Supplementary Material

Alokendu Mazumder[1], Tirthajit Baruah[1*], Bhartendu Kumar[2*], Rishab Sharma[3], Vishwajeet Pattanaik[1]
Punit Rathore[1]
[1]Indian Institute of Science, Bengaluru, [2]TCS Research, Bengaluru
[3]Dayananda Sagar College of Engineering, Bengaluru

1. Experiment Parameters

1.1. Dataset

This paper encompasses a range of experiments conducted using the MNIST and CelebA datasets. To ensure uniformity and facilitate comparisons, all images from the MNIST dataset were resized to dimensions of 32×32 pixels. Likewise, with the CelebA dataset, images were initially center-cropped to 148×148 pixels and subsequently resized to 64×64 pixels.

1.2. Model Architecture

The encoder and decoder architectures for each experiment are detailed below. The notation $Conv_n$ and $ConvT_n$ signify a convolutional and transposed-convolutional layer with an output channel dimension of n respectively. All convolutional layers employ a 4×4 kernel size with a stride of 2 and padding of 1. FC_n denotes a fully connected network with an output dimension of n.

Table 1. Architecture of encoder and decoder for MNIST and CelebA dataset.

Datasets	MNIST	CelebA
Encoder	$x \in \mathbb{R}^{32\times32\times1}$ $\rightarrow Conv_{32} \rightarrow ReLU$ $\rightarrow Conv_{64} \rightarrow ReLU$ $\rightarrow Conv_{128} \rightarrow ReLU$ $\rightarrow Conv_{256} \rightarrow ReLU$ Flatten 1024 $\rightarrow FC_{128} \rightarrow z \in \mathbb{R}^{128}$	$x \in \mathbb{R}^{64\times64\times3}$ $\rightarrow Conv_{128} \rightarrow ReLU$ $\rightarrow Conv_{256} \rightarrow ReLU$ $\rightarrow Conv_{512} \rightarrow ReLU$ $\rightarrow Conv_{1024} \rightarrow ReLU$ Flatten 16,384 $\rightarrow FC_{256} \rightarrow z \in \mathbb{R}^{256}$
Decoder	$z \in \mathbb{R}^{128}$ FC_{8096} Reshape to $8 \times 8 \times 128$ $\rightarrow ConvT_{64} \rightarrow ReLU$ $\rightarrow ConvT_{32} \rightarrow ReLU$ $\rightarrow ConvT_{3} \rightarrow Tanh$ $\hat{x} \in \mathbb{R}^{32\times32\times1}$	$z \in \mathbb{R}^{512}$ FC_{65536} Reshape to $8 \times 8 \times 1024$ $\rightarrow ConvT_{512} \rightarrow ReLU$ $\rightarrow ConvT_{256} \rightarrow ReLU$ $\rightarrow ConvT_{128} \rightarrow ReLU$ $\rightarrow ConvT_{3} \rightarrow Tanh$ $\hat{x} \in \mathbb{R}^{64\times64\times3}$

*denotes equal contribution

1.3. Hyperparameter Settings

Our model underwent training based on the hyperparameter settings provided below.

Table 2. The hyperparameters for each experiment are elaborated in the following table. The determination of the number of epochs was guided by the aim of attaining a stage of converged reconstruction error.

Dataset	MNIST	CelebA
Batch Size	32	32
Epochs	50	100
Training Examples	60,000	16,2079
Test Examples	10,000	20,000
Dimension of Latent Space	128	128
Learning Rate	10^{-3}	10^{-3}
λ	10^{-3}	10^{-5}

2. Theoretical Analysis

In this section, we provide a detailed proof for each of the theorems introduced in our paper. Before delving into the proof explanations, we'll establish an understanding of the symbols and terms that will be employed throughout the proofs.

2.1. Notations

1. *For our proposed model:*

   - We denote the combined parameter of encoder ($\mathbf{E}$), decoder ($\mathbf{D}$) and the matrix between encoder and decoder ($\mathbf{M}$) by $\mathbf{w}$. Hence, from now onwards, $\mathbf{w}$ is the parameter set of our model.
   - $\mathbf{w}_t$ denotes the parameter of our model at t^{th} iteration.
   - $\mathbf{w}^*$ denotes the parameter of our model after convergence.
   - We denote the loss function of our model as:

   $$\mathcal{L}(\mathbf{w}) = \mathcal{L}(\mathbf{E}, \mathbf{D}, \mathbf{M}) = \underbrace{\|\mathbf{D} - (\mathbf{M}(\mathbf{E}(x))\|_2^2}_{\mathcal{L}_{mse}(\mathbf{w}) = \mathcal{L}_{mse}(\mathbf{E}, \mathbf{D}, \mathbf{M})} + \|\mathbf{M}\|_* \tag{1}$$

   or, in short hand $\mathcal{L}(\mathbf{w}) = \mathcal{L}_{mse}(\mathbf{w}) + \|\mathbf{M}\|_*$.

2. *For ADAM Optimizer:*

   - The ADAM update for our model can be written as:

   $$\mathbf{w}_{t+1} = \mathbf{w}_t - \alpha(V_t^{1/2} + diag(\epsilon\mathbb{I}))^{-1}\mathbf{m}_t \tag{2}$$

   where , $\mathbf{m}_t = \beta_1\mathbf{m}_{t-1} + (1-\beta_1)\nabla\mathcal{L}(\mathbf{w}_t)$, $\mathbf{v}_t = \beta_2\mathbf{v}_{t-1} + (1-\beta_2)(\nabla\mathcal{L}(\mathbf{w}_t))^2$, $V_t = diag(\mathbf{v}_t)$ is a diagonal matrix, $\beta_1, \beta_2 \in (0,1)$ and $\epsilon > 0$.
   - $\alpha > 0$, is the constant step size.
   - One can clearly see from the equation $\mathbf{v}_t = \beta_2\mathbf{v}_{t-1} + (1-\beta_2)(\nabla\mathcal{L}(\mathbf{w}_t))^2$ that $\mathbf{v}_t$ will be always non-negative. Also, the term ϵ in $diag(\epsilon\mathbb{I})$ will always keep the matrix (diagonal matrix) $(V_t^{1/2} + diag(\epsilon\mathbb{I}))^{-1}$ positive definite (PD).
   - From now onward, to avoid using too much terms in derivation, we will denote the matrix $(V_t^{1/2} + diag(\epsilon\mathbb{I}))^{-1}$ as $\mathbf{A}_t$.
   - Hence, the ADAM update in Eq.(2) will now look like this.

   $$\mathbf{w}_{t+1} = \mathbf{w}_t - \alpha\mathbf{A}_t\mathbf{m}_t \tag{3}$$

   - We will denote the gradient of the loss function as $\nabla\mathcal{L}(\mathbf{w})$ for simplicity in rest of our proof.

2.2. Proofs

Theorem 1. *Let the loss function $\mathcal{L}(\boldsymbol{E}, \boldsymbol{D}, \boldsymbol{M})$ be $K-$Lipchitz and let $\gamma < \infty$ be an upper bound on the norm of the gradient of $\mathcal{L}$. Then the following holds for the deterministic version (when batch size = total dataset) of Algorithm (1):*

For any $\sigma > 0$ if we let $\alpha = \sqrt{2(\mathcal{L}(\boldsymbol{E}_0, \boldsymbol{D}_0, \boldsymbol{M}_0) - \mathcal{L}(\boldsymbol{E}^*, \boldsymbol{D}^*, \boldsymbol{M}^*))/K\delta^2 T}$, then there exists a natural number $T(\sigma, \delta)$ (depends on σ and δ) such that $\|\mathcal{L}(\boldsymbol{E}_t, \boldsymbol{D}_t, \boldsymbol{M}_t)\|_2 \leq \sigma$ for some $t \geq T(\sigma, \delta)$, where $\delta^2 = \frac{\gamma^2}{\epsilon^2}$.

Proof. We aim to prove Theorem (1) with contradiction. Let $\|\nabla\mathcal{L}(\mathbf{w}_t)\|_2 > \sigma > 0$ for all $t \in \{1, 2, \dots\}$. Using Lipchitz continuity, we can write:

$$\begin{aligned}\mathcal{L}(\mathbf{w}_{t+1}) - \mathcal{L}(\mathbf{w}_t) \leq\ & \nabla\mathcal{L}(\mathbf{w}_t)^T(\mathbf{w}_{t+1} - \mathbf{w}_t) + \frac{K}{2}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2 \\ \leq & -\alpha\nabla\mathcal{L}(\mathbf{w}_t)^T(\mathbf{A}_t\mathbf{m}_t) + \frac{K}{2}\alpha^2\|\mathbf{A}_t\mathbf{m}_t\|_2^2 \end{aligned} \tag{4}$$

One can clearly see that $\mathbf{A}_t$ is *positive definite* (PD). From here, we will find an upper bound and lower bound on the last and first terms of RHS of Eq.(4), respectively.

Consider the term $\|\mathbf{A}_t\mathbf{m}_t\|_2$. We have $\lambda_{max}(\mathbf{A}_t) \leq \frac{1}{\epsilon + \min\limits_{1\leq i\leq|\mathbf{v}_t|}\sqrt{(\mathbf{v}_t)_i}}$. Further we note that recursion of $\mathbf{v}_t$ can be solved as $\mathbf{v}_t = (1-\beta_2)\sum_{j=1}^{t}\beta_2^{t-j}(\nabla\mathcal{L}(\mathbf{w}_j))^2$. Now we define $\rho_t = \min\limits_{1\leq j\leq t, 1\leq k\leq|\mathbf{v}_t|}(\nabla\mathcal{L}(\mathbf{w}_j)^2)_k$. This gives us the following:

$$\lambda_{max}(\mathbf{A}_t) \leq \frac{1}{\epsilon + \sqrt{(1-\beta_2^t)\rho_t}} \tag{5}$$

The equation of $\mathbf{m}_t$ without recursion is $\mathbf{m}_t = (1-\beta_1)\sum_{j=1}^{t}\beta_1^{t-j}\nabla\mathcal{L}(\mathbf{w}_j)$. Let us define $\gamma_t = \max\limits_{1\leq j\leq t}\|\nabla\mathcal{L}(\mathbf{w}_j)\|$ then by using triangle inequality, we have $\|\mathbf{m}_t\|_2 \leq (1-\beta_1^t)\gamma_t$. We can rewrite $\|\mathbf{A}_t\mathbf{m}_t\|_2$ as:

$$\|\mathbf{A}_t\mathbf{m}_t\|_2 \leq \frac{(1-\beta_1^t)\gamma_t}{\epsilon + \sqrt{\rho_t(1-\beta_2^t)}} \leq \frac{(1-\beta_1^t)\gamma_t}{\epsilon} \leq \frac{\gamma_t}{\epsilon} \tag{6}$$

Taking $\gamma_{t-1} = \gamma_t = \gamma$ and plugging Eq.(6) in Eq.(4):

$$\mathcal{L}(\mathbf{w}_{t+1}) - \mathcal{L}(\mathbf{w}_t) \leq -\alpha\nabla\mathcal{L}(\mathbf{w}_t)^T(\mathbf{A}_t\mathbf{m}_t) \ + \frac{K}{2}\alpha^2\frac{\gamma^2}{\epsilon^2} \tag{7}$$

Now, we will investigate the term $\nabla\mathcal{L}(\mathbf{w}_t)^T(\mathbf{A}_t\mathbf{m}_t)$ separately, *i.e.* we will find a lower bound on this term. To analyze this, we define the following sequence of functions:

$$\begin{aligned} P_j - \beta_1 P_{j-1} =\ & \nabla\mathcal{L}(\mathbf{w}_t)^T\mathbf{A}_t(\mathbf{m}_j - \beta_1\mathbf{m}_{j-1}) \\ =\ & (1-\beta_1)\nabla\mathcal{L}(\mathbf{w}_t)^T(\mathbf{A}_t\nabla\mathcal{L}(\mathbf{w}_j)) \end{aligned}$$

At $j = t$, we have:

$$P_t - \beta_1 P_{t-1} \geq\ (1-\beta_1)\|\nabla\mathcal{L}(\mathbf{w}_t)\|_2^2\lambda_{min}(\mathbf{A}_t)$$

Let us (again) define $\gamma_{t-1} = \max\limits_{1\leq j\leq t-1}\|\nabla\mathcal{L}(\mathbf{w}_j)\|_2$, and $\forall j \in \{1, 2, \dots t-1\}$:

$$P_j - \beta_1 P_{j-1} \geq\ -(1-\beta_1)\|\nabla\mathcal{L}(\mathbf{w}_t)\|_2\gamma_{t-1}\lambda_{max}(\mathbf{A}_t)$$

Now, we note the following identity:

$$P_t - \beta_1^t P_0 = \sum_{j=1}^{t-1}\beta_1^j(P_{t-j} - \beta_1 P_{t-j-1})$$

Now, we use the lower bounds proven on $P_j - \beta_1 P_{j-1}\ \forall j \in \{1, 2, \dots t-1\}$ and $P_t - \beta_1 P_{t-1}$ to lower bound the above sum as:

$$\begin{aligned}
P_t - \beta_1^t P_0 \geq\ & (1-\beta_1)\|\nabla\mathcal{L}(\mathbf{w}_t)\|_2^2\lambda_{min}(\mathbf{A}_t) - (1-\beta_1)\|\nabla\mathcal{L}(\mathbf{w}_t)\|_2\gamma_{t-1}\lambda_{max}(\mathbf{A}_t)\sum_{j=0}^{t-1}\beta_1^j \\
\geq\ & (1-\beta_1)\|\nabla\mathcal{L}(\mathbf{w}_t)\|_2^2\lambda_{min}(\mathbf{A}_t) - (\beta_1-\beta_1^t)\|\nabla\mathcal{L}(\mathbf{w}_t)\|_2\gamma_{t-1}\lambda_{max}(\mathbf{A}_t) \\
\geq\ & \|\nabla\mathcal{L}(\mathbf{w}_t)\|_2^2\left((1-\beta_1)\lambda_{min}(\mathbf{A}_t) - \frac{(\beta_1-\beta_1^t)\gamma_{t-1}\lambda_{max}(\mathbf{A}_t)}{\|\nabla\mathcal{L}(\mathbf{w}_t)\|_2}\right) \\
\geq\ & \|\nabla\mathcal{L}(\mathbf{w}_t)\|_2^2\left((1-\beta_1)\lambda_{min}(\mathbf{A}_t) - \frac{(\beta_1-\beta_1^t)\gamma_{t-1}\lambda_{max}(\mathbf{A}_t)}{\sigma}\right) \quad \text{(From Contradiction)}
\end{aligned} \tag{8}$$

The inequality in Eq.(8) will be maintained as the term $\left((1-\beta_1)\lambda_{min}(\mathbf{A}_t) - \frac{(\beta_1-\beta_1^t)\gamma_{t-1}\lambda_{max}(\mathbf{A}_t)}{\sigma}\right)$ is lower bounded by some positive constant c. We will show this later in **Extension 1**.

Hence, we let $\left((1-\beta_1)\lambda_{min}(\mathbf{A}_t) - \frac{(\beta_1-\beta_1^t)\gamma_{t-1}\lambda_{max}(\mathbf{A}_t)}{\sigma}\right) \geq c > 0$ and put $P_0 = 0$ (from definition and initial conditions) in the above equation and get:

$$P_t = \nabla\mathcal{L}(\mathbf{w}_t)^T(\mathbf{A}_t\mathbf{m}_t) \geq\ c\|\nabla\mathcal{L}(\mathbf{w}_t)\|_2^2 \tag{9}$$

Now we are done with computing the bounds on the terms in Eq.(4). Hence, we combine Eq.(9) with Eq.(7) to get:

$$\mathcal{L}(\mathbf{w}_{t+1}) - \mathcal{L}(\mathbf{w}_t) \leq\ -\alpha c\|\nabla\mathcal{L}(\mathbf{w}_t)\|_2^2 + \frac{K}{2}\alpha^2\frac{\gamma^2}{\epsilon^2}$$

Let $\delta^2 = \frac{\gamma^2}{\epsilon^2}$ for simplicity. We have:

$$\begin{aligned}
\mathcal{L}(\mathbf{w}_{t+1}) - \mathcal{L}(\mathbf{w}_t) \leq\ & -\alpha c\|\nabla\mathcal{L}(\mathbf{w}_t)\|_2^2 + \frac{K}{2}\alpha^2\delta^2 \\
\alpha c\|\nabla\mathcal{L}(\mathbf{w}_t)\|_2^2 \leq\ & \mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_{t+1}) + \frac{K}{2}\alpha^2\delta^2 \\
\|\nabla\mathcal{L}(\mathbf{w}_t)\|_2^2 \leq\ & \frac{\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_{t+1})}{\alpha c} + \frac{K\alpha\delta^2}{2c}
\end{aligned} \tag{10}$$

From Eq.(10), we have the following inequalities:

$$\begin{cases}
\|\nabla\mathcal{L}(\mathbf{w}_0)\|_2^2 \leq \dfrac{\mathcal{L}(\mathbf{w}_0) - \mathcal{L}(\mathbf{w}_1)}{\alpha c} + \dfrac{K\alpha\delta^2}{2c} \\
\|\nabla\mathcal{L}(\mathbf{w}_1)\|_2^2 \leq \dfrac{\mathcal{L}(\mathbf{w}_1) - \mathcal{L}(\mathbf{W}_2)}{\alpha c} + \dfrac{K\alpha\delta^2}{2c} \\
\vdots \\
\|\nabla\mathcal{L}(\mathbf{w}_{T-1})\|_2^2 \leq \dfrac{\mathcal{L}(\mathbf{w}_{T-1}) - \mathcal{L}(\mathbf{w}_t)}{\alpha c} + \dfrac{K\alpha\delta^2}{2c}
\end{cases}$$

Summing up all the inequalities presented above, we obtain:

$$\sum_{t=0}^{T-1}\|\nabla\mathcal{L}(\mathbf{w}_t)\|_2^2 \leq \frac{\mathcal{L}(\mathbf{w}_0) - \mathcal{L}(\mathbf{w}_t)}{\alpha c} + \frac{K\alpha\delta^2 T}{2c}$$

The inequality remains valid if we substitute $\|\nabla\mathcal{L}(\mathbf{w}_t)\|_2^2$ with $\min\limits_{0\leq t\leq T-1}\|\nabla\mathcal{L}(\mathbf{w}_t)\|_2^2$ within the summation on the left-hand

side (LHS).

$$\min_{0\le t\le T-1}\|\nabla\mathcal{L}(\mathbf{w}_t)\|_2^2 T \le \frac{\mathcal{L}(\mathbf{w}_0)-\mathcal{L}(\mathbf{w}^*)}{\alpha c}+\frac{K\alpha\delta^2 T}{2c}$$
$$\min_{0\le t\le T-1}\|\nabla\mathcal{L}(\mathbf{w}_t)\|_2^2 \le \frac{\mathcal{L}(\mathbf{w}_0)-\mathcal{L}(\mathbf{w}^*)}{\alpha c T}+\frac{K\alpha\delta^2}{2c}$$
$$\min_{0\le t\le T-1}\|\nabla\mathcal{L}(\mathbf{w}_t)\|_2^2 \le \frac{1}{\sqrt{T}}\left(\frac{\mathcal{L}(\mathbf{w}_0)-\mathcal{L}(\mathbf{w}^*)}{cb}+\frac{K\delta^2 b}{2c}\right)$$

where $b=\alpha\sqrt{T}$. We set $b=\sqrt{2(\mathcal{L}(\mathbf{w}_0)-\mathcal{L}(\mathbf{w}^*)\delta^2)/K\delta^2}$, and we have:

$$\min_{0\le t\le T-1}\|\nabla\mathcal{L}(\mathbf{w}_t)\|_2 \le \left(\frac{2K\delta^2}{T}(\mathcal{L}(\mathbf{w}_0)-\mathcal{L}(\mathbf{w}^*))\right)^{\frac{1}{4}}$$

When $T\ge\left(\frac{2K\delta^2}{\sigma^4}(\mathcal{L}(\mathbf{w}_0)-\mathcal{L}(\mathbf{w}^*))\right)=T(\sigma,\delta)$, we will have $\min_{0\le t\le T-1}\|\nabla\mathcal{L}(\mathbf{w}_t)\|_2\le\sigma$ which will contradict the assumption, *i.e.* ($\|\nabla\mathcal{L}(\mathbf{w}_t)\|_2>\sigma$ for all $t\in\{1,2,\dots\}$). Hence, completing the proof.
From the above analysis, one can clearly see that the convergence rate is $\mathcal{O}(1/T^{1/4})$ □

Theorem 2. *Given any set of i.i.d x, $x_1,x_2,\dots,x_N\in\mathbb{R}^l$, we denote $d_{max}^{\boldsymbol{E}^*\boldsymbol{M}^*}=\max_{1\le j\le N} d^{\boldsymbol{E}^*\boldsymbol{M}^*}(x,x_j)$ and $d_{min}^{\boldsymbol{E}^*\boldsymbol{M}^*}=\min_{1\le j\le N} d^{\boldsymbol{E}^*\boldsymbol{M}^*}(x,x_j)$, then we always have the conditional probability:*

$$\mathbb{P}\left(\frac{d_{max}^{\boldsymbol{E}^*\boldsymbol{M}^*}-d_{min}^{\boldsymbol{E}^*\boldsymbol{M}^*}}{d_{min}^{\boldsymbol{E}^*\boldsymbol{M}^*}}\ge\Theta(\mathcal{D},\lambda)\middle|\lambda>0\right)=1 \tag{11}$$

where $d^{\boldsymbol{E}^*\boldsymbol{M}^*}(x,x_j)=\frac{\|\boldsymbol{M}^*(\boldsymbol{E}^*(x))-\boldsymbol{M}^*(\boldsymbol{E}^*(x_i))\|_2}{rank(\boldsymbol{M}^*)}$, $\mathcal{D}$ denotes the training dataset and $\Theta(\mathcal{D},\lambda)$ depends on the training set and regularization penalty parameter λ.

Proof. As $\mathbf{w}^*$ is learned from Algorithm (1), we always have:

$$\mathcal{L}(\mathbf{w}^*) \quad\le\quad \mathcal{L}(\mathbf{w}_0)$$

where, $\mathcal{L}(\mathbf{w}_0)$ is loss of our model at 0^{th} epoch. Hence,

$$\begin{aligned}
\mathcal{L}_{mse}(\mathbf{w}^*)+\lambda\|\mathbf{M}^*\|_* &\le \mathcal{L}_{mse}(\mathbf{w}_0)+\lambda\|\mathbf{M}_0\|_*\\
\lambda\|\mathbf{M}^*\|_* &\le \mathcal{L}_{mse}(\mathbf{w}_0)-\mathcal{L}_{mse}(\mathbf{w}^*)+\lambda\|\mathbf{M}_0\|_*\\
\|\mathbf{M}^*\|_* &\le \frac{1}{\lambda}\left(\mathcal{L}_{mse}(\mathbf{w}_0)-\mathcal{L}_{mse}(\mathbf{w}^*)\right)+\|\mathbf{M}_0\|_*\\
\|\mathbf{M}^*\|_* &\le \frac{1}{\lambda}\left(c_1-c_2\right)+c_3
\end{aligned} \tag{12}$$

where, $c_1=\mathcal{L}_{mse}(\mathbf{w}_0)$, $c_2=\mathcal{L}_{mse}(\mathbf{w}^*)$, and $c_3=\|\mathbf{M}_0\|_*$. Now, from Eq.(12) we can estimate an upperbound on the rank of matrix $\mathbf{M}^*$:

$$rank(\mathbf{M}^*) \quad\le\quad c\left(\frac{1}{\lambda}\left(c_1-c_2\right)+c_3\right) \quad \text{(where } c\in\mathbb{R}^+\text{)} \tag{13}$$

Using the definition of $d_{max}^{\mathbf{E}^*\mathbf{M}^*}$ and $d_{min}^{\mathbf{E}^*\mathbf{M}^*}$, we have:

$$
\begin{aligned}
\frac{d_{max}^{\mathbf{E}^*\mathbf{M}^*} - d_{min}^{\mathbf{E}^*\mathbf{M}^*}}{d_{min}^{\mathbf{E}^*\mathbf{M}^*}} &= \frac{\underset{i\in[n]}{max}\frac{\|\mathbf{M}^*(\mathbf{E}^*(x))-\mathbf{M}^*(\mathbf{E}^*(x_i))\|_2}{rank(\mathbf{M}^*)} - \underset{i\in[n]}{min}\frac{\|\mathbf{M}^*(\mathbf{E}^*(x))-\mathbf{M}^*(\mathbf{E}^*(x_i))\|_2}{rank(\mathbf{M}^*)}}{\underset{i\in[n]}{min}\frac{\|\mathbf{M}^*(\mathbf{E}^*(x))-\mathbf{M}^*(\mathbf{E}^*(x_i))\|_2}{rank(\mathbf{M}^*)}} \\
&= \frac{\underset{i\in[n]}{max}\frac{\|\mathbf{M}^*(\mathbf{E}^*(x))-\mathbf{M}^*(\mathbf{E}^*(x_i))\|_2}{rank(\mathbf{M}^*)}}{\underset{i\in[n]}{min}\frac{\|\mathbf{M}^*(\mathbf{E}^*(x))-\mathbf{M}^*(\mathbf{E}^*(x_i))\|_2}{rank(\mathbf{M}^*)}} - 1 \\
&\geq \frac{\underset{i\in[n]}{max}\frac{\|\mathbf{M}^*(\mathbf{E}^*(x))-\mathbf{M}^*(\mathbf{E}^*(x_i))\|_2}{c\left(\frac{1}{\lambda}(c_1-c_2)+c_3\right)}}{\underset{i\in[n]}{min}\frac{\|\mathbf{M}^*(\mathbf{E}^*(x))-\mathbf{M}^*(\mathbf{E}^*(x_i))\|_2}{rank(\mathbf{M}^*)}} - 1 \quad \text{(Using Eq.(15))} \\
&\geq \frac{L(\mathcal{D})}{c\left(\frac{1}{\lambda}(c_1-c_2)+c_3\right)} \quad \left(here, L(\mathcal{D}) = \frac{\underset{i\in[n]}{max}\|\mathbf{M}^*(\mathbf{E}^*(x)) - \mathbf{M}^*(\mathbf{E}^*(x_i))\|_2}{\underset{i\in[n]}{min}\frac{\|\mathbf{M}^*(\mathbf{E}^*(x))-\mathbf{M}^*(\mathbf{E}^*(x_i))\|_2}{rank(\mathbf{M}^*)}}\right) \\
&\geq \frac{\lambda L(\mathcal{D})}{c(c_1-c_2)+\lambda c c_3} \\
&\geq \frac{\lambda L(\mathcal{D})}{cc_1} = \Theta(\lambda, \mathcal{D}) > 0 \quad \text{when } \lambda > 0
\end{aligned}
$$

hence, completing the proof. □

Proposition 1. *The rank of the latent space follows* $\mathcal{O}(1/\lambda)$.

Proof. Let $\mathbf{E}^*$ denote the trained encoder of our model and let $x \in \mathbb{R}^{m\times n\times c}$ be an image with dimension $m \times n$ and c number of channels. Let $y = \mathbf{E}^*(x)$, then we can define the latent space of our model (LoRAE) as:

$$z = \mathbf{M}^* y = \mathbf{M}^*(\mathbf{E}^*(x)) \tag{14}$$

We define the rank of the latent space as the number of non-zero singular values of the covariance matrix of latent space, i.e $\mathbb{E}_{\mathcal{D}}[zz^T]$. We can write:

$$\mathbb{E}_{\mathcal{D}}[zz^T] = \mathbb{E}_{\mathcal{D}}[\mathbf{M}^* yy^T \mathbf{M}^{*T}] \tag{15}$$

Eq.(13) from Theorem 2 states that:

$$rank(\mathbf{M}^*) \leq \frac{c}{\lambda}(c_1-c_2) + cc_3 \quad (\text{where } c \in \mathbb{R}^+)$$

As $\mathbf{M}^*$ is deterministic in Eq.(15), the covariance matrix can be re-written as $\mathbf{M}^*\mathbb{E}_{\mathcal{D}}\left[yy^T\right]\mathbf{M}^{*T}$. An upper bound on the rank of $\mathbf{M}^*\mathbb{E}_{\mathcal{D}}\left[xx^T\right]\mathbf{M}^{*T}$ is the upper bound on the rank of $\mathbf{M}^*$. Thus, from Eq.(13) of Theorem 2, this analysis gives an upper bound on the rank of latent space as $\mathcal{O}(1/\lambda)$.

□

Extention 1. *The term* $\left((1-\beta_1)\lambda_{min}(\boldsymbol{A}_t) - \frac{(\beta_1-\beta_1^t)\gamma_{t-1}\lambda_{max}(\boldsymbol{A}_t)}{\sigma}\right)$ *from Eq.(8) is always non-negetive.*

Proof. We can construct a lower bound on $\lambda_{min}(\mathbf{A}_t)$ and an upper bound on $\lambda_{min}(\mathbf{A}_t)$ as follows:

$$\lambda_{min}(\mathbf{A}_t) \geq \frac{1}{\epsilon + \sqrt{\underset{1\leq j\leq|\mathbf{v}_t|}{max}(\mathbf{v}_t)_j}} \tag{16}$$

$$\lambda_{max}(\mathbf{A}_t) \leq \frac{1}{\epsilon + \sqrt{\underset{1\leq j\leq|\mathbf{v}_t|}{min}(\mathbf{v}_t)_j}} \tag{17}$$

We remember that $\mathbf{v}_t$ can be rewritten as $\mathbf{v}_t = \beta_2\mathbf{v}_{t-1} + (1-\beta_2)(\nabla\mathcal{L}(\mathbf{w}_t))^2$, solving this recursion and defining $\rho_t = \underset{1\leq j\leq t,1\leq k\leq|\mathbf{v}_t|}{min}(\nabla\mathcal{L}(\mathbf{w}_j)^2)_k$ and taking $\gamma_{t-1} = \gamma_t = \gamma$ we have:

$$\lambda_{min}(\mathbf{A}_t) \geq \frac{1}{\epsilon + \sqrt{(1-\beta_2^t)\gamma^2}}$$

$$\lambda_{max}(\mathbf{A}_t) \leq \frac{1}{\epsilon + \sqrt{(1-\beta_2^t)\rho_t}}$$

Where, $\gamma_{t-1} = \underset{1\leq j\leq t-1}{max}\|\nabla\mathcal{L}(\mathbf{w}_j)\|_2$, and $\forall j \in \{1,2,\dots t-1\}$. Setting $\rho_t = 0$, we can rewrite the term $\left((1-\beta_1)\lambda_{min}(\mathbf{A}_t) - \frac{(\beta_1-\beta_1^t)\gamma_{t-1}\lambda_{max}(\mathbf{A}_t)}{\sigma}\right)$ as:

$$\begin{aligned}
\left((1-\beta_1)\lambda_{min}(\mathbf{A}_t) - \frac{(\beta_1-\beta_1^t)\gamma_{t-1}\lambda_{max}(\mathbf{A}_t)}{\sigma}\right) &\geq \left(\frac{(1-\beta_1)}{\epsilon+\gamma\sqrt{(1-\beta_2^t)}} - \frac{(\beta_1-\beta_1^t)\gamma}{\epsilon\sigma}\right) \\
&\geq \frac{\epsilon\sigma(1-\beta_1) - \gamma(\beta_1-\beta_1^t)(\epsilon+\gamma\sqrt{(1-\beta_2^t)})}{\epsilon\sigma(\epsilon+\gamma\sqrt{(1-\beta_2^t)})} \\
&\geq \gamma(\beta_1-\beta_1^t)\frac{\epsilon\left(\frac{\sigma(1-\beta_1)}{\gamma(\beta_1-\beta_1^t)} - 1\right) - \gamma\sqrt{(1-\beta_2^t)}}{\epsilon\sigma(\epsilon+\gamma\sqrt{(1-\beta_2^t)})} \\
&\geq \gamma(\beta_1-\beta_1^t)\left(\frac{\sigma(1-\beta_1)}{\gamma(\beta_1-\beta_1^t)} - 1\right)\frac{\epsilon - \left(\frac{\gamma\sqrt{(1-\beta_2^t)}}{\frac{(1-\beta_1\sigma)}{(\beta_1-\beta_1^t)\gamma}-1}\right)}{\epsilon\sigma(\epsilon+\gamma\sqrt{(1-\beta_2^t)})}
\end{aligned} \tag{18}$$

By definition $\beta_1 \in (0,1)$ and hence $(\beta_1-\beta_1^t) \in (0,\beta_1)$. This implies that $\frac{(1-\beta_1)\sigma}{(\beta_1-\beta_1^t)\gamma} > \frac{(1-\beta_1)\sigma}{\beta_1\gamma} > 1$ where the last inequality follows due to the choice of σ as stated in the beginning of this theorem. This allows us to define a constant $\frac{(1-\beta_1)\sigma}{\beta_1\gamma} - 1 := \psi_1 > 0$ such that $\frac{(1-\beta_1)\sigma}{(\beta_1-\beta_1^t)\gamma} - 1 > \psi_1$. Similarly, our definition of delta allows us to define another constant $\psi_2 > 0$ to get:

$$\left(\frac{\gamma\sqrt{(1-\beta_2^t)}}{\frac{(1-\beta_1\sigma)}{(\beta_1-\beta_1^t)\gamma}-1}\right) < \frac{\gamma}{\psi_1} = \epsilon - \psi_2 \tag{19}$$

Putting Eq.(19) in Eq.(18), we get:

$$\left((1-\beta_1)\lambda_{min}(\mathbf{A}_t) - \frac{(\beta_1-\beta_1^t)\gamma_{t-1}\lambda_{max}(\mathbf{A}_t)}{\sigma}\right) \geq \left(\frac{\gamma(\beta_1-\beta_1^2)\psi_1\psi_2}{\epsilon\sigma(\epsilon+\sigma)}\right) = c > 0$$

□