Appendices

A1 Supplementary Material

A1 .1. Supplementary Proofs

Lemma 1. *If assumption 1 is true, and encoding activation function $s_e(.)$ has first derivative in $[0,1]$, then $\partial \mathcal{J}_{AE}/\partial b_{e_j} \in [-2\sigma_r\sqrt{n}\|\mathbf{W}_j\|, 2\sigma_r\sqrt{n}\|\mathbf{W}_j\|]$.*

Proof. For squared loss function $\mathcal{J}_{AE}$,

$$\frac{\partial \mathcal{J}_{AE}}{\partial b_{e_j}} = 2\mathbb{E}_{\mathbf{x}}\left[\frac{\partial s_e(a_j)}{\partial a_j}\left(\mathbf{x} - \mathbf{W}^T s_e(\mathbf{W}\mathbf{x} + \mathbf{b}_e)\right)^T \mathbf{W}_j\right] = 2\mathbb{E}_{\mathbf{x}}\left[\frac{\partial s_e(a_j)}{\partial a_j}\mathbf{r}_{\mathbf{x}}^T\mathbf{W}_j\right] \tag{15}$$

where $a_j = \mathbf{W}_j^T\mathbf{x} + b_j$. Since $\frac{\partial s_e(a_j)}{\partial a_j} \in [0,1]$,

$$\mathbb{E}_{\mathbf{x}}\left[\frac{\partial s_e(a_j)}{\partial a_j}\mathbf{r}_{\mathbf{x}}^T\mathbf{W}_j\right] \leq \mathbb{E}_{\mathbf{x}}\left[\frac{\partial s_e(a_j)}{\partial a_j}\|\mathbf{r}_{\mathbf{x}}\|\|\mathbf{W}_j\|\right] \leq \|\mathbf{W}_j\|.\mathbb{E}_{\mathbf{x}}\left[\|\mathbf{r}_{\mathbf{x}}\|\right] \tag{16}$$

Let $r_{\mathbf{x}}$ denote any one of the elements of $\mathbf{r}_{\mathbf{x}}$. Since each element of $\mathbf{r}_{\mathbf{x}}$ is *i.i.d.* from assumption 1 and $\mathbf{r}_{\mathbf{x}} \in \mathbb{R}^n$, using Jensen's inequality, $\mathbb{E}_{\mathbf{x}}\left[\|\mathbf{r}_{\mathbf{x}}\|_2\right] \leq \sqrt{n\mathbb{E}_{\mathbf{x}}[r_{\mathbf{x}}^2]} = \sqrt{n}\sigma_r$. Thus,

$$\mathbb{E}_{\mathbf{x}}\left[\frac{\partial s_e(a_j)}{\partial a_j}\mathbf{r}_{\mathbf{x}}^T\mathbf{W}_j\right] \leq \sqrt{n}\sigma_r\|\mathbf{W}_j\| \tag{17}$$

which leads to $\frac{\partial \mathcal{J}_{AE}}{\partial b_{e_j}} \leq 2\sigma_r\sqrt{n}\|\mathbf{W}_j\|$. We can similarly prove in the other direction get the desired bound. $\square$

Theorem 1. *Let $\{\mathbf{W}^t \in \mathbb{R}^{m\times n}, \mathbf{b}_e^t \in \mathbb{R}^m\}$ be the parameters of a regularized auto-encoder ($\lambda > 0$)*

$$\mathcal{J}_{RAE} = \mathcal{J}_{AE} + \lambda\mathcal{R}(\mathbf{W}, \mathbf{b}_e) \tag{18}$$

*at training iteration t with regularization term $\mathcal{R}(\mathbf{W}, \mathbf{b}_e)$, activation function $s_e(.)$ and define pre-activation $a_j^t = \mathbf{W}_j^t\mathbf{x} + b_{e_j}^t$ (thus $h_j^t = s_e(a_j^t)$). **If** $\lambda\frac{\partial \mathcal{R}}{\partial b_{e_j}} > 2\sigma_r\sqrt{n}\|\mathbf{W}_j\|$, where $j \in \{1, 2, \ldots, m\}$, **then** updating $\{\mathbf{W}^t, \mathbf{b}_e^t\}$ along the negative gradient of $\mathcal{J}_{RAE}$, results in $\mathbb{E}_{\mathbf{x}}[a_j^{t+1}] < \mathbb{E}_{\mathbf{x}}[a_j^t]$ **and** $\mathrm{Var}[a_j^{t+1}] = \|\mathbf{W}_j^{t+1}\|^2$ **for** all $t \geq 0$.*

Proof. At iteration $t+1$,

$$a_j^{t+1} = a_j^t - \eta\frac{\partial \mathcal{J}_{RAE}}{\partial \mathbf{W}_j}\mathbf{x} - \eta\frac{\partial \mathcal{J}_{RAE}}{\partial b_{e_j}} \tag{19}$$

for any step size η. Expanding $\mathcal{J}_{RAE}$, we get,

$$a_j^{t+1} = a_j^t - \eta\frac{\partial \mathcal{J}_{AE}}{\partial \mathbf{W}_j}\mathbf{x} - \eta\frac{\partial \mathcal{J}_{AE}}{\partial b_{e_j}} - \eta\lambda\frac{\partial \mathcal{R}}{\partial \mathbf{W}_j}\mathbf{x} - \eta\lambda\frac{\partial \mathcal{R}}{\partial b_{e_j}} \tag{20}$$

Thus taking expectation over $\mathbf{x}$ on both sides we get,

$$\mathbb{E}_{\mathbf{x}}\left[a_j^{t+1}\right] = \mathbb{E}_{\mathbf{x}}\left[a_j^t\right] - \eta\frac{\partial \mathcal{J}_{AE}}{\partial b_{e_j}} - \eta\lambda\frac{\partial \mathcal{R}}{\partial b_{e_j}} \tag{21}$$

Notice the terms containing $\frac{\partial \mathcal{J}_{AE}}{\partial \mathbf{W}_j}$ and $\frac{\partial \mathcal{R}}{\partial \mathbf{W}_j}$ in equation 20 disappear because both terms are already a function of expectation over $\mathbf{x}$ (see various auto-encoder regularizations) when we deal with expected cost function. Thus these terms are linear in $\mathbf{x}$ and hence taking an expectation results in 0.

From lemma 1, $\frac{\partial \mathcal{J}_{AE}}{\partial b_{e_j}} \geq -2\epsilon\sqrt{n}\|\mathbf{W}_j\|$, thus if $\lambda\frac{\partial \mathcal{R}}{\partial b_{e_j}} > 2\sigma_r\sqrt{n}\|\mathbf{W}_j\|$, then $\mathbb{E}_{\mathbf{x}}[a_j^{t+1}] < \mathbb{E}_{\mathbf{x}}[a_j^t]$.

Finally, $\mathrm{Var}[a_j^{t+1}] = \mathbb{E}_{\mathbf{x}}[a_j^{t+1} - \mathbb{E}_{\mathbf{x}}[a_j^{t+1}]]^2 = \mathbb{E}_{\mathbf{x}}[\mathbf{W}_j^{t+1}\mathbf{x}]^2 = \|\mathbf{W}_j^{t+1}\|^2$ $\square$

Corollary 1. ***If*** *s_e is a non-decreasing activation function with first derivative in $[0,1]$* ***and*** *$\mathcal{R} = \sum_{j=1}^{m} f(\mathbb{E}_{\mathbf{x}}[h_j])$ for any monotonically increasing function $f(.)$,* ***then*** *$\exists \lambda > 0$ such that updating $\{\mathbf{W}^t, \mathbf{b}_e^t\}$ along the negative gradient of $\mathcal{J}_{RAE}$ results in $\mathbb{E}_{\mathbf{x}}[a_j^{t+1}] \leq \mathbb{E}_{\mathbf{x}}[a_j^t]$* ***and*** *$\mathrm{Var}[a_j^{t+1}] = \|\mathbf{W}_j^{t+1}\|^2$* ***for*** *all $t \geq 0$.*

Proof. We need one additional argument other than theorem 1. $\frac{\partial \mathcal{R}}{\partial b_{e_j}} = \frac{\partial f(\mathbb{E}_{\mathbf{x}}[h_j])}{\partial \mathbb{E}_{\mathbf{x}}[h_j]} \mathbb{E}_{\mathbf{x}}\left[\frac{\partial h_j}{\partial a_j}\right]$. Since both $s_e(.)$ and $f(.)$ are non-decreasing functions, $\frac{\partial \mathcal{R}}{\partial b_{e_j}} \geq 0$ in all cases. □

Corollary 2. ***If*** *s_e is a non-decreasing convex activation function with first derivative in $[0,1]$* ***and*** *$\mathcal{R} = \mathbb{E}_{\mathbf{x}}\left[\sum_{j=1}^{m}\left(\left(\frac{\partial h_j}{\partial a_j}\right)^q \|\mathbf{W}_j^t\|_2^p\right)\right]$, $q \in \mathbb{N}$, $p \in \mathbb{W}$,* ***then*** *$\exists \lambda > 0$ such that updating $\{\mathbf{W}^t, \mathbf{b}_e^t\}$ along the negative gradient of $\mathcal{J}_{RAE}$, results in $\mathbb{E}_{\mathbf{x}}[a_j^{t+1}] \leq \mathbb{E}_{\mathbf{x}}[a_j^t]$* ***and*** *$\mathrm{Var}[a_j^{t+1}] = \|\mathbf{W}_j^{t+1}\|^2$* ***for*** *all $t \geq 0$.*

Proof. We need one additional argument other than theorem 1. $\frac{\partial \mathcal{R}}{\partial b_{e_j}} = \mathbb{E}_{\mathbf{x}}\left[q\left(\frac{\partial h_j}{\partial a_j}\right)^{q-1} \frac{\partial^2 h_j}{\partial a_j^2}\frac{\partial a_j}{\partial b_{e_j}}\|\mathbf{W}_j^t\|_2^p\right]$. Since $s_e(.)$ is a non-decreasing convex function, both $\frac{\partial^2 s_e(a_j)}{\partial a_j^2} \geq 0$ and $\frac{\partial s_e(a_j)}{\partial a_j} \geq 0 \; \forall a_j \in \mathbb{R}$. Finally, $\frac{\partial a_j}{\partial b_{e_j}} = 1$ by definition. Thus $\frac{\partial \mathcal{R}}{\partial b_{e_j}} \geq 0$ in all cases. □

Theorem 2. *Let p_j^t denote a lower bound of $\Pr(h_j^t \leq \delta_{\min})$ at iteration t and $s_e(.)$ be a non-decreasing function with first derivative in $[0,1]$.* ***If*** *$\|\mathbf{W}_j^t\|_2$ is upper bounded independent of λ* ***then*** *$\exists S \subseteq \mathbb{R}^+$* ***and*** *$\exists T_{\min}, T_{\max} \in \mathbb{N}$* ***such that*** *$p_j^{t+1} \geq p_j^t \; \forall \lambda \in S$, $T_{\min} \leq t \leq T_{\max}$.*

Proof. From theorem 1, $\mathbb{E}[a_j^{t+1}] < \mathbb{E}[a_j^t] \; \forall t \geq 0$. Define $a_{\min}$ such that $\delta_{\min} = \max_{a_{\min}} s_e(a_{\min})$. Thus $\exists T_{\min} \in \mathbb{N}$, such that $\forall t \geq T_{\min}$, $\mathbb{E}[a_j^t] < a_{\min}$. Then in the case of non-decreasing activation functions, using Chebyshev's bound,

$$\begin{aligned}\Pr(h_j^t \leq \delta_{\min}) = \Pr(a_j^t \leq a_{\min}) &\geq \Pr(|a_j^t - \mathbb{E}[a_j^t]| \leq a_{\min} - \mathbb{E}[a_j^t]) \\ &\geq 1 - \frac{\mathrm{Var}[a_j^t]}{(a_{\min} - \mathbb{E}[a_j^t])^2}\end{aligned} \tag{22}$$

Thus $p_j^t := 1 - \frac{\mathrm{Var}[a_j^t]}{(a_{\min} - \mathbb{E}[a_j^t])^2}$ lower bounds $\Pr(h_j^t \leq \delta_{\min}) \; \forall t \geq T_{\min}$. Now consider the difference

$$D(t) := \frac{\mathrm{Var}[a_j^{t+1}]}{(a_{\min} - \mathbb{E}[a_j^{t+1}])^2} - \frac{\mathrm{Var}[a_j^t]}{(a_{\min} - \mathbb{E}[a_j^t])^2} \tag{23}$$

and recall that

$$\mathbb{E}_{\mathbf{x}}\left[a_j^{t+1}\right] = \mathbb{E}_{\mathbf{x}}\left[a_j^t\right] - \eta\frac{\partial \mathcal{J}_{AE}}{\partial b_{e_j}} - \eta\lambda\frac{\partial \mathcal{R}}{\partial b_{e_j}} \tag{24}$$

where both the step size η and $\frac{\partial \mathcal{R}}{\partial b_{e_j}}$ are positive and $\partial \mathcal{J}_{AE}/\partial b_{e_j} \in [-2\sigma_r\sqrt{n}\|\mathbf{W}_j\|, 2\sigma_r\sqrt{n}\|\mathbf{W}_j\|]$. Thus, since $\mathrm{Var}[a_j] = \|\mathbf{W}_j^t\|^2$, we can always choose a fixed $S \subseteq \mathbb{R}^+$ such that $D(t) \leq 0 \; \forall \lambda \in S$ and $T_{\min} \leq t \leq T_{\max}$. □

Theorem 3. *Let $\{\mathbf{W}, \mathbf{b}_e\}$ represent the parameters of a DAE with squared loss, linear decoding, and i.i.d. Gaussian corruption with zero mean and σ^2 variance, at any point of training over data sampled from distribution $\mathcal{D}$. Let $a_j := \mathbf{W}_j\mathbf{x} + b_{e_j}$ so that $h_j = s_e(a_j)$ corresponding to sample $\mathbf{x} \sim \mathcal{D}$. Then,*

$$\begin{aligned}\mathcal{J}_{DAE} = \mathcal{J}_{AE} + \sigma^2\mathbb{E}_{\mathbf{x}}\Bigg[&\sum_{j=1}^{m}\left(\left(\frac{\partial h_j}{\partial a_j}\right)^2\|\mathbf{W}_j\|_2^4\right) + \sum_{\substack{j,k=1 \\ j\neq k}}^{m}\left(\frac{\partial h_j}{\partial a_j}\frac{\partial h_k}{\partial a_k}(\mathbf{W}_j^T\mathbf{W}_k)^2\right) \\ &+ \sum_{i=1}^{n}\left((\mathbf{b}_d + \mathbf{W}^T\mathbf{h} - \mathbf{x})^T\mathbf{W}^T\left(\frac{\partial^2\mathbf{h}}{\partial \mathbf{a}^2}\odot\mathbf{W}^i\odot\mathbf{W}^i\right)\right)\Bigg] + o(\sigma^2)\end{aligned} \tag{25}$$

where $\frac{\partial^2 \mathbf{h}}{\partial \mathbf{a}^2} \in \mathbb{R}^m$ *is the element-wise* 2^{nd} *derivative of* $\mathbf{h}$ *w.r.t.* $\mathbf{a}$ *and* $\odot$ *is element-wise product.*

Proof. Using 2^{nd} order Taylor's expansion of the loss function, we get

$$\ell(\mathbf{x}, f_d(f_e(\tilde{\mathbf{x}}))) = \ell(\mathbf{x}, f_d(f_e(\mu_{\mathbf{x}}))) + (\tilde{\mathbf{x}} - \mu_{\mathbf{x}})^T \nabla_{\tilde{\mathbf{x}}} \ell + \frac{1}{2} (\tilde{\mathbf{x}} - \mu_{\mathbf{x}})^T \nabla^2_{\tilde{\mathbf{x}}} \ell \ (\tilde{\mathbf{x}} - \mu_{\mathbf{x}}) + o(\sigma^2) \tag{26}$$

where $\mu_{\mathbf{x}} = \mathbf{x}$. since we assume zero mean Gaussian noise. Thus taking the expectation of this approximation over noise yields

$$\mathbb{E}[\ell(\mathbf{x}, f_d(f_e(\tilde{\mathbf{x}})))] = \mathbb{E}[\ell(\mathbf{x}, f_d(f_e(\mu_{\mathbf{x}})))] + \frac{1}{2} tr(\Sigma_{\mathbf{x}} \nabla^2_{\tilde{\mathbf{x}}} \ell) + o(\sigma^2) \tag{27}$$

where $\Sigma_{\mathbf{x}} := \mathbb{E}[(\tilde{\mathbf{x}} - \mu_{\mathbf{x}})(\tilde{\mathbf{x}} - \mu_{\mathbf{x}})^T]$. Since the corruption is $i.i.d.$, assume the covariance $\Sigma_{\mathbf{x}} = \sigma^2 \mathbf{I}$, where $\mathbf{I}$ is the identity matrix.

Taking expectation over $\mathbf{x}$, we can rewrite equation (27) as

$$\mathcal{J}_{DAE} = \mathcal{J}_{AE} + \mathbb{E}_{\mathbf{x}} \left[\frac{1}{2} \sigma^2 \sum_{i=1}^{n} \frac{\partial^2 \ell}{\partial \tilde{x}_i^2} \right] + o(\sigma^2) \tag{28}$$

Expanding the second order term in the above equation, we get

$$\frac{\partial^2 \ell}{\partial \tilde{x}_i^2} = \frac{\partial \mathbf{h}}{\partial \tilde{x}_i}^T \frac{\partial^2 \ell}{\partial \mathbf{h}^2} \frac{\partial \mathbf{h}}{\partial \tilde{x}_i} + \frac{\partial \ell}{\partial \mathbf{h}}^T \frac{\partial^2 \mathbf{h}}{\partial \tilde{x}_i^2} \tag{29}$$

For linear decoding and squared loss,

$$\frac{\partial \ell}{\partial \mathbf{h}}^T \frac{\partial^2 \mathbf{h}}{\partial \tilde{x}_i^2} = \sum_{i=1}^{n} \left((\mathbf{b}_d + \mathbf{W}^T \mathbf{h} - \mathbf{x})^T \mathbf{W}^T \left(\frac{\partial^2 \mathbf{h}}{\partial \mathbf{a}^2} \odot \mathbf{W}^i \odot \mathbf{W}^i \right) \right) \tag{30}$$

where $\frac{\partial^2 \mathbf{h}}{\partial \mathbf{a}^2} \in \mathbb{R}^m$ is the element-wise 2^{nd} derivative of $\mathbf{h}$ *w.r.t.* $\mathbf{a}$, $\odot$ represents element-wise product and $\mathbf{W}^i$ denotes the i^{th} column of $\mathbf{W}$. Let vector $\mathbf{d_h} \in \mathbb{R}^m$ be defined such that $d_{h_j} = \frac{\partial h_j}{\partial a_j} \ \forall j \in \{1, 2, \ldots, m\}$. Then,

$$\sum_{i=1}^{n} \frac{\partial \mathbf{h}}{\partial \tilde{x}_i}^T \frac{\partial^2 \ell}{\partial \mathbf{h}^2} \frac{\partial \mathbf{h}}{\partial \tilde{x}_i} = 2 \sum_{j=1}^{n} \sum_{k=1}^{n} \left((\mathbf{d_h} \odot (\mathbf{W})^j)^T (\mathbf{W})^k \right)^2 \tag{31}$$

where $(\mathbf{W})^j$ represents the j^{th} column of $\mathbf{W}$ and $\odot$ denotes element-wise product. Let $\mathbf{D_h} = \text{diag}(\mathbf{d_h})$. Then,

$$\sum_{j=1}^{n} \sum_{k=1}^{n} \left((\mathbf{d_h} \odot (\mathbf{W})^j)^T (\mathbf{W})^k \right)^2 = \|(\mathbf{D_h} \mathbf{W})^T \mathbf{W}\|_F^2 \tag{32}$$

Finally, using the cyclic property of trace operator, we get, $\|(\mathbf{D_h} \mathbf{W})^T \mathbf{W}\|_F^2 = \text{tr}(\mathbf{W}^T \mathbf{D_h} \mathbf{W} \mathbf{W}^T \mathbf{D_h} \mathbf{W}) = \text{tr}(\mathbf{D_h} \mathbf{W} \mathbf{W}^T \mathbf{D_h} \mathbf{W} \mathbf{W}^T)$. Thus DAE objective becomes,

$$\begin{aligned} \mathcal{J}_{DAE} = \mathcal{J}_{AE} + \sigma^2 \mathbb{E}_{\mathbf{x}} \Big[\text{tr}(\mathbf{D_h} \mathbf{W} \mathbf{W}^T \mathbf{D_h} \mathbf{W} \mathbf{W}^T) + \\ \sum_{i=1}^{n} \left((\mathbf{b}_d + \mathbf{W}^T \mathbf{h} - \mathbf{x})^T \mathbf{W}^T \left(\frac{\partial^2 \mathbf{h}}{\partial \mathbf{a}^2} \odot \mathbf{W}^i \odot \mathbf{W}^i \right) \right) \Big] + o(\sigma^2) \end{aligned} \tag{33}$$

Upon expansion of the second term above, we get the final form. □

Remark 3. *Let $\{\mathbf{W} \in \mathbb{R}^{m \times n}, \mathbf{b}_e \in \mathbb{R}^m\}$ represent the parameters of a Marginalized De-noising Auto-Encoder (mDAE) with $s_e(.)$ activation function, linear decoding, squared loss and $\sigma^2_{\mathbf{x}i} = \lambda \; \forall i \in \{1, \ldots, n\}$, at any point of training over data sampled from some distribution $\mathcal{D}$. Let $a_j := \mathbf{W}_j\mathbf{x} + b_{e_j}$ so that $h_j = s_e(a_j)$ corresponding to sample $\mathbf{x} \sim \mathcal{D}$. Then,*

$$\mathcal{J}_{mDAE} = \mathcal{J}_{AE} + \lambda \mathbb{E}_{\mathbf{x}} \left[\sum_{j=1}^{m} \left(\left(\frac{\partial h_j}{\partial a_j} \right)^2 \|\mathbf{W}_j\|_2^4 \right) \right] \tag{34}$$

Proof. For linear decoding and squared loss, $\frac{\partial^2 \ell}{\partial {h_j}^2} = 2\|\mathbf{W}_j\|_2^2$ and $\frac{\partial h_j}{\partial \tilde{\mathbf{x}}_i} = \frac{\partial h_j}{\partial a_j} W_{ji}$. Thus

$$\begin{aligned} &\frac{1}{2} \sum_{i=1}^{n} \sigma^2_{\mathbf{x}i} \sum_{j=1}^{m} \frac{\partial^2 \ell}{\partial {h_j}^2} \left(\frac{\partial h_j}{\partial \tilde{\mathbf{x}}_i} \right)^2 = \sum_{i=1}^{n} \lambda \sum_{j=1}^{m} \|\mathbf{W}_j\|_2^2 \left(\frac{\partial h_j}{\partial a_j} W_{ji} \right)^2 \\ &= \lambda \sum_{j=1}^{m} \|\mathbf{W}_j\|_2^2 \left(\frac{\partial h_j}{\partial a_j} \right)^2 \sum_{i=1}^{n} W_{ji}^2 = \lambda \sum_{j=1}^{m} \left(\frac{\partial h_j}{\partial a_j} \right)^2 \|\mathbf{W}_j\|_2^4 \end{aligned} \tag{35}$$

$\square$