Sparse Linear Networks with a Fixed Butterfly Structure: Theory and Practice (Supplementary Material)

Nir Ailon[1] **Omer Leibovitch**[1] **Vineet Nair**[1]

[1]Technion Israel Institute of Technology
[1,3]{nailon, vineet}@cs.technion.ac.il
[2]leibovitch@campus.technion.ac.il

1 BUTTERFLY DIAGRAM FROM SECTION 1

Figure 1 referred to in the introduction is given here.

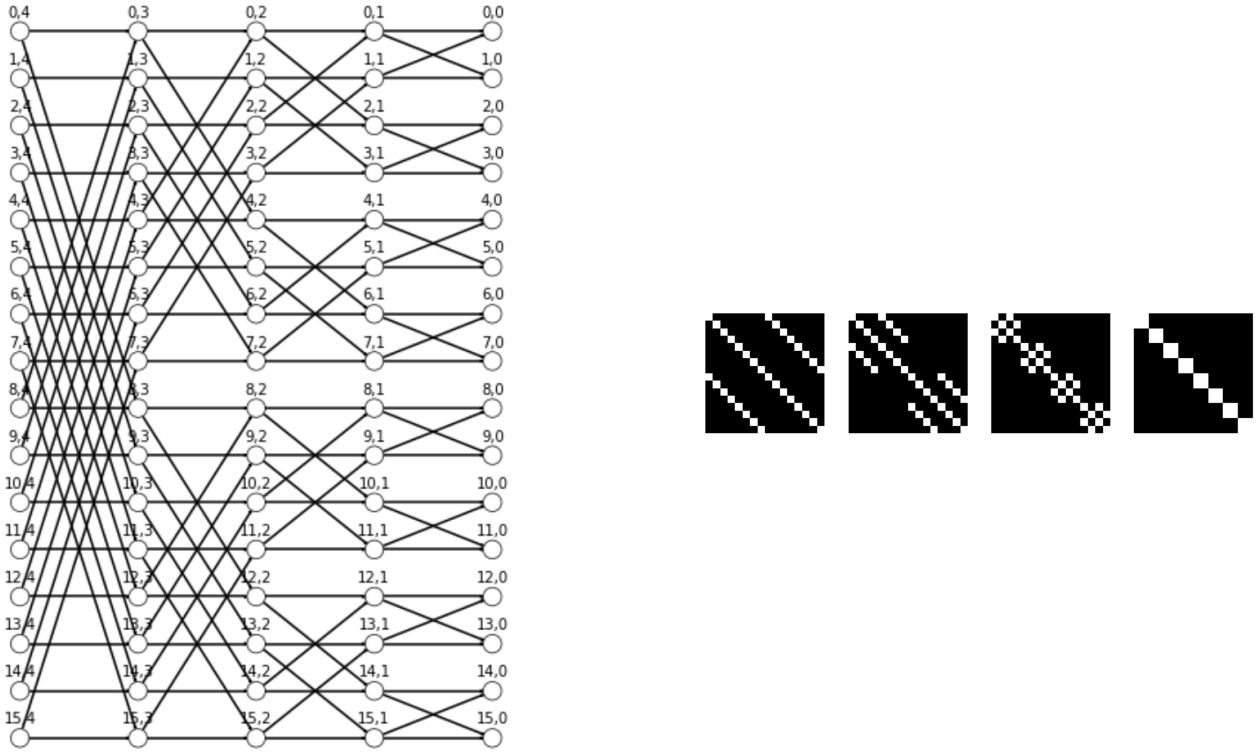

Figure 1: A 16×16 butterfly network represented as a 4-layered graph on the left, and as product of 4 sparse matrices on the right. The white entries are the non-zero entries of the matrices.

2 PROOF OF PROPOSITION 1

The proof of the proposition will use the following well known fact (Lemma 2.1 below) about FJLT (more generally, JL) distributions (see Ailon and Chazelle [2009], Ailon and Liberty [2009], Krahmer and Ward [2011]).

Lemma 2.1. *Let $\mathbf{x} \in \mathbb{R}^n$ be a unit vector, and let $J \in \mathbb{R}^{k \times n}$ be a matrix drawn from an FJLT distribution. Then for all $\epsilon < 1$ with probability at least $1 - e^{-\Omega(k\epsilon^2)}$:*

$$\|\mathbf{x} - J^T J\mathbf{x}\| \leq \epsilon \, . \tag{1}$$

By Lemma 2.1 we have that with probability at least $1 - e^{-\Omega(k_1\epsilon^2)}$,

$$\|\mathbf{x} - J_1^T J_1\mathbf{x}\| \leq \epsilon\|\mathbf{x}\| = \epsilon \, . \tag{2}$$

Henceforth, we condition on the event $\|\mathbf{x} - J_1^T J_1\mathbf{x}\| \leq \epsilon\|\mathbf{x}\|$. Therefore, by the definition of spectral norm $\|W\|$ of W:

$$\|W\mathbf{x} - WJ_1^T J_1\mathbf{x}\| \leq \epsilon\|W\| \, . \tag{3}$$

Now apply Lemma 2.1 again on the vector $WJ_1^T J_1\mathbf{x}$ and transformation J_2 to get that with probability at least $1 - e^{-\Omega(k_2\epsilon^2)}$,

$$\|WJ_1^T J_1\mathbf{x} - J_2^T J_2 WJ_1^T J_1\mathbf{x}\| \leq \epsilon\|WJ_1^T J_1\mathbf{x}\|. \tag{4}$$

Henceforth, we condition on the event $\|WJ_1^T J_1\mathbf{x} - J_2^T J_2 WJ_1^T J_1\mathbf{x}\| \leq \epsilon\|WJ_1^T J_1\mathbf{x}\|$. To bound the last right hand side, we use the triangle inequality together with (3):

$$\|WJ_1^T J_1\mathbf{x}\| \leq \|W\mathbf{x}\| + \epsilon\|W\| \leq \|W\|(1+\epsilon). \tag{5}$$

Combining (4) and (5) gives:

$$\|WJ_1^T J_1\mathbf{x} - J_2^T J_2 WJ_1^T J_1\mathbf{x}\| \leq \epsilon\|W\|(1+\epsilon). \tag{6}$$

Finally,

$$\begin{aligned} \|J_2^T J_2 WJ_1^T J_1\mathbf{x} - W\mathbf{x}\| &= \|(J_2^T J_2 WJ_1^T J_1\mathbf{x} - WJ_1^T J_1\mathbf{x}) + (WJ_1^T J_1\mathbf{x} - W\mathbf{x})\| \\ &\leq \epsilon\|W\|(1+\epsilon) + \epsilon\|W\| \\ &= \|W\|\epsilon(2+\epsilon) \leq 3\|W\|\epsilon \, , \end{aligned} \tag{7}$$

where the first inequality is from the triangle inequality together with (3) and (6), and the second inequality is from the bound on ϵ. The proposition is obtained by adjusting the constants hiding inside the $\Omega()$ notation in the exponent in the proposition statement.

3 PROOF OF THEOREM 1

We first note that our result continues to hold even if B in the theorem is replaced by any structured matrix. For example the result continues to hold if B is an $\ell \times n$ matrix with one non-zero entry per column, as is the case with a random sparse sketching matrix Clarkson and Woodruff [2009]. We also compare our result with that Baldi and Hornik [1989], Kawaguchi [2016].

Comparison with Baldi and Hornik [1989] and Kawaguchi [2016]: The critical points of the encoder-decoder network are analyzed in Baldi and Hornik [1989]. Suppose the eigenvalues of $YX^T(XX^T)^{-1}XY^T$ are $\gamma_1 > \ldots > \gamma_m > 0$ and $k \leq m \leq n$. Then they show that corresponding to a critical point there is an $I \subseteq [m]$ such that the loss at this critical point is equal to $\text{tr}(YY^T) - \sum_{i \in I}\gamma_i$, and the critical point is a local/global minima if and only if $I = [k]$. Kawaguchi [2016] later generalized this to prove that a local minima is a global minima for an arbitrary number of hidden layers in a linear neural network if $m \leq n$. Note that since $\ell \leq n$ and $m \leq n$ in Theorem 1, replacing X by BX in Baldi and Hornik [1989] or Kawaguchi [2016] does not imply Theorem 1 as it is.

Next, we introduce a few notation before delving into the proof. Let $r = (\overline{Y} - Y)^T$, and $\text{vec}(r) \in \mathbb{R}^{md}$ is the entries of r arranged as a vector in column-first ordering, $(\nabla_{\text{vec}(D^T)}\mathcal{L}(\overline{Y}))^T \in \mathbb{R}^{mk}$ and $(\nabla_{\text{vec}(E^T)}\mathcal{L}(\overline{Y}))^T \in \mathbb{R}^{k\ell}$ denote the partial derivative of $\mathcal{L}(\overline{Y})$ with respect to the parameters in $\text{vec}(D^T)$ and $\text{vec}(E^T)$ respectively. Notice that $\nabla_{\text{vec}(D^T)}\mathcal{L}(\overline{Y})$ and $\nabla_{\text{vec}(E^T)}\mathcal{L}(\overline{Y})$ are row vectors of size mk and $k\ell$ respectively. Also, let P_D denote the projection matrix of D, and hence if D is a matrix with full column-rank then $P_D = D(D^T \cdot D)^{-1} \cdot D^T$. The $n \times n$ identity matrix is denoted as I_n, and for convenience of notation let $\tilde{X} = B \cdot X$. First we prove the following lemma which gives an expression for D and E if $\nabla_{\text{vec}(D^T)}\mathcal{L}(\overline{Y})$ and $\nabla_{\text{vec}(E^T)}\mathcal{L}(\overline{Y})$ are zero.

Lemma 3.1 (Derivatives with respect to D and E)**.**

1. $\nabla_{\text{vec}(D^T)}\mathcal{L}(\overline{Y}) = vec(r)^T(I_m \otimes (E \cdot \tilde{X})^T)$, *and*
2. $\nabla_{\text{vec}(E^T)}\mathcal{L}(\overline{X}) = vec(r)^T(D \otimes \tilde{X})^T$

Proof. 1. Since $\mathcal{L}(\overline{Y}) = \frac{1}{2}\text{vec}(r)^T \cdot \text{vec}(r)$,

$$\begin{aligned}\nabla_{\text{vec}(D^T)}\mathcal{L}(\overline{Y}) = \text{vec}(r)^T \cdot \nabla_{\text{vec}(D^T)}\text{vec}(r) \;&=\; \text{vec}(r)^T(\text{vec}_{(D^T)}(\tilde{X}^T \cdot E^T \cdot D^T)) \\ = \text{vec}(r)^T(I_m \otimes (E \cdot \tilde{X})^T) \cdot \nabla_{\text{vec}(D^T)}\text{vec}(D^T) \;&=\; \text{vec}(r)^T(I_m \otimes (E \cdot \tilde{X})^T)\end{aligned}$$

2. Similarly,

$$\begin{aligned}\nabla_{\text{vec}(E^T)}\mathcal{L}(\overline{Y}) = \text{vec}(r)^T \cdot \nabla_{\text{vec}(E^T)}\text{vec}(r) \;&=\; \text{vec}(r)^T(\text{vec}_{(E^T)}(\tilde{X}^T \cdot E^T \cdot D^T)) \\ = \text{vec}(r)^T(D \otimes \tilde{X}^T) \cdot \nabla_{\text{vec}(E^T)}\text{vec}(E^T) \;&=\; \text{vec}(r)^T(D \otimes \tilde{X}^T)\end{aligned}$$

□

Assume the rank of D is equal to p. Hence there is an invertible matrix $C \in \mathbb{R}^{k\times k}$ such that $\tilde{D} = D \cdot C$ is such that the last $k-p$ columns of $\tilde{D}$ are zero and the first p columns of $\tilde{D}$ are linearly independent (via Gauss elimination). Let $\tilde{E} = C^{-1} \cdot E$. Without loss of generality it can be assumed $\tilde{D} \in \mathbb{R}^{d\times p}$, and $\tilde{E} \in \mathbb{R}^{p\times d}$, by restricting restricting $\tilde{D}$ to its first p columns (as the remaining are zero) and $\tilde{E}$ to its first p rows. Hence, $\tilde{D}$ is a full column-rank matrix of rank p, and $DE = \tilde{D}\tilde{E}$. Claims 3.1 and 3.2 aid us in the completing the proof of the theorem. First the proof of theorem is completed using these claims, and at the end the two claims are proved.

Claim 3.1 (Representation at the critical point)**.**

1. $\tilde{E} = (\tilde{D}^T\tilde{D})^{-1}\tilde{D}^T Y\tilde{X}^T(\tilde{X} \cdot \tilde{X}^T)^{-1}$
2. $\tilde{D}\tilde{E} = P_{\tilde{D}} Y\tilde{X}^T(\tilde{X} \cdot \tilde{X}^T)^{-1}$

Claim 3.2. 1. $\tilde{E}B\tilde{D} = (\tilde{E}BY\tilde{X}^T\tilde{E}^T)(\tilde{E}\tilde{X}\tilde{X}^T\tilde{E}^T)^{-1}$

2. $P_{\tilde{D}}\Sigma = \Sigma P_{\tilde{D}} = P_{\tilde{D}}\Sigma P_{\tilde{D}}$

We denote $\Sigma(B)$ as Σ for convenience. Since Σ is a real symmetric matrix, there is an orthogonal matrix U consisting of the eigenvectors of Σ, such that $\Sigma = U \wedge U^T$, where $\wedge$ is a $m \times m$ diagonal matrix whose first ℓ diagonal entries are $\lambda_1, \ldots, \lambda_\ell$ and the remaining entries are zero. Let $u_1, \ldots, u_m$ be the columns of U. Then for $i \in [\ell]$, u_i is the eigenvector of Σ corresponding to the eigenvalue λ_i, and $\{u_{\ell+1}, \ldots, u_{d_y}\}$ are the eigenvectors of Σ corresponding to the eigenvalue 0.

Note that $P_{U^T\tilde{D}} = U^T\tilde{D}(\tilde{D}^TU^TU\tilde{D})^{-1}\tilde{D}^TU = U^TP_{\tilde{D}}U$, and from part two of Claim 3.2 we have

$$(UP_{U^T\tilde{D}}U^T)\Sigma = \Sigma(UP_{U^T\tilde{D}}U^T) \tag{8}$$

$$U \cdot P_{U^T\tilde{D}} \wedge U^T = U \wedge P_{U^T\tilde{D}}U^T \tag{9}$$

$$P_{U^T\tilde{D}}\wedge = \wedge P_{U^T\tilde{D}} \tag{10}$$

Since $P_{U^T\tilde{D}}$ commutes with $\wedge$, $P_{U^T\tilde{D}}$ is a block-diagonal matrix comprising of two blocks P_1 and P_2: the first block P_1 is an $\ell \times \ell$ diagonal block, and P_2 is a $(m-\ell) \times (m-\ell)$ matrix. Since $P_{U^T\tilde{D}}$ is orthogonal projection matrix of rank p its eigenvalues are 1 with multiplicity p and 0 with multiplicity $m-p$. Hence at most p diagonal entries of P_1 are 1 and the remaining are 0. Finally observe that

$$\begin{aligned}\mathcal{L}(\overline{Y}) &= \text{tr}((\overline{Y}-Y)(\overline{Y}-Y)^T) \\ &= \text{tr}(YY^T) - 2\text{tr}(\overline{Y}Y^T) + \text{tr}(\overline{YY}^T) \\ &= \text{tr}(YY^T) - 2\text{tr}(P_{\tilde{D}}\Sigma) + \text{tr}(P_{\tilde{D}}\Sigma P_{\tilde{D}}) \\ &= \text{tr}(YY^T) - \text{tr}(P_{\tilde{D}}\Sigma)\end{aligned}$$

The second line in the above equation follows using the fact that $\text{tr}(\overline{Y}Y^T) = \text{tr}(Y\overline{Y}^T)$, the third line in the above equation follows by substituting $\overline{Y} = P_{\tilde{D}} Y\tilde{X}^T \cdot (\tilde{X} \cdot \tilde{X}^T)^{-1} \cdot \tilde{X}$ (from part two of Claim 3.1), and the last line follows from part two of Claim 3.2. Substituting $\Sigma = U \wedge U^T$, and $P_{\tilde{D}} = UP_{U^T\tilde{D}}U^T$ in the above equation we have,

$$\begin{aligned}\mathcal{L}(\overline{Y}) &= \text{tr}(YY^T) - \text{tr}(UP_{U^T\tilde{D}} \wedge U^T) \\ &= \text{tr}(YY^T) - \text{tr}(P_{U^T\tilde{D}}\wedge)\end{aligned}$$

The last line the above equation follows from the fact that $\text{tr}(UP_{U^T\tilde{D}} \wedge U^T) = \text{tr}(P_{U^T\tilde{D}} \wedge U^TU) = \text{tr}(P_{U^T\tilde{D}}\wedge)$. From the structure of $P_{U^T\tilde{D}}$ and $\wedge$ it follows that there is a subset $I \subseteq [\ell]$, $|I| \leq p$ such that $\text{tr}(P_{U^T\tilde{D}}\wedge) = \sum_{i\in I}\lambda_i$. Hence, $\mathcal{L}(\overline{Y}) = \text{tr}(YY^T) - \sum_{i\in I}\lambda_i$.

Since $P_{\tilde{D}} = UP_{U^T\tilde{D}}U^T$, there is a $p \times p$ invertible matrix M such that

$$\tilde{D} = (U \cdot V)_{I'} \cdot M \ \text{ , and } \ \tilde{E} = M^{-1}(V^TU^T)_{I'}Y\tilde{X}^T(\tilde{X}\tilde{X}^T)^{-1}$$

where V is a block-diagonal matrix consisting of two blocks V_1 and V_2: V_1 is equal to I_ℓ, and V_2 is an $(m-\ell)\times(m-\ell)$ orthogonal matrix, and I' is such that $I \subseteq I'$ and $|I'| = p$. The relation for $\tilde{E}$ in the above equation follows from part one of Claim 3.1. Note that if $I' \subseteq [\ell]$, then $I = I'$, that is I consists of indices corresponding to eigenvectors of non-zero eigenvalues.

Recall that $\tilde{D}$ was obtained by truncating the last $k-p$ zero rows of DC, where C was a $k \times k$ invertible matrix simulating the Gaussian elimination. Let $[M|O_{p\times(k-p)}]$ denoted the $p \times k$ matrix obtained by augmenting the columns of M with $(k-p)$ zero columns. Then

$$D = (UV)_{I'}[M|O_{p\times(k-p)}]C^{-1} \ .$$

Similarly, there is a $p \times (k-p)$ matrix N such that

$$E = C[\tfrac{M^{-1}}{N}]((UV)_{I'})^TY\tilde{X}^T(\tilde{X}\tilde{X}^T)^{-1}$$

where $[\frac{M^{-1}}{N}]$ denotes the $k \times p$ matrix obtained by augmenting the rows of M^{-1} with the rows of N. Now suppose $I \neq [k]$, and hence $I' \neq [k]$. Then we will show that there are matrices D' and E' arbitrarily close to D and E respectively such that if $Y' = D'E'\tilde{X}$ then $\mathcal{L}(Y') < \mathcal{L}(\overline{Y})$. There is an $a \in [k] \setminus I'$, and $b \in I'$ such that $\lambda_a > \lambda_b$ (λ_b could also be zero). Denote the columns of the matrix UV as $\{v_1, \ldots, v_m\}$, and observe that $v_i = u_i$ for $i \in [\ell]$ (from the structure of V). For $\epsilon > 0$ let $u'_b = (1+\epsilon^2)^{-\frac{1}{2}}(v_b + \epsilon u_a)$. Define U' as the matrix which is equal to UV except that the column vector v_b in UV is replaced by u'_b in U'. Since $a \in [k] \subseteq [\ell]$ and $a \notin I'$, $v_a = u_a$ and $(U'_{I'})^TU'_{I'} = I_p$. Define

$$D' = U'_{I'}[M|O_{p\times(k-p)}]C^{-1} \ \text{ , and } \ E' = C[\tfrac{M^{-1}}{N}](U'_{I'})^TY\tilde{X}^T(\tilde{X}\tilde{X}^T)^{-1}$$

and let $Y' = D'E'\tilde{X}$. Now observe that, $D'E' = U'_{I'}(U_{I'})^TY\tilde{X}^T(\tilde{X}\tilde{X}^T)^{-1}$, and that

$$\mathcal{L}(Y') = \text{tr}(YY^T) - \sum_{i\in I}\lambda_i - \frac{\epsilon^2}{1+\epsilon^2}(\lambda_a - \lambda_b) = \mathcal{L}(\overline{Y}) - \frac{\epsilon^2}{1+\epsilon^2}(\lambda_a - \lambda_b)$$

Since ϵ can be set arbitrarily close to zero, it can be concluded that there are points in the neighbourhood of $\overline{Y}$ such that the loss at these points are less than $\mathcal{L}(\overline{Y})$. Further, since $\mathcal{L}$ is convex with respect to the parameters in D (respectively E), when the matrix E is fixed (respectively D is fixed) $\overline{Y}$ is not a local maximum. Hence, if $I \neq [k]$ then $\overline{Y}$ represents a saddle point, and in particular $\overline{Y}$ is local/global minima if and only if $I = [k]$.

Proof of Claim 3.1. Since $\nabla_{\text{vec}(E^T)}\mathcal{L}(\overline{X})$ is equal to zero, from the second part of Lemma 3.1 the following holds,

$$\begin{aligned}&\tilde{X}(Y-\overline{Y})^TD = \tilde{X}Y^TD - \tilde{X}\overline{Y}^TD = 0 \\ \Rightarrow\ &\tilde{X}\tilde{X}^TE^TD^TD = \tilde{X}Y^TD\end{aligned}$$

Taking transpose on both sides

$$\Rightarrow D^TDE\tilde{X}\tilde{X}^T = D^TY\tilde{X}^T \tag{11}$$

Substituting DE as $\tilde{D}\tilde{E}$ in Equation 11, and multiplying Equation 11 by C^T on both the sides from the left, Equation 12 follows.

$$\Rightarrow \tilde{D}^T\tilde{D}\tilde{E}\tilde{X}\tilde{X}^T = \tilde{D}^T Y\tilde{X}^T \tag{12}$$

Since $\tilde{D}$ is full-rank, we have

$$\tilde{E} = (\tilde{D}^T\tilde{D})^{-1}\tilde{D}^T Y\tilde{X}^T(\tilde{X}\tilde{X}^T)^{-1}. \tag{13}$$

and,

$$\tilde{D}\tilde{E} = P_{\tilde{D}} Y\tilde{X}^T(\tilde{X}\tilde{X}^T)^{-1} \tag{14}$$

□

Proof of Claim 3.2. Since $\nabla_{\text{vec}(D^T)}\mathcal{L}(\overline{Y})$ is zero, from the first part of Lemma 3.1 the following holds,

$$E\tilde{X}(Y-\overline{Y})^T = E\tilde{X}Y^T - E\tilde{X}\cdot\overline{Y}^T = 0$$

$$\Rightarrow E\tilde{X}\tilde{X}^T E^T D^T = E\tilde{X}Y^T \tag{15}$$

Substituting $E^T\cdot D^T$ as $\tilde{E}^T\cdot\tilde{D}^T$ in Equation 11, and multiplying Equation 15 by C^{-1} on both the sides from the left Equation 16 follows.

$$\tilde{E}\tilde{X}\tilde{X}^T\tilde{E}^T\tilde{D}^T = \tilde{E}\tilde{X}Y^T \tag{16}$$

Taking transpose of the above equation we have,

$$\tilde{D}\tilde{E}\tilde{X}\tilde{X}^T\tilde{E}^T = Y\tilde{X}^T\tilde{E}^T \tag{17}$$

From part 1 of Claim 3.1, it follows that $\tilde{E}$ has full row-rank, and hence $\tilde{E}\tilde{X}\tilde{X}^T\tilde{E}^T$ is invertible. Multiplying the inverse of $\tilde{E}\tilde{X}\tilde{X}^T\tilde{E}^T$ from the right on both sides and multiplying $\tilde{E}B$ from the left on both sides of the above equation we have,

$$\tilde{E}B\tilde{D} = (\tilde{E}BY\tilde{X}^T\tilde{E}^T)(\tilde{E}\tilde{X}\tilde{X}^T\tilde{E}^T)^{-1} \tag{18}$$

This proves part one of the claim. Moreover, multiplying Equation 17 by $\tilde{D}^T$ from the right on both sides

$$\begin{aligned}\tilde{D}\tilde{E}\tilde{X}\tilde{X}^T\tilde{E}^T\tilde{D}^T &= Y\tilde{X}^T\tilde{E}^T\tilde{D}^T\\ \Rightarrow (P_{\tilde{D}}Y\tilde{X}^T(\tilde{X}\tilde{X}^T)^{-1})(\tilde{X}\tilde{X}^T)((\tilde{X}\tilde{X}^T)^{-1}\tilde{X}Y^T P_{\tilde{D}}) &= Y\tilde{X}^T((\tilde{X}\tilde{X}^T)^{-1}\tilde{X}Y^T\cdot P_{\tilde{D}})\\ \Rightarrow P_{\tilde{D}}Y\tilde{X}^T(\tilde{X}\tilde{X}^T)^{-1}\tilde{X}Y^T P_{\tilde{D}} &= Y\tilde{X}^T(\tilde{X}\tilde{X}^T)^{-1}\tilde{X}Y^T\cdot P_{\tilde{D}}\end{aligned}$$

The second line the above equation follows by substituting $\tilde{D}\tilde{E} = P_{\tilde{D}}Y\tilde{X}^T(\tilde{X}\tilde{X}^T)^{-1}$ (from part 2 of Claim 3.1). Substituting $\Sigma = Y\tilde{X}^T(\tilde{X}\tilde{X}^T)^{-1}\tilde{X}Y^T$ in the above equation we have

$$P_{\tilde{D}}\Sigma P_{\tilde{D}} = \Sigma\cdot P_{\tilde{D}}$$

Since $P_{\tilde{D}}^T = P_{\tilde{D}}$, and $\Sigma^T = \Sigma$, we also have $\Sigma P_{\tilde{D}} = P_{\tilde{D}}\Sigma$. □

4 ADDITIONAL TABLES AND PLOTS RELATED TO DENSE LAYER REPLACEMENT

4.1 PLOTS FROM SECTION 5.1

Figure 2 displays the number of parameter in the original model and the butterfly model. Figure 3 reports the results for the NLP tasks done as part of experiment in Section 5.1. Figures 4 and 5 reports the training and inference times required for the original model and the butterfly model in each of the experiments. The training and and inference times in Figures 4 and 5 are averaged over 100 runs. Figure 6 is the same as the right part of Figure 3 but here we compare the test accuracy of the original and butterfly model for the the first 20 epochs.

4.2 PLOTS FROM SECTION 5.2

Figure 7 reports the losses for the Gaussian 2, Olivetti, and Hyper data matrices.

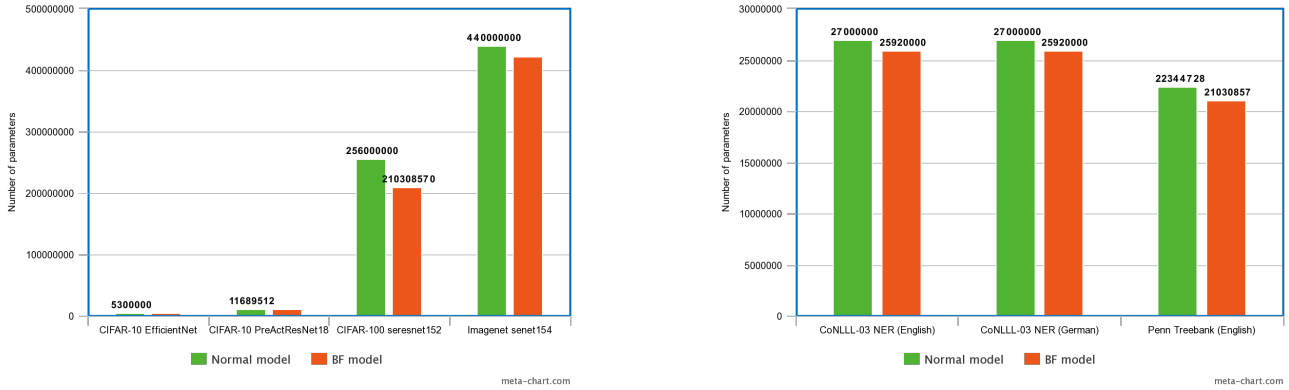

Figure 2: Total number of parameters in the original model and the butterfly model; Left: Vision data, Right: NLP

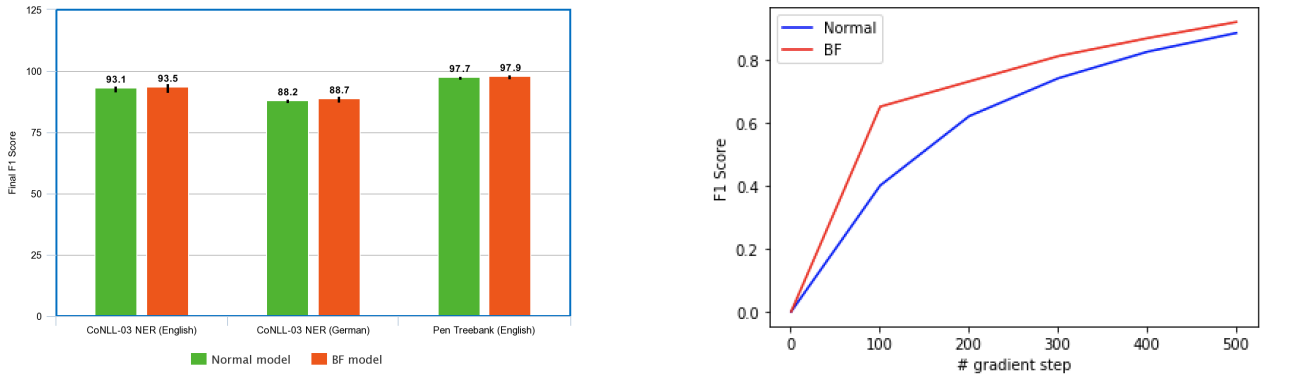

Figure 3: Left: F1 comparison in the first few epochs with different models on CoNLL-03 Named Entity Recognition (English) with the flair's Sequence Tagger, Right: Final F1 Score for different NLP models and data sets.

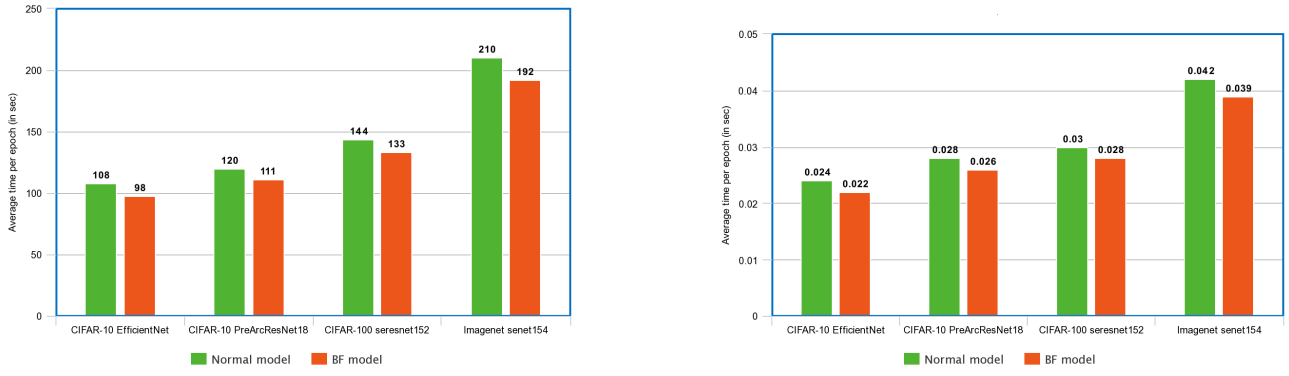

Figure 4: Training/Inference times for Vision Data; Left: Training time, Right: Inference time

5 ADDITIONAL PLOTS RELATED TO SKETCHING

In this section we state a few additional cases that were done as part of the experiment in Section 6. Figure 8 compares the test errors of the different methods in the extreme case when $k = 1$. Figure 9 compares the test errors of the different methods for various values of ℓ. Figure 10 shows the test error for $\ell = 20$ and $k = 10$ during the training phase on HS-SOD. Observe that the butterfly learned is able to surpass sparse learned after a merely few iterations. Finally Table 1 compares the test error for different values of ℓ and k.

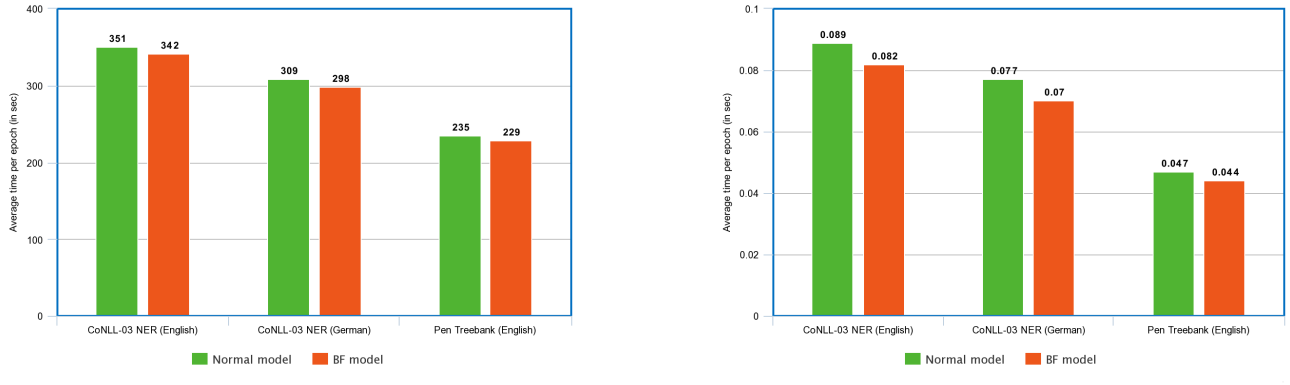

Figure 5: Training/Inference times for NLP; Left: Training time, Right: Inference time

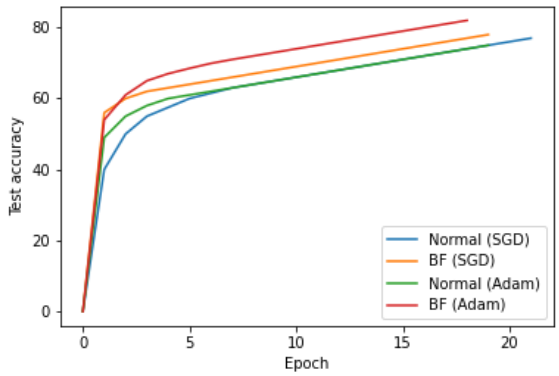

Figure 6: Comparison of test accuracy in the first 20 epochs with different models and optimizers on CIFAR-10 with PreActResNet18

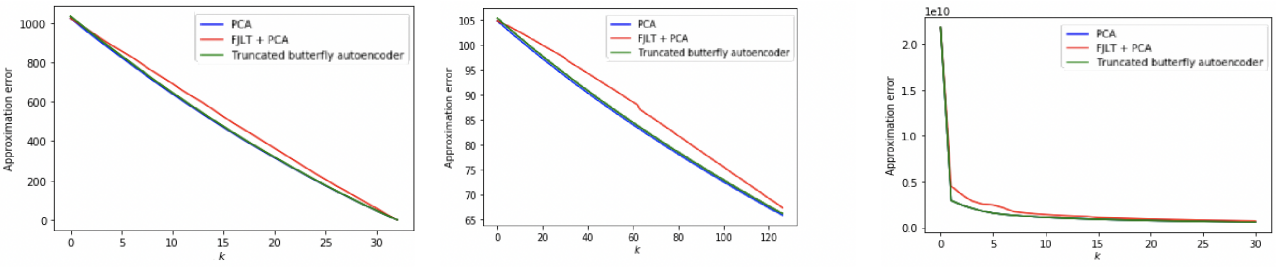

Figure 7: Approximation error on data matrix with various methods for various values of k. From left to right: Gaussian 2, Olivetti, Hyper

6 BOUND ON NUMBER OF EFFECTIVE WEIGHTS IN TRUNCATED BUTTERFLY NETWORK

A butterfly network for dimension n, which we assume for simplicity to be an integral power of 2, is $\log n$ layers deep. Let p denote the integer $\log n$. The set of nodes in the first (input) layer will be denoted here by $V^{(0)}$. They are connected to the set of n nodes $V^{(1)}$ from the next layer, and so on until the nodes $V^{(p)}$ of the output layer. Between two consecutive layers $V^{(i)}$ and $V^{(i+1)}$, there are $2n$ weights, and each node in $V^{(i)}$ is adjacent to exactly two nodes from $V^{(i+1)}$.

When truncating the network, we discard all but some set $S^{(p)} \subseteq V^{(p)}$ of at most ℓ nodes in the last layer. These nodes are

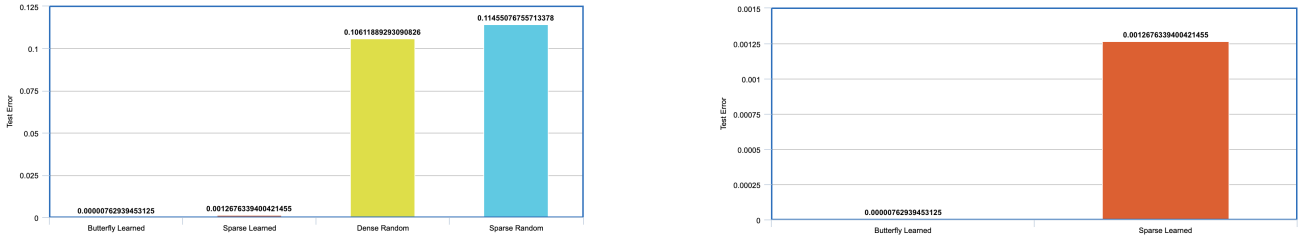

Figure 8: Test errors on HS-SOD for $\ell = 20$ and $k = 1$, zoomed on butterfly and sparse learned in the right

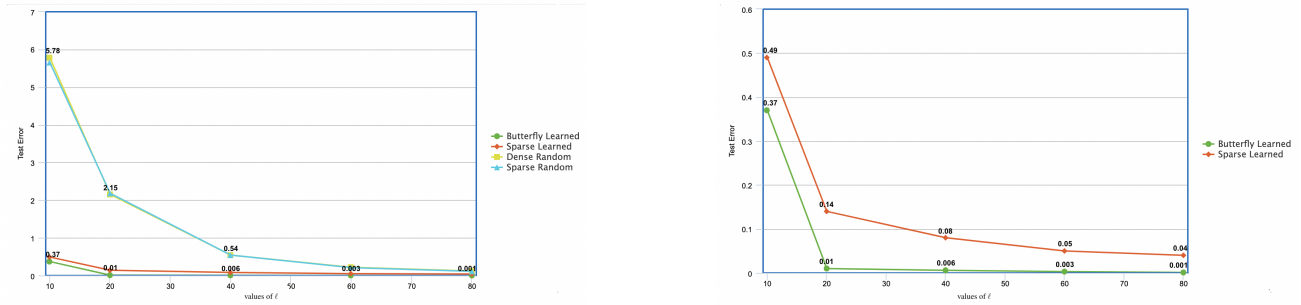

Figure 9: Test error when $k = 10$, $\ell = [10, 20, 40, 60, 80]$ on HS-SOD, zoomed on butterfly and sparse learned in the right

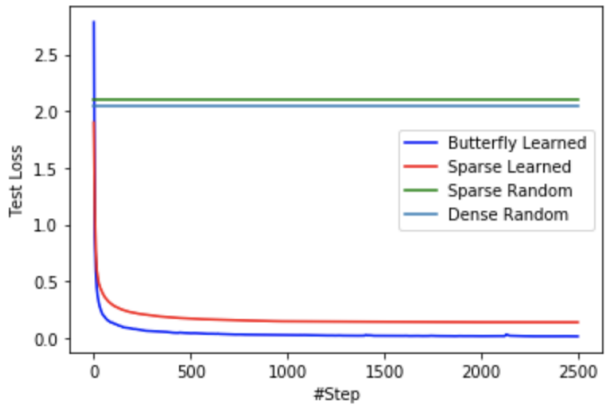

Figure 10: Test error when $k = 10$, $\ell = 20$ during the training phase on HS-SOD

connected to a subset $S^{(p-1)} \subseteq V^{(p-1)}$ of at most 2ℓ nodes from the preceding layer using at most 2ℓ weights. By induction, for all $i \geq 0$, the set of nodes $S^{(p-i)} \subseteq V^{(p-i)}$ is of size at most $2^i \cdot \ell$, and is connected to the set $S^{(p-i-1)} \subseteq V^{(p-i-1)}$ using at most $2^{i+1} \cdot \ell$ weights.

Now take $k = \lceil \log_2(n/\ell) \rceil$. By the above, the total number of weights that can participate in a path connecting some node in $S^{(p)}$ with some node in $V^{(p-k)}$ is at most $2\ell + 4\ell + \cdots + 2^k\ell \leq 4n$.

From the other direction, the total number of weights that can participate in a path connecting any node from $V^{(0)}$ with any node from $V^{(p-k)}$ is $2n$ times the number of layers in between, or more precisely:

$$2n(p-k) = 2n(\log_2 n - \lceil \log_2(n/\ell) \rceil) \leq 2n(\log_2 n - \log_2(n/\ell) + 1) = 2n(\log \ell + 1) \, .$$

The total is $2n \log \ell + 6n$, as required.

k, ℓ, Sketch	Hyper	Cifar-10	Tech
1, 5, Butterfly	**0.0008**	**0.173**	**0.188**
1, 5, Sparse	0.003	1.121	1.75
1, 5, Random	0.661	4.870	3.127
1, 10, Butterfly	**0.0002**	**0.072**	**0.051**
1, 10, Sparse	0.002	0.671	0.455
1, 10, Random	0.131	1.82	1.44
10, 10, Butterfly	**0.031**	**0.751**	**0.619**
10, 10, Sparse	0.489	6.989	7.154
10, 10, Random	5.712	26.133	18.805
10, 20, Butterfly	**0.012**	**0.470**	**0.568**
10, 20, Sparse	0.139	3.122	3.134
10, 20, Random	2.097	9.216	8.22
10, 40, Butterfly	**0.006**		**0.111**
10, 40, Sparse	0.081		0.991
10, 40, Random	0.544		3.304
20, 20, Butterfly	**0.058**		**1.38**
20, 20, Sparse	0.229		8.14
20, 20, Random	4.173		15.268
20, 40, Butterfly	**0.024**		**0.703**
20, 40, Sparse	0.247		3.441
20, 40, Random	1.334		6.848
30, 30, Butterfly	**0.027**		**1.25**
30, 30, Sparse	0.749		7.519
30, 30, Random	3.486		13.168
30, 60, Butterfly	**0.014**		**0.409**
30, 60, Sparse	0.331		2.993
30, 60, Random	2.105		5.124

Table 1: Test error for different ℓ and k