Supplementary Materials

A Proofs of Theorem 1

We basically follow the arguments in the convergence of Euclidean SAM [50], but the details are totally different.

Lemma 1 (Properties of Retraction Smoothness)**.** *Let* $w^* = \mathrm{R}_w\big(\rho \mathrm{grad}\mathcal{L}(w)\big)$ *and the curve* $\gamma(t) = \mathrm{R}_w(t\eta)$ *with the endpoints* $\gamma(0) = w$ *and* $\gamma(1) = w^*$*. Then, we have*

$$\langle \mathcal{T}(\gamma)_{w^*}^{w} \mathrm{grad}\mathcal{L}(w^*), \mathrm{grad}\mathcal{L}(w)\rangle \geq (1 - \rho L_{\mathrm{R}})\|\mathrm{grad}\mathcal{L}(w)\|_w^2$$

Proof. By the condition (C-5), we have

$$\|\mathcal{T}(\gamma)_{w^*}^{w} \mathrm{grad}\mathcal{L}(w^*) - \mathrm{grad}\mathcal{L}(w)\|_w \leq L_{\mathrm{R}}\|\eta\|_w$$

By the Cauchy-Schwarz inequality, we have

$$\begin{aligned}
|\langle \mathcal{T}(\gamma)_{w^*}^{w} \mathrm{grad}\mathcal{L}(w^*) - \mathrm{grad}\mathcal{L}(w), \eta\rangle_w| &\leq \|\mathcal{T}(\gamma)_{w^*}^{w} \mathrm{grad}\mathcal{L}(w^*) - \mathrm{grad}\mathcal{L}(w)\|_w \|\eta\|_w \\
&\leq L_{\mathrm{R}}\|\eta\|_w^2 \\
&\leq \rho^2 L_{\mathrm{R}}\|\mathrm{grad}\mathcal{L}(w)\|_w^2
\end{aligned}$$

Therefore, we obtain

$$\langle \mathcal{T}(\gamma)_{w^*}^{w} \mathrm{grad}\mathcal{L}(w^*) - \mathrm{grad}\mathcal{L}(w), \rho \mathrm{grad}\mathcal{L}(w)\rangle_w \geq -\rho^2 L_{\mathrm{R}}\|\mathrm{grad}\mathcal{L}(w)\|_w^2$$

Removing the constant ρ, the above inequality becomes

$$\langle \mathcal{T}(\gamma)_{w^*}^{w} \mathrm{grad}\mathcal{L}(w^*) - \mathrm{grad}\mathcal{L}(w), \mathrm{grad}\mathcal{L}(w)\rangle_w \geq -\rho L_{\mathrm{R}}\|\mathrm{grad}\mathcal{L}(w)\|_w^2$$

Lastly, we arrive at the final result as

$$\begin{aligned}
\langle \mathcal{T}(\gamma)_{w^*}^{w} \mathrm{grad}\mathcal{L}(w^*), \mathrm{grad}\mathcal{L}(w)\rangle_w &= \langle \mathcal{T}(\gamma)_{w^*}^{w} \mathrm{grad}\mathcal{L}(w^*) - \mathrm{grad}\mathcal{L}(w), \mathrm{grad}\mathcal{L}(w)\rangle_w + \|\mathrm{grad}\mathcal{L}(w)\|_w^2 \\
&\geq (1 - \rho L_{\mathrm{R}})\|\mathrm{grad}\mathcal{L}(w)\|_w^2
\end{aligned}$$

□

In the next lemma, we will show the alignment of the true Riemannian gradient and the true Riemannian SAM gradient.

Lemma 2 (Alignment of the true Riemannian gradient and the true Riemannian SAM gradient)**.** *Let us denote the stochastic Riemannian gradient at time t by* $\mathrm{grad}\mathcal{L}_t(w) = \frac{1}{b}\sum_{i\in J_t} \mathrm{grad}\mathcal{L}(w; x_i) \in \mathrm{T}_w\mathcal{M}$ *and* $w^{adv} = \mathrm{R}_w\big(\rho \mathrm{grad}\mathcal{L}_t(w)\big)$*. Further, let* $\gamma(t) = \mathrm{R}_w(t\eta)$ *be a retraction curve with* $\gamma(0) = w$ *and* $\gamma(1) = w^{adv}$*. Then, we have the following inequality*

$$\mathbb{E}\Big[\Big\langle \mathcal{T}(\gamma)_{w^{adv}}^{w} \mathrm{grad}\mathcal{L}_t(w^{adv}), \mathrm{grad}\mathcal{L}(w)\Big\rangle_w\Big] \geq \Big(\frac{1}{2} - \rho L_{\mathrm{R}} - 3\rho^2 L_{\mathrm{R}}^2\Big)\Big\|\mathrm{grad}\mathcal{L}(w)\Big\|_w^2 - \frac{2\rho^2 L_{\mathrm{R}}^2 \sigma^2}{b}$$

Proof. Let $w^* = \mathrm{R}_w\big(\rho \mathrm{grad}\mathcal{L}(w)\big)$ evaluated on the loss function. We first add and subtract $\langle \mathcal{T}(\zeta)_{w^*}^{w} \mathrm{grad}\mathcal{L}_t(w^*), \mathrm{grad}\mathcal{L}(w)\rangle_w$ where $\zeta(t) = \mathrm{R}_w(t\xi)$ is a retraction curve where $\zeta(0) = w$ and $\zeta(1) = w^*$.

$$\begin{aligned}
\langle \mathcal{T}(\gamma)_{w^{adv}}^{w} \mathrm{grad}\mathcal{L}_t(w^{adv}), \mathrm{grad}\mathcal{L}(w)\rangle_w &= \underbrace{\langle \mathcal{T}(\gamma)_{w^{adv}}^{w} \mathrm{grad}\mathcal{L}_t(w^{adv}) - \mathcal{T}(\zeta)_{w^*}^{w} \mathrm{grad}\mathcal{L}_t(w^*), \mathrm{grad}\mathcal{L}(w)\rangle_w}_{T_1} \\
&\quad + \underbrace{\langle \mathcal{T}(\zeta)_{w^*}^{w} \mathrm{grad}\mathcal{L}_t(w^*), \mathrm{grad}\mathcal{L}(w)\rangle_w}_{T_2}
\end{aligned}$$

We will bound two terms, T_1 and T_2, separately. Regarding the term T_1, we derive

$$\begin{aligned}
-T_1 &= -\langle \mathcal{T}(\gamma)_{w^{adv}}^{w} \mathrm{grad}\mathcal{L}_t(w^{adv}) - \mathcal{T}(\zeta)_{w^*}^{w} \mathrm{grad}\mathcal{L}_t(w^*), \mathrm{grad}\mathcal{L}(w)\rangle_w \\
&\leq \frac{1}{2}\Big\|\mathcal{T}(\gamma)_{w^{adv}}^{w} \mathrm{grad}\mathcal{L}_t(w^{adv}) - \mathcal{T}(\zeta)_{w^*}^{w} \mathrm{grad}\mathcal{L}_t(w^*)\Big\|_w^2 + \frac{1}{2}\Big\|\mathrm{grad}\mathcal{L}(w)\Big\|_w^2 \\
&\leq \Big\|\mathcal{T}(\gamma)_{w^{adv}}^{w} \mathrm{grad}\mathcal{L}_t(w^{adv}) - \mathrm{grad}\mathcal{L}_t(w)\Big\|_w^2 + \Big\|\mathcal{T}(\zeta)_{w^*}^{w} \mathrm{grad}\mathcal{L}_t(w^*) - \mathrm{grad}\mathcal{L}_t(w)\Big\|_w^2 + \frac{1}{2}\Big\|\mathrm{grad}\mathcal{L}(w)\Big\|_w^2 \\
&\leq L_{\mathrm{R}}^2\|\rho \mathrm{grad}\mathcal{L}_t(w)\|_w^2 + L_{\mathrm{R}}^2\|\rho \mathrm{grad}\mathcal{L}(w)\|_w^2 + \frac{1}{2}\|\mathrm{grad}\mathcal{L}(w)\|_w^2 \\
&\leq \rho^2 L_{\mathrm{R}}^2\Big(2\|\mathrm{grad}\mathcal{L}_t(w) - \mathrm{grad}\mathcal{L}(w)\|_w^2 + 2\|\mathrm{grad}\mathcal{L}(w)\|_w^2\Big) + \Big(\frac{1}{2} + \rho^2 L_{\mathrm{R}}^2\Big)\|\mathrm{grad}\mathcal{L}(w)\|_w^2 \\
&\leq \frac{2\rho^2 L_{\mathrm{R}}^2 \sigma^2}{b} + \Big(\frac{1}{2} + 3\rho^2 L_{\mathrm{R}}^2\Big)\|\mathrm{grad}\mathcal{L}(w)\|_w^2
\end{aligned}$$

From the above inequality, we could finally bound the term T_1 as

$$T_1 \geq -\frac{2\rho^2 L_{\mathrm{R}}^2 \sigma^2}{b} - \Big(\frac{1}{2} + 3\rho^2 L_{\mathrm{R}}^2\Big)\|\mathrm{grad}\mathcal{L}(w)\|_w^2$$

Regarding the term T_2, we just use the lemma as

$$T_2 = \langle \mathcal{T}(\zeta)_{w^*}^{w} \mathrm{grad}\mathcal{L}_t(w^*), \mathrm{grad}\mathcal{L}(w)\rangle_w \geq (1-\rho L_{\mathrm{R}})\|\mathrm{grad}\mathcal{L}(w)\|_w^2$$

Hence, we arrive at

$$\mathbb{E}\Big[\Big\langle \mathcal{T}(\gamma)_{w^{adv}}^{w} \mathrm{grad}\mathcal{L}_t(w^{adv}), \mathrm{grad}\mathcal{L}(w)\Big\rangle_w\Big] \geq \Big(\frac{1}{2} - \rho L_{\mathrm{R}} - 3\rho^2 L_{\mathrm{R}}^2\Big)\Big\|\mathrm{grad}\mathcal{L}(w)\Big\|_w^2 - \frac{2\rho^2 L_{\mathrm{R}}^2\sigma^2}{b}$$

□

According to Algorithm 1, we follow the notation as

$$\mathrm{grad}\mathcal{L}_t(w) = \frac{1}{b}\sum_{i\in I_t} \mathrm{grad}\ell_i(w)$$

$$w_t^{adv} = \mathrm{R}_{w_t}\big(\rho\,\mathrm{grad}\mathcal{L}_t(w_t)\big)$$

We assume the stochastic m-SAM where the same batch is used for both inner and outer updates.

Lemma 3 (Descent inequality). *Under the assumptions in Theorem 1, we have*

$$\mathbb{E}\big[\mathcal{L}(w_{t+1})\big] \leq \mathbb{E}\big[\mathcal{L}(w_t)\big] - \frac{3\alpha}{8}\mathbb{E}\big[\|\mathrm{grad}\mathcal{L}(w_t)\|_{w_t}^2\big] + \frac{\alpha^2 L_S \sigma^2}{b} + \frac{2\alpha^2 L_S^3 \rho^2\sigma^2}{b} + \frac{2\alpha\rho^3 L_{\mathrm{R}}^2\sigma^2}{b}$$

Proof. Using the condition (C-4), we have

$$\begin{aligned}
\mathcal{L}(w_{t+1}) &= \mathcal{L}\Bigg(\mathrm{R}_{w_t}\Big(-\alpha\mathcal{T}(\gamma)_{w_t^{adv}}^{w_t}\mathrm{grad}\mathcal{L}_t(w_t^{adv})\Big)\Bigg)\\
&\leq \mathcal{L}(w_t) - \alpha\Big\langle \mathrm{grad}\mathcal{L}(w_t), \mathcal{T}(\gamma)_{w_t^{adv}}^{w_t}\mathrm{grad}\mathcal{L}_t(w_t^{adv})\Big\rangle_{w_t} + \frac{\alpha^2 L_S}{2}\Big\|\mathcal{T}(\gamma)_{w_t^{adv}}^{w_t}\mathrm{grad}\mathcal{L}_t(w_t^{adv})\Big\|_{w_t}^2
\end{aligned}$$

For the last term in RHS, we can bound as

$$\begin{aligned}
\Big\|\mathcal{T}(\gamma)_{w_t^{adv}}^{w_t}\mathrm{grad}\mathcal{L}_t(w_t^{adv})\Big\|_{w_t}^2 &= -\|\mathrm{grad}\mathcal{L}(w_t)\|_{w_t}^2 + \Big\|\mathcal{T}(\gamma)_{w_t^{adv}}^{w_t}\mathrm{grad}\mathcal{L}_t(w_t^{adv}) - \mathrm{grad}\mathcal{L}(w_t)\Big\|_{w_t}^2\\
&\quad + 2\Big\langle \mathcal{T}(\gamma)_{w_t^{adv}}^{w_t}\mathrm{grad}\mathcal{L}_t(w_t^{adv}), \mathrm{grad}\mathcal{L}(w_t)\Big\rangle_{w_t}
\end{aligned}$$

Again, we have

$$\begin{aligned}
\mathcal{L}(w_{t+1}) &\leq \mathcal{L}(w_t) - \alpha\Big\langle \mathrm{grad}\mathcal{L}(w_t), \mathcal{T}(\gamma)_{w_t^{adv}}^{w_t}\mathrm{grad}\mathcal{L}_t(w_t^{adv})\Big\rangle_{w_t} + \frac{\alpha^2 L_S}{2}\Big\|\mathcal{T}(\gamma)_{w_t^{adv}}^{w_t}\mathrm{grad}\mathcal{L}_t(w_t^{adv})\Big\|_{w_t}^2\\
&= \mathcal{L}(w_t) - \frac{\alpha^2 L_S}{2}\|\mathrm{grad}\mathcal{L}(w_t)\|_{w_t}^2 + \frac{\alpha^2 L_S}{2}\Big\|\mathcal{T}(\gamma)_{w_t^{adv}}^{w_t}\mathrm{grad}\mathcal{L}_t(w_t^{adv}) - \mathrm{grad}\mathcal{L}(w_t)\Big\|_{w_t}^2\\
&\quad - \alpha(1-\alpha L_S)\Big\langle \mathcal{T}(\gamma)_{w_t^{adv}}^{w_t}\mathrm{grad}\mathcal{L}_t(w_t^{adv}), \mathrm{grad}\mathcal{L}(w_t)\Big\rangle_{w_t}\\
&\leq \mathcal{L}(w_t) - \frac{\alpha^2 L_S}{2}\|\mathrm{grad}\mathcal{L}(w_t)\|_{w_t}^2 + \alpha^2 L_S\Big\|\mathcal{T}(\gamma)_{w_t^{adv}}^{w_t}\mathrm{grad}\mathcal{L}_t(w_t^{adv}) - \mathrm{grad}\mathcal{L}_t(w_t)\Big\|_{w_t}^2\\
&\quad + \alpha^2 L_S\|\mathrm{grad}\mathcal{L}_t(w_t) - \mathrm{grad}\mathcal{L}(w_t)\|_{w_t}^2 - \alpha(1-\alpha L_S)\Big\langle \mathcal{T}(\gamma)_{w_t^{adv}}^{w_t}\mathrm{grad}\mathcal{L}_t(w_t^{adv}), \mathrm{grad}\mathcal{L}(w_t)\Big\rangle_{w_t}\\
&\leq \mathcal{L}(w_t) - \frac{\alpha^2 L_S}{2}\|\mathrm{grad}\mathcal{L}(w_t)\|_{w_t}^2 + \alpha^2 L_S^3\rho^2\|\mathrm{grad}\mathcal{L}_t(w_t)\|_{w_t}^2 + \alpha^2 L_S\|\mathrm{grad}\mathcal{L}_t(w_t) - \mathrm{grad}\mathcal{L}(w_t)\|_{w_t}^2\\
&\quad - \alpha(1-\alpha L_S)\Big\langle \mathcal{T}(\gamma)_{w_t^{adv}}^{w_t}\mathrm{grad}\mathcal{L}_t(w_t^{adv}), \mathrm{grad}\mathcal{L}(w_t)\Big\rangle_{w_t}\\
&\leq \mathcal{L}(w_t) - \frac{\alpha^2 L_S}{2}\|\mathrm{grad}\mathcal{L}(w_t)\|_{w_t}^2 + 2\alpha^2 L_S^3\rho^2\|\mathrm{grad}\mathcal{L}(w_t)\|_{w_t}^2 + 2\alpha^2 L_S^3\rho^2\|\mathrm{grad}\mathcal{L}_t(w_t) - \mathrm{grad}\mathcal{L}(w_t)\|_{w_t}^2\\
&\quad + \alpha^2 L_S\|\mathrm{grad}\mathcal{L}_t(w_t) - \mathrm{grad}\mathcal{L}(w_t)\|_{w_t}^2 - \alpha(1-\alpha L_S)\Big\langle \mathcal{T}(\gamma)_{w_t^{adv}}^{w_t}\mathrm{grad}\mathcal{L}_t(w_t^{adv}), \mathrm{grad}\mathcal{L}(w_t)\Big\rangle_{w_t}\\
&= \mathcal{L}(w_t) - \frac{\alpha^2 L_S(1-4L_S^2\rho^2)}{2}\|\mathrm{grad}\mathcal{L}(w_t)\|_{w_t}^2 + \alpha^2 L_S(1+2L_S^2\rho^2)\|\mathrm{grad}\mathcal{L}_t(w_t) - \mathrm{grad}\mathcal{L}(w_t)\|_{w_t}^2\\
&\quad - \alpha(1-\alpha L_S)\Big\langle \mathcal{T}(\gamma)_{w_t^{adv}}^{w_t}\mathrm{grad}\mathcal{L}_t(w_t^{adv}), \mathrm{grad}\mathcal{L}(w_t)\Big\rangle_{w_t}
\end{aligned}$$

Taking the expectation on both sides, we have

$$
\begin{aligned}
\mathbb{E}\big[\mathcal{L}(w_{t+1})\big] \leq \mathbb{E}\big[\mathcal{L}(w_t)\big] &- \frac{\alpha^2 L_S(1-4L_S^2\rho^2)}{2}\mathbb{E}\big[\|\mathrm{grad}\mathcal{L}(w_t)\|_{w_t}^2\big] + \frac{\alpha^2 L_S(1+2L_S^2\rho^2)\sigma^2}{b} \\
&- \alpha(1-\alpha L_S)\mathbb{E}\left[\left\langle \mathcal{T}(\gamma)_{w_t^{adv}}^{w_t}\mathrm{grad}\mathcal{L}_t(w_t^{adv}), \mathrm{grad}\mathcal{L}(w_t)\right\rangle_{w_t}\right] \\
\mathbb{E}\big[\mathcal{L}(w_t)\big] &- \frac{\alpha^2 L_S(1-4L_S^2\rho^2)}{2}\mathbb{E}\big[\|\mathrm{grad}\mathcal{L}(w_t)\|_{w_t}^2\big] + \frac{\alpha^2 L_S(1+2L_S^2\rho^2)\sigma^2}{b} \\
&- \alpha(1-\alpha L_S)\left[\Big(\frac{1}{2} - \rho L_{\mathrm{R}} - 3\rho^2 L_{\mathrm{R}}^2\Big)\mathbb{E}\big[\|\mathrm{grad}\mathcal{L}(w_t)\|_{w_t}^2\big] - \frac{2\rho^2 L_{\mathrm{R}}^2\sigma^2}{b}\right]
\end{aligned}
$$

For sufficiently large number of total iteration T, the condition $\rho \leq \frac{1}{4\widetilde{L}}$ is easily satisfied where $\widetilde{L} = \max\{L_{\mathrm{R}}, L_S\}$ (defined in Theorem 1). Hence, we obtain

$$
\begin{aligned}
\frac{3\alpha}{8}\mathbb{E}\big[\|\mathrm{grad}\mathcal{L}(w_t)\|_{w_t}^2\big] &\leq \mathbb{E}\big[\mathcal{L}(w_t)\big] - \mathbb{E}\big[\mathcal{L}(w_{t+1})\big] + \frac{\alpha^2 L_S(1+2L_S^2\rho^2)\sigma^2}{b} + \frac{2\alpha(1-\alpha L_S)\rho^3 L_{\mathrm{R}}^2\sigma^2}{b} \\
&\leq \mathbb{E}\big[\mathcal{L}(w_t)\big] - \mathbb{E}\big[\mathcal{L}(w_{t+1})\big] + \frac{\alpha^2 L_S\sigma^2}{b} + \frac{2\alpha^2 L_S^3\rho^2\sigma^2}{b} + \frac{2\alpha\rho^3 L_{\mathrm{R}}^2\sigma^2}{b}
\end{aligned}
$$

By telescoping the above inequality from $t = 0 \sim T-1$, we arrive at

$$
\mathbb{E}\big[\|\mathrm{grad}\mathcal{L}(\widetilde{w})\|_{\widetilde{w}}^2\big] = \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\big[\|\mathrm{grad}\mathcal{L}(w_t)\|_{w_t}^2\big] \leq \frac{8\Delta}{3\alpha T} + \frac{8\alpha^2 L_S\sigma^2}{3b} + \frac{16\alpha^2 L_S^3\rho^2\sigma^2}{3b} + \frac{16\alpha\rho^3 L_{\mathrm{R}}^2\sigma^2}{3b}
$$

Under the step size condition $\alpha_t = \frac{1}{\sqrt{T}\widetilde{L}}$ and $\rho_t = \frac{1}{T^{1/6}\widetilde{L}}$, we finally get

$$
\mathbb{E}\big[\|\mathrm{grad}\mathcal{L}(\widetilde{w})\|_{\widetilde{w}}^2\big] \leq \frac{Q_1\widetilde{L}\Delta}{\sqrt{T}} + \frac{Q_2\sigma^2}{b\sqrt{T}} + \frac{Q_3\sigma^2}{bT^{5/6}}
$$

for appropriate constants $\{Q_i\}_{i=1}^3$. $\square$

B Hyperparameter Details

We use the almost same hyperparameters in the study [51] and implement our experiments in Section 5 upon its official implementation. For completeness, we summarize the hyperparameter configurations in Table 4 and Table 5.

Table 4: Hyperparameter configurations for knowledge graph completion.

	WN18RR		**FB15k-237**	
Dimension	32	β	32	β
Batch Size	1000	1000	500	500
Negative Samples	50	50	50	50
Margin	8.0	8.0	8.0	8.0
Epochs	1000	1000	500	500
Max Norm	1.5	2.5	1.5	1.5
Max Scaler	3.5	2.5	2.5	2.5
Learning Rate	0.005	0.003	0.003	0.003
Gradient Norm	0.5	0.5	0.5	0.5

Table 5: Hyperparameter configurations for machine translation.

Hyperparameter	**IWSLT'14**	**WMT'14**
GPU Numbers	4	4
Embedding Dimension d	64	64
Feed-forward Dimension	256	256
Batch Type	Token	Token
Batch Size	10240	10240
Gradient Accumulation Steps	1	1
Training Steps	40000	200000
Dropout	0.0	0.1
Attention Dropout	0.1	0.0
Max Gradient Norm	0.5	0.5
Warmup Steps	8000	6000
Decay Method	noam	noam
Label Smoothing	0.1	0.1
Layer Number	6	6
Head Number	4	8
Learning Rate	5.0	5.0
Adam β_2	0.998	0.998