq 0 \text{ and } x + y - 2z > 0 \\
y + 5z & \text{if } x - y \geq 0 \text{ and } x + y - 2z < 0 \\
y + 5z & \text{if } x = y = z
\end{cases}
\]

(91)

So, we wish to establish five things:

1. \( f \) is well-defined.
2. \( f \) is continuous.
3. \( f \) is homogeneous.
4. \( f \) has partial derivatives at \((1, 1, 1)\).

Figure 7: A visual representation of the eight regions of the piecewise linear homogeneous function \( f : \mathbb{R}^3 \to \mathbb{R} \). We are looking along the axis \( x = y = z \).

5. Euler’s formula does not hold at \((1, 1, 1)\).

To confirm \( f \) is well-defined, we must confirm that it is defined everywhere. If one considers Figure 7, one will notice that we have defined each region, starting from the top region and going clockwise, and then defined the center. Each region contains its counterclockwise edge, but not its clockwise edge, and none contain the center.

Thus, each point is in exactly one of the cases.

Next, we must establish continuity. First, we establish it at the center. If \( x = y = z \), then clearly \( 7z - x = 7z - y \), and \( y + 5z = x + 5z \). If \( x = y = z \), then \( x + y - 2z = 0 \), so \( 7z - x = 7z - x - (x + y - 2z) = x + 5z \).

Thus, all four linear functions equal each other along the line \( x = y = z \). Simple algebra shows us, along the plane \( x - y = 0 \), \( 7z - x = 7z - y \), because \( 7z - x = 7z - x + (x - y) = 7z - y \). Similarly, we can look at each boundary, and prove equality.

Notice that homogeneity is pretty straightforward. Since each constraint is in itself a linear inequality with no constants, given any \((x, y, z), (\lambda x, \lambda y, \lambda z)\) has the exact same constraints hold. Since within each region, the function is linear, then it is also homogeneous.

At the point \((1, 1, 1)\), notice that:

1. if you move in either direction along the \( x \) axis, \( f(x, y, z) = 7z - y \), so \( \frac{\partial f(x, 1, 1)}{\partial x} = 0 \).
2. if you move in either direction along the \( y \) axis, \( f(x, y, z) = 7z - x \), so \( \frac{\partial f(1, y, 1)}{\partial y} = 0 \).
3. and if you move in either direction along the z axis, you will stay where \( x - y = 0 \), so \( f(x, y, z) = 7z - x, \) so \( \frac{\partial f(1,1,z)}{\partial z} \bigg|_{z=1} = 7. \)

Note that \( f(1, 1, 1) = 6 \). However:

\[
\frac{\partial f(x, 1, 1)}{\partial x} \bigg|_{x=1} + \frac{\partial f(1, y, 1)}{\partial y} \bigg|_{y=1} + \frac{\partial f(1, 1, z)}{\partial z} \bigg|_{z=1} = 0 \times 1 + 0 \times 1 + 7 \times 1 = 7. \quad (92)
\]

Thus, since \( 6 \neq 7 \), this function, while homogeneous and continuous, has a point where it is partially differentiable, but Euler’s formula does not hold.

With further effort, we can also show that the function can be constructed using \textit{relu} gates. Thus, in a very crucial way, Euler’s function requires total differentiability, justifying the large role this concept of total features has in our theoretical analysis. However, as our empirical analysis shows, we can pretty much ignore this concern in practice. In Appendix [I], we give a sufficient condition (based on a complex proof) to determine if the model is total on an input.

F PROOFS FOR HOMOGENEOUS FUNCTIONS

Before continuing, it is important to note that when defining the homogeneous property of a model family, we directly appealed to a property of differentiability due to the technical issues described in higher dimensional homogenous functions in Appendix [E].

We now prove Theorem 20:

**Theorem 20** If a model family \( M \) is homogeneous on the parameter set \( S' \), \( M \in M \), \( L' \) is the family of linear models associated with \( S' \) given \( M \), and \( S' \) is total on the training data given \( M \), then there exists \( N \in L' \) such that \( M \) and \( N \) are equivalent on the training data.

**Proof:** Define \( M = \{ M_w \}_{w \in \mathbb{R}^s} \), and \( w \in \mathbb{R}^s \) such that \( M_w = M \). Given some \( x \) in the training data, since \( M \) is homogeneous with respect to \( S' \), and \( S' \) is total on \( x \) given \( M_w \), then by Lemma [17]

\[
M_w(x) = \sum_{s \in S'} \frac{\partial M_s(x)}{\partial w} w_s. \quad (93)
\]

Since \( f_s(x) = \frac{\partial M_w(x)}{\partial w_s} \) is the feature generated by \( s \) given \( M \), if \( x \) is in the training data, then \( \{s\} \subseteq S' \) is total on \( x \) given \( M_w \), and \( f_s(x) = \frac{\partial M_w(x)}{\partial w} \), so:

\[
M_w(x) = \sum_{s \in S'} w_s f_s(x). \quad (94)
\]

Note \( L' = L(\{ f_s \}_{s \in S'}) \), and if \( \{ N_{w'} \}_{w \in \mathbb{R}^s} = L' \), then for the \( w' \in \mathbb{R}^s \) where \( w'_s = w_s \) for all \( s \in S' \):

\[
N_{w'}(x) = \sum_{s \in S'} w'_s f_s(x) \quad (95)
\]

\[
= \sum_{s \in S'} w_s f_s(x) \quad (96)
\]

\[
= M_w(x), \quad (97)
\]

so \( M(x) = N_{w'}(x) \) on the training data, and \( N_{w'} \) is in the family of linear models associated with \( S' \) given \( M \).

We now prove Theorem 23:

**Theorem 23** Given a problem \( \mathcal{P} \) and a family of feedforward network models \( \mathcal{D}(V, E, I; \{ g_i \}_{i \in I}, o^*, A) = \{ M_w \}_{w \in \mathbb{R}^s} \), where all \( a \in A \) are homogeneous; if \( B \) and \( T \) is an ordered cut of the feedforward network, such that \( E' \) is the cut set, then:

1. \( \{ M_w \}_{w \in \mathbb{R}^s} \) is homogeneous on \( E' \);
2. for some \( w \in \mathcal{E}^F \), if \( E' \) is total on the training data given \( M_w \), \( L' \) is the family of linear models associated with \( E' \) given \( M_w \), and \( \frac{\partial L_M(M_w)}{\partial w} = 0 \) for all \( e \in E' \), then \( M_w \) is equal on the training data to any optimal model in \( L' \);
3. for some \( w \in \mathcal{E}^F \), if \( L' \) is the family of linear models associated with \( E \) given \( M_w \), if \( E \) is total on the training data given \( M_w \), and \( \frac{\partial L_M(M_w)}{\partial w} = 0 \) for all \( e \in E \), then \( M_w \) is equal on the training data to any optimal model in \( L' \).

**Proof:** We need only prove that the model family is homogeneous: the other two results follow from Theorem 20, Lemma 13, and Lemma 14. We need to construct a function from the inputs passing through the cut set to the second part of the function. First, define \( E'' = E - E' \), and given \( w \), define \( w'' \in \mathcal{E}^F \) such that for all \( e \in E'' \), \( w''_e = w_e \). Thus, for all \( t \in T \), define \( d_t, w'' : \mathcal{E}^F \rightarrow \mathbb{R} \) recursively (using the partial ordering of the vertices in the network) such that for all \( z \in \mathcal{E}^F \):

\[
d_t, w''(z) = \alpha_t \left( \sum_{u: (u, t) \in E'} z(u, t) + \sum_{u: (u, t) \in E''} d_{t, w''}(z) w''(u, t) \right). \quad (98)
\]

We subcribe this by \( w'' \) to indicate that \( d_t, w'' \) is independent of the weights in \( E' \). An important point to note is that for any \( x \in X \), for any \( b \in B \), \( c_b, w(x) \) does not depend upon the weight of any edge in \( E' \). For any
Lemma 14, the other results hold.

We have established that $M$.

Recall the main theorem about ResNets and CNNs: $s$.

We can use this formulation to prove $M(x)$ is homogeneous in $w'$. First, we can prove recursively that $d_{t,w'}(z)$ is a homogeneous function of $z$: since all $a \in A$ are homogeneous, it is a homogeneous function of the sum of a set of projections (and all projections are homogeneous) and a set of homogeneous functions multiplied by a constant. To prove that the overall function is homogeneous, we introduce $\lambda \geq 0$, $v \in \mathbb{R}^{E'}$, $v' \in \mathbb{R}^{E''}$, and $v'' \in \mathbb{R}^{E'''}$, such that $v'_e = v_s$ for all $s \in E'$, $v''_e = v_s$ for all $s \in E''$, $v''_e = v''_e$, and $v' = \lambda w'$. Note that:

$$M_v(x) = c_{v',w'}(x)$$
$$= d_{t,w'}(v' \circ g_{w'}(x))$$
$$= d_{t,w'}((\lambda w') \circ g_{w'}(x))$$
$$= d_{t,w'}((\lambda v') \circ g_{w'}(x)).$$

Since $d_{t,w'}$ is homogeneous:

$$M_v(x) = \lambda d_{t,w'}(w' \circ g_{w'}(x))$$
$$= \lambda M_v(x).$$

We have established that $M_w(x)$ is homogeneous with respect to $w'$. Thus, from Theorem 29, Lemma 13 and Lemma 14, the other results hold.

G PROOFS FOR RESNETS AND CNNS

Recall the main theorem about ResNets and CNNs:

**Theorem 24** For a problem $\mathcal{P}$ and family of RC feedforward network models $\mathcal{R}(V,E,I,\{g_i\}_{i \in I},o^*,A,w^*)$ denoted $\{Q_v\}_{v \in \mathbb{R}^n}$, where all $a \in A$ are homogeneous: if $E_1 \ldots E_n$ is the partition of the dynamic parameters, and $B$ and $T$ are a well-behaved cut such that there is an $S \subseteq \{1 \ldots n\}$ where $E' = \bigcup_{s \in S} E_s$ is the cut set; then

1. $\{Q_v\}_{v \in \mathbb{R}^n}$ is a homogeneous model family with respect to $S$;
2. if for some $v \in \mathbb{R}^n$, $\frac{\partial L_P(Q_v)}{\partial v} = 0$ for all $s \in S$, the set $S$ is total on the training data given $Q_v$, and $L'$ is the family of linear models associated with $\{1 \ldots n\}$ given $Q_v$, then $Q_v$ is equal on the training data to any optimal $N'$ in $L'$.

3. if for some $v \in \mathbb{R}^n$, $\frac{\partial L_P(Q_v)}{\partial v} = 0$ for all $s \in \{1 \ldots n\}$, the set $\{1 \ldots n\}$ is total on the training data given $Q_v$, and $L'$ is the family of linear models associated with $\{1 \ldots n\}$ given $Q_v$, then $Q_v$ is equal on the training data to any optimal $N'$ in $L'$.

**Proof:** We need only prove that the model family is homogeneous: as with Theorem 29, the other two results follow from Theorem 29, Lemma 13, and Lemma 14. We can define $S' = \{1 \ldots n\} - S$, and $v'' \in \mathbb{R}^{S'}$ such that $v''_s = v_s$ for all $s \in S''$. We can define $v' \in \mathbb{R}^S$ such that $v'_s = v_s$ when $s \in S$. We can define $E'' = E - E'$, and $w'' : \mathbb{R}^{S''} \to \mathbb{R}^{E''}$ such that for all $e \in E''$, $w''(v''_e) = w''_e$ if $e \in E''$, and $w''_e = v''_e$ otherwise.

As before, we will define $d_{t,w''} : \mathbb{R}^{E''} \to \mathbb{R}$, and we will create a matrix $G^{v''} : X \to \mathbb{R}^{E''}$. As before, for all $t \in T$, for all $z \in \mathbb{R}^{E''}$, we define $d_{t,w''}$ recursively:

$$d_{t,w''}(z) = a_t \left( \sum_{u_t \in E''} z(u_t) \right)$$
$$+ \sum_{u_t \in E''} d_{t,w''}(z)w''(u_t)(v''_e) \right)$$

Since, for all $b \in B$, $c_{b,w''}(v(x))$ does not depend upon parameters in the cut set, we can write $G^{v''}(x)_c,s = 0$ if $\pi(c) \neq s$, and otherwise $G^{v''}(x)_c,s = c_{b,w''}(v(x))$. Crucially, neither $d$ nor $G$ is a function of the parameters in $S$. We can now write the activation energy of a node in the top as a combination of $d$, $G$, and $v'$:

$$c_{t,w''}(v(x)) = d_{t,w''}(G^{v''}(x)v')$$

Notice that $G^{v''}(x)v'$ is the product of a matrix and a vector, and $(G^{v''}(x)v')(b,t) = v''_{s}(b,t)c_{b,w''}(v(x))$. Moreover, this relies on the fact that the cut set does not include $E''$, as this would make the activation energies on the cut set an affine function of $v'$ instead of a linear one. Considering the output of the model is $c_{t,w''}(v(x))$ yields:

$$Q_v(x) = d_{t,w''}(G^{v''}(x)v')$$

Consider any $\lambda > 0$. Define $y \in \mathbb{R}^n$, where for all $s \in S''$, $y_s = u_s$, and for all $s \in S$, $y_s = \lambda u_s$. If we define $y'_e \in \mathbb{R}^S$ such that $y'_e = y_s$ for all $s \in S$, then we can write:

$$Q_y(x) = d_{t,w''}(G^{v''}(x)y')$$

$$Q_y(x) = d_{t,w''}(G^{v''}(x)(\lambda v'))$$

$$Q_y(x) = d_{t,w''}(G^{v''}(x)v')$$

As we argued in the proof of Theorem 29, $d_{t,w''}(z)$ is a homogeneous function of $z$. So:

$$Q_y(x) = \lambda d_{t,w''}(G^{v''}(x)v')$$

$$Q_y(x) = \lambda Q_v(x)$$

Thus, the RC class is homogeneous with respect to $S$. ■
H REGULARIZATION AND RESTRICTIONS

Now we prove Theorem 26.

Theorem 26 Given a problem \( \mathcal{P} \), a set \( S \), a subset \( S' \subseteq S \), a model family \( \mathcal{M} = \{ M_w \}_{w \in \mathbb{R}^S} \) that is homogeneous with respect to \( S' \), a strictly convex regularization function \( R : \mathcal{M} \rightarrow \mathbb{R} \) that is additively separable with respect to \( S' \), and a model \( M_w \) where for all \( s \in S' \):
\[
\frac{\partial[L_P(M_w) + R(M_w)]}{\partial w_s} = 0,
\]
if \( S' \) is total on the training data given \( M_w \), \( \mathcal{L}' = \{ N_v \}_{v \in \mathbb{R}^S} \) is the family of linear models associated with \( S' \) given \( M_w \), and a new regularizer \( R' : \mathcal{L}' \rightarrow \mathbb{R} \) is defined such that \( R'(N_v) = (R')^{S'}(v) \), then \( M_w \) must be equal on the training data to any optimal \( N \in \mathcal{L}' \) given the supervised learning problem and regularization \( R' \).

Proof: Define \( v' = R S' \) such that \( v'_s = w_s \) for all \( s \in S' \). For all \( s \in S' \), define \( f_s \) to be the feature generated by \( s \) given \( M_w \). We will prove the result by showing that:

1. For all \( s \in S' \), for all \( x \) in the training data, \( f_s(x) = \frac{\partial M_w(x)}{\partial w_s} \).
2. For all \( x \) in the training data, \( N_{v'}(x) = M_w(x) \).
3. For all \( s \in S' \), \( \frac{\partial L_P(N_{v'})}{\partial v'_s} = \frac{\partial L_P(M_w)}{\partial w_s} \).
4. For all \( s \in S' \), \( \frac{\partial R'(N_{v'})}{\partial v'_s} = \frac{\partial R(M_w)}{\partial w_s} \).
5. Now, we know that \( \frac{\partial[L_P(N_{v'}) + R'(N_{v'})]}{\partial v'_s} = 0 \). Then, we use this to prove that \( N_{v'} \) is the unique optimal solution in \( \mathcal{L}' \).

We begin by proving item 1. By definition, \( f_s(x) = \frac{\partial M_w(x)}{\partial w_s} \). Since \( S' \) is total on the training data given \( M_w \), \( M_w \) is partially differentiable with respect to \( w_s \), so \( f_s(x) = \frac{\partial M_w(x)}{\partial w_s} \).

Next we prove item 2 (c.f. Theorem 20). Note that, for any \( x \in X \):
\[
N_{v'}(x) = \sum_{s \in S'} v'_s f_s(x).
\]
By the definition of \( v' \):
\[
N_{v'}(x) = \sum_{s \in S'} w_s f_s(x).
\]
By item 1 if \( x \) is in the training data:
\[
N_{v'}(x) = \sum_{s \in S'} w_s \frac{\partial M_w(x)}{\partial w_s}.
\]
Because the model family \( \mathcal{M} \) is homogeneous, and \( S' \) is total on the training data given \( M_w \):
\[
N_{v'}(x) = M_w(x).
\]

Next, we prove item 3. Choose an arbitrary \( s \in S' \):
\[
\frac{\partial L_P(N_{v'})}{\partial v'_s} = \sum_{i=1}^m \frac{\partial L_P(y_i, \hat{y}'_i)}{\partial \hat{y}'_i} \bigg|_{\hat{y}'_i = N_{v'}(x_i)} f_s(x_i),
\]
By item 1:
\[
\frac{\partial L_P(N_{v'})}{\partial v'_s} = \sum_{i=1}^m \frac{\partial L_P(y_i, \hat{y}'_i)}{\partial \hat{y}'_i} \bigg|_{\hat{y}'_i = N_{v'}(x_i)} \frac{\partial M_w(x_i)}{\partial w_s}.
\]

Next, we prove item 4. Choose an arbitrary \( s \in S' \). Note that \( \frac{\partial R(M_w)}{\partial w_s} = \frac{\partial R'(w)}{\partial w_s} = \frac{\partial R'}{\partial v'_s} \frac{\partial v'_s}{\partial w_s} \). Since \( v' = \pi_S \rightarrow S'(w) \), \( \frac{\partial R'}{\partial v'_s} \frac{\partial v'_s}{\partial w_s} = \frac{\partial R'}{\partial v'_s} \). Note also by the definition of \( R' \):
\[
\frac{\partial R(M_w)}{\partial w_s} = \frac{\partial R'(N_{v'})}{\partial v'_s}.
\]
We prove item 4. By item 3 and item 4 for any \( s \in S' \):
\[
\frac{\partial[L_P(N_{v'}) + R'(N_{v'})]}{\partial v'_s} = \frac{\partial[L_P(M_w) + R(M_w)]}{\partial w_s}.
\]
Then, by assumption:
\[
\frac{\partial[L_P(N_{v'}) + R'(N_{v'})]}{\partial v'_s} = \frac{\partial[L_P(M_w) + R(M_w)]}{\partial w_s} = 0.
\]
Now, define a function \( g : \mathbb{R}^S' \rightarrow \mathbb{R} \) such that for all \( v \in \mathbb{R}^S' \), \( g(v) = L_P(N_v) + R'(N_v) \). Notice that the first part is convex, and the second part is strictly convex, implying \( g \) is strictly convex. Notice that \( \nabla g(v') = 0 \). Thus, \( v' \) is a minimum of \( g \). Moreover, since \( g \) is strictly convex, it can have no more than one minimum. So \( N_{v'} \) is the unique optimal model, and it is equivalent to \( M_w \).
In this section, we want to focus on when the concept of totality applies in deep networks. Specifically, are there easy rules of thumb to determine whether a feature set is total?

I.1 TOTALITY AND DIFFERENTIABILITY

While we invented the term "totality", it is very similar to the concept of differentiability. Specifically, consider $g : \mathbb{R}^S \rightarrow \mathbb{R}$, where for all $h \in \mathbb{R}^S$, $g(h) = M_{w+h}(x)$. Then, $M_w(x)$ is total if $g(h)$ is differentiable at zero.

I.2 TOTALITY AND FEEDFORWARD NETWORKS

Let’s focus on feedforward networks with relu gates, as those are relatively simple and have nice homogeneous properties. Specifically, assume that each input and the output have identity transformations, and the internal nodes are relu gates.

Notice that the features will always exist for feedforward networks. If one considers the output of a feedforward network as a function of weights of the edges given a fixed input, the output is a piecewise polynomial function.

The relu function has a single point of non-differentiability at zero. If we “avoid” this point, we will have a total model on an example. More formally, let’s take apart deep networks in a different way. Specifically, for all $v \in V - I$, $w \in \mathbb{R}^E$, define $k_{v,w} : X \rightarrow \mathbb{R}$ such that:

$$k_{v,w}(x) = \sum_{u : (u,v) \in E} c_{u,v}(w_{u,v}) \Delta (u,v).$$  (126)

Notice that, for all $v \in V - I$:

$$c_{v,w}(x) = a_w(k_{v,w}(x)).$$  (127)

In this section we assume $a_w$ is the identity for $v \in I \cup \{o^*\}$ and $a_w$ is a relu function elsewhere. Observe that, for some $w \in \mathbb{R}^E$, for some $x \in X$, if $k_{v,w}(x) \neq 0$ for all $v \in V - I$, then $E$ is total for $x$ given $M_w$. However, notice that if $k_{v,w}(x) = 0$ for some $v \in V - I$, that does not mean that the features are not total. Specifically, if $k_{v,w}(x) = 0$, but for all $u \in V$ where $(v,u) \in E$, $w_{v,u} = 0$, then effectively the node $v$ has no effect. Thus, we will define a node to be soft given $M_w$ and $x$.

If $k_{v,w}(x) \neq 0$ or $\frac{\partial^2 M_w(x)}{\partial k_{v,w}(x)} = 0$. Otherwise, we will say $v$ is hard. Notice that any node where all weights on outgoing edges are zero is soft.

**Lemma 31** Given a set $E' \subseteq E$, given a model $M_w$ and an example $x \in X$, if for all $(u,v) \in E'$, for all $v' \geq v$ (in the sense that there exists a path from $v$ to $v'$), $v'$ is soft, then $E'$ is total on $M_w$.

**Proof:** At a high level, we construct differentiable functions, starting with one that maps the input of $o^*$ to its output, and then incorporating more and more nodes and edges until we have incorporated all of $E'$. These functions will not be differentiable everywhere, simply where we need them to be.

Consider $V'$ to be the set of all vertices $v' \in V$ where there is a $(u,v) \in E'$ where $v \leq v'$. Now, without loss of generality, assume $E' = \{(u,v) \in E' : u \in V'\}$. Then, we know that all $v' \in V'$ are soft. Without loss of generality, assume $o^* \in V'$.

Now, we can arrange a total ordering on the vertices in $V'$ (in the opposite direction of the partial ordering induced by the graph), beginning with $o^*$, and we will denote the ordering $v_1 \ldots v_j$, such that there is no path from $v_j$ to $v_i$ for all $i > j$, and $v_1 = o^*$. We will denote $V_i = \{v_1 \ldots v_i\}$ (such that $V_j = V'$), and $E_i = \{(u,v) \in E : u \in V_i\}$. Define $w^{i} \in \mathbb{R}^{E_i}$ such that $w^{i}_{(u,v)} = w_{(u,v)}$ if $(u,v) \in E_i$. Define $\alpha^{i}_{v_1} = k_{v_1,w}(x)$. We define $d^{i} : \mathbb{R}^{V_i} \times \mathbb{R}^{E_i} \rightarrow \mathbb{R}$ such that $d^{i}((\alpha,\theta)) = a_{v_1}(\alpha) = \alpha$. Thus, $d^{i}(\alpha^{i}_{v_1}) = c_{v_1,w}(x)$, and $d^{i}$ is differentiable with respect to its arguments.

Now, we recursively define $d^{2} \ldots d^{i}$. Given $d^{i}$, we define $m^{i} : \mathbb{R}^{V_{i+1}} \times \mathbb{R}^{E_{i+1}} \rightarrow \mathbb{R}$ such that $(m^{i}(\alpha,w^{i}))_{v} = \alpha_{v} + \alpha_{v+1}(w^{i+1}_{v+1})$ if $(v_{i+1},v) \in E$ and $m^{i}(\alpha,w^{i}) = \alpha_{v}$ if $(v_{i+1},v) \notin E$. Moreover, for any sets $S$, and $T \subseteq S$, for any $v \in \mathbb{R}^S$, we define $\Pi^T(v)$ such that $\Pi^T(v)_{i} = v_{i}$ for all $i \in T$. Then, we can define $d^{i+1} : \mathbb{R}^{V_{i+1}} \times \mathbb{R}^{E_{i+1}} \rightarrow \mathbb{R}$ such that $d^{i+1}(\alpha,w^{i}) = d^{i}(m^{i}(\alpha,w^{i}),\Pi^{E_{i}}(w^{i}))$.

Recursively, we define $\alpha^{i+1} \in \mathbb{R}^{V_{i+1}}$ such that $\alpha^{i+1}_{v_{i+1}} = k_{v_{i+1},w}(x)$ and for all $v \in V_{i}$, $\alpha^{i+1}_{v} = \alpha^{i}_{v}$ otherwise. Thus, observe that $d^{i+1}(\alpha^{i+1},w^{i+1}) = d^{i}(\alpha^{i},w^{i})$, so recursively $d^{i+1}(\alpha^{i+1},w^{i+1}) = c_{v_{i+1},w}(x)$.

Moreover, recursively one can establish that for any $w^{i} \in \mathbb{R}^E$ where $\Pi^{E-E_{i+1}}(w^{i}) = \Pi^{E-E_{i+1}}(w)$, $d^{i+1}(\alpha^{i+1},\Pi^{E_{i}}(w^{i})) = c_{v_{i+1},w}(x)$.

Now, we must show differentiability. The inductive hypothesis is $d^{i}$ is differentiable, or more formally, for

---

12The following may not be obvious: however, it is a corollary of Lemma 31.

13Note that $E_i$ is the emptyset, as there are no outgoing edges from $o^*$.  

---

### Reference

- **Lemma 31**"
\[ h \in \mathbb{R}^{V_i} \times \mathbb{R}^{E_i} : \]
\[ \lim_{h \to 0} \frac{d^i(\alpha^i + \Pi^V_i(h), w^i + \Pi^E_i(h))}{\|h\|} = 0 \]  
(128)

First, we know that \( v_l \) is soft. If \( k_{v_{l+1}, w}(x) \neq 0 \), then \( \alpha_{v_{l+1}}^{i+1} \neq 0 \). If \( \alpha_{v_{l+1}}^{i+1} < 0 \), then in a region around \( \alpha_{v_{l+1}}^{i+1}, w^{i+1}_l, m^i((\alpha_{v_{l+1}}^{i+1}, w^{i+1}_l) + h) = \alpha^i + \Pi^V_i(h) \), i.e. is linear. If \( \alpha_{v_{l+1}}^{i+1} > 0 \), then in a region around \( \alpha_{v_{l+1}}^{i+1}, w^{i+1}_l, m^i((\alpha_{v_{l+1}}^{i+1}, w^{i+1}_l) + h) = \alpha_l^i + h_v + h_{v_{l+1}}(w^{i+1}_{(v_{l+1},v)} + h_{(v_{l+1},v)}) \) if \( (v_{l+1},v) \in E \), and \( m^i((\alpha_{v_{l+1}}^{i+1}, w^{i+1}_l) + h_v) = \alpha_l^i + h_v \) otherwise (again, linear).

Thus, if \( k_{v_{l+1}, w}(x) \neq 0 \), then \( \alpha_{v_{l+1}}^{i+1} \neq 0 \), in a region around \( \alpha_{v_{l+1}}^{i+1} \) and \( w^i \), \( m^i \) is linear. Then \( d^{i+1} \) is the composition of a differentiable function, a linear function, a projection, and therefore differentiable at \( \alpha_{v_{l+1}}^{i+1} \) and \( w^i \).

On the other hand, if \( k_{v_{l+1}, w}(x) = 0 \), then \( \alpha_{v_{l+1}}^{i+1} = 0 \). Since \( v_l \) is soft, \( \frac{\partial^i m^i}{\partial \alpha_{v_{l+1}}^{i+1}}(x, \alpha^i, w^i) = 0 \). This implies \( \frac{\partial^i d^{i+1}(\alpha_{v_{l+1}}^{i+1}, w^{i+1}_l)}{\partial \alpha_{v_{l+1}}^{i+1}} = 0 \). Using the chain rule for directional derivatives:
\[ 0 = \frac{\partial^i d^{i+1}(\alpha_{v_{l+1}}^{i+1}, w^{i+1}_l)}{\partial \alpha_{v_{l+1}}^{i+1}} \]
(129)
\[ 0 = \frac{\partial^i d^i(\alpha^i, w^i)}{\partial \alpha_{v_{l+1}}^{i+1}} \]
(130)

Where \( g = \frac{\partial^i m^i(\alpha_{v_{l+1}}^{i+1}, w^{i+1}_l)}{\partial \alpha_{v_{l+1}}^{i+1}} \), so \( g_v = w^{i+1}_{v_{l+1}, v} \) if \( (v_{l+1},v) \in E \), and \( g_v = 0 \) otherwise. If \( J^i \) is the derivative of \( d^i \) at \( \alpha^i, w^i \), then, because \( v_{l+1} \) is soft and \( \alpha_{v_{l+1}}^{i+1} = 0 \):
\[ \sum_{v \in V_l} J^i_v g_v = 0 \]
(131)
\[ \sum_{(v_{l+1},v) \in E} J^i_{v_{l+1}} w = 0 \]
(132)

In order to prove differentiability of \( d^{i+1} \), we introduce a function \( c^i : \mathbb{R}^{V_i} \times \mathbb{R}^{E_i} \to \mathbb{R} \) quantifying the error of the derivative of \( d^i \), such that for any \( h^\alpha \in \mathbb{R}^{V_i}, h^w \in \mathbb{R}^{E_i}, c(h^\alpha, h^w) = d^i(\alpha^i + h^\alpha, w^i + h^w) - \left( \sum_{v \in V_l} J^i_v h^\alpha_v + \sum_{e \in E} J^i_e h^w_e \right) \).

Thus, \( \lim_{h^\alpha, h^w \to 0} \frac{c(h^\alpha, h^w)}{\|h^\alpha, h^w\|} = 0 \), where \( \|h^\alpha, h^w\| = \sqrt{\|h^\alpha\|^2 + \|h^w\|^2} \).

First, although \( m^i \) is not differentiable when \( \alpha_{v_{l+1}}^{i+1} = 0 \), the derivative is “almost” the linear projection operator \( \Pi^V_i \), because the partial derivative of \( \alpha_{v_{l+1}}^{i+1} = 0 \). We can prove this using a proof similar to the proof of the chain rule. We will denote the following \( \eta \), and will endeavor to prove it exists and is zero.

\[ \eta = \lim_{h^\alpha, h^w \to 0} \frac{d^{i+1}(\alpha^{i+1} + h^\alpha, w^{i+1} + h^w)}{\|h^\alpha, h^w\|} \]
(133)
\[ - \frac{d^{i+1}(\alpha^{i+1}, w^{i+1}) + \sum_{v \in V_l} J^i_v h^\alpha_v + \sum_{e \in E} J^i_e h^w_e}{\|h^\alpha, h^w\|} \]
(134)

The last sum is a linear operator: notice that in this case, we are effectively showing \( J^{i+1} = J^i \). First, we note that \( d^{i+1}(\alpha^{i+1}, w^{i+1}) = d^i(\alpha^i, w^i) \):

\[ \eta = \lim_{h^\alpha, h^w \to 0} \frac{d^{i+1}(\alpha^{i+1} + h^\alpha, w^{i+1} + h^w)}{\|h^\alpha, h^w\|} \]
(135)
\[ - \frac{d^i(\alpha^i, w^i) + \sum_{v \in V_l} J^i_v h^\alpha_v + \sum_{e \in E} J^i_e h^w_e}{\|h^\alpha, h^w\|} \]
(136)

Furthermore, we study the first term separately:

\[ d^{i+1}(\alpha^{i+1} + h^\alpha, w^{i+1} + h^w) = d^i(m^i(\alpha^{i+1} + h^\alpha, w^{i+1} + h^w), \Pi^E_i(w^{i+1} + h^w)) \]
(137)

Write \( a = a_{\alpha^{i+1}} \), which is a relu function, because \( v_{l+1} \neq \alpha^* \), and \( v_{l+1} \notin I \). Using \( e \) we get:

\[ d^{i+1}(\alpha^{i+1} + h^\alpha, w^{i+1} + h^w) \]
\[ = d^i(m^i(\alpha^{i+1} + h^\alpha, w^{i+1} + h^w), \Pi^E_i(w^{i+1} + h^w)) + \sum_{v \in V_l} J^i_v h^\alpha_v \]
\[ + \sum_{(v_{l+1},v) \in E} J^i_{v_{l+1}} a(h^\alpha_{v_{l+1}})(w_{v_{l+1},v} + h^w_{v_{l+1},v}) \]
\[ + \sum_{e \in E} J^i_e h^w_e \]
(138)

Focusing on the most complex term:

\[ \sum_{(v_{l+1},v) \in E} J^i_{v_{l+1}} a(h^\alpha_{v_{l+1}})(w_{v_{l+1},v} + h^w_{v_{l+1},v}) \]
\[ = a(h^\alpha_{v_{l+1}}) \sum_{(v_{l+1},v) \in E} J^i_{v_{l+1}} w_{v_{l+1},v} + a(h^\alpha_{v_{l+1}}) \sum_{e \in E} J^i_e h^w_e \]
(139)

Note that \( \sum_{(v_{l+1},v) \in E} J^i_{v_{l+1}} w_{v_{l+1},v} = 0 \), so:

\[ \sum_{(v_{l+1},v) \in E} J^i_{v_{l+1}} a(h^\alpha_{v_{l+1}})(w_{v_{l+1},v} + h^w_{v_{l+1},v}) \]
\[ = a(h^\alpha_{v_{l+1}}) \sum_{e \in E} J^i_e h^w_e \]
(140)
We next consider the term \( \epsilon_0 \):

\[
\lim_{h^\alpha, h^w \to 0} \sum_{(v_i+1,v) \in E} J_i^\alpha (h_{v_i+1}) (w_{v_i+1,v} + h^w_{v_i+1,v}) \frac{h^\alpha, h^w}{\|h^\alpha, h^w\|} = \lim_{h^\alpha, h^w \to 0} \frac{a(h_{v_i+1}^\alpha)}{\|h^\alpha, h^w\|} \sum_{(v_i+1,v) \in E} J_i^\alpha (h_{v_i+1}) \frac{h^w_{v_i+1,v}}{\|h^\alpha, h^w\|} \tag{141}
\]

Taking the absolute value:

\[
\lim_{h^\alpha, h^w \to 0} \left| \sum_{(v_i+1,v) \in E} J_i^\alpha (h_{v_i+1}^\alpha) (w_{v_i+1,v} + h^w_{v_i+1,v}) \right| \frac{\|h^\alpha, h^w\|}{\|h^\alpha, h^w\|} = \lim_{h^\alpha, h^w \to 0} \frac{|a(h_{v_i+1}^\alpha)|}{\|h^\alpha, h^w\|} \sum_{(v_i+1,v) \in E} J_i^\alpha \frac{|h^w_{v_i+1,v}|}{\|h^\alpha, h^w\|} \tag{142}
\]

Since \( |a(h_{v_i+1}^\alpha)| \leq |h_{v_i+1}^\alpha| : \)

\[
\lim_{h^\alpha, h^w \to 0} \frac{\sum_{(v_i+1,v) \in E} J_i^\alpha (h_{v_i+1}^\alpha) (w_{v_i+1,v} + h^w_{v_i+1,v})}{\|h^\alpha, h^w\|} = \lim_{h^\alpha, h^w \to 0} \frac{h_{v_i+1}^\alpha}{\|h^\alpha, h^w\|} \sum_{(v_i+1,v) \in E} J_i^\alpha \frac{|h^w_{v_i+1,v}|}{\|h^\alpha, h^w\|} \tag{143}
\]

The quadratic terms in the numerator mean that the limit is zero, because for any \( i, j \in \{1 \ldots n\} \lim_{z \to 0} \frac{z_i z_j}{z^2} = 0. \)

\[
\lim_{h^\alpha, h^w \to 0} \frac{\sum_{(v_i+1,v) \in E} J_i^\alpha (h_{v_i+1}^\alpha) (w_{v_i+1,v} + h^w_{v_i+1,v})}{\|h^\alpha, h^w\|} = 0 \lim_{h^\alpha, h^w \to 0} \frac{\epsilon_i (m_i^0 (\alpha^{i+1} + h^\alpha, w^{i+1} + h^w), \Pi_i (w^{i+1} + h^w))]}{\|h^\alpha, h^w\|} = \frac{\epsilon_i (m_i^0 (\alpha^{i+1} + h^\alpha, w^{i+1} + h^w), \Pi_i (w^{i+1} + h^w))]}{\|h^\alpha, h^w\|} = 0 \tag{150}
\]

So, the limit of the absolute value:

\[
\lim_{h^\alpha, h^w \to 0} \frac{\epsilon_i (m_i^0 (\alpha^{i+1} + h^\alpha, w^{i+1} + h^w), \Pi_i (w^{i+1} + h^w))]}{\|h^\alpha, h^w\|} = \frac{\epsilon_i (m_i^0 (\alpha^{i+1} + h^\alpha, w^{i+1} + h^w), \Pi_i (w^{i+1} + h^w))]}{\|h^\alpha, h^w\|} = 0 \tag{151}
\]

Now we have established that:

\[
\lim_{h^\alpha, h^w \to 0} \frac{d_i^0 (\alpha^{i+1} + h^\alpha, w^{i+1} + h^w)}{\|h^\alpha, h^w\|} = \lim_{h^\alpha, h^w \to 0} \frac{d_i^0 (\alpha^i, w^i)}{\|h^\alpha, h^w\|} + \frac{\sum_{c \in E_i} J_i^c h^w_c}{\|h^\alpha, h^w\|} + \frac{\sum_{c \in E_i} J_i^c h^w_c}{\|h^\alpha, h^w\|} \tag{152}
\]
Plugging this into $\eta$ yields $\eta = 0$, implying that $d^{i+1}$ is differentiable. Recursively, we have established that $d^j$ is differentiable. However, $E_j$ is a subset of $E'$: by construction, $V' = V_j$ are the nodes that are at the end of the edges in $E'$, not at the beginning, and $E_j$ are the edges that can be reached from $V_j$. The final step is to construct a function $f : \mathbb{R}^{E'} \to \mathbb{R}$ as a function defined on all the relevant edges. Define $E^* = E' - E_j$. For all $(u, v) \in E^*$, define $\beta_{(u,v)} = c_{u,w}(x)$. Define $w^* \in \mathbb{R}^{E'}$ such that $w^*_e = w_e$ for all $e \in E'$. Then, we define $m^* : \mathbb{R}^{E'} \to \mathbb{R}^{V_j}$ such that for all $v \in V_j$, for all $w' \in \mathbb{R}^{E'}$

$$m^*(w') = \sum_{(u,v) \in E^*} w'_{(u,v)} \beta_{(u,v)} \quad (153)$$

Now we can define $f(w') = d^j(m^*(w'), \Pi_{E_j}(w'))$. Note that $f(w^*) = M_w(x)$: moreover, for any $w' \in \mathbb{R}^{E'}$, if $\Pi^{E-E'}(w') = \Pi^{E-E'}(w)$, then $f(\Pi^{E-E'}(w')) = M_{w'}(x)$. Finally, since $d^j$ is differentiable and $m^*$ is linear, then $f$ is differentiable at $w^*$. This implies that $E'$ is total on $x$ given $M_w(x)$.