A Algorithm

In Algorithm 2, we also apply Ouroboros to Adam, an adaptive variant of SGD. We update first moment vectors and second moment vectors using gradients computed by the Ouroboros algorithm so that the updates in modules can be parallelized.

Algorithm 2 Ouroboros + Adam

Require:
Initial weights $w^0 = [w^0_{\mathcal{G}(1)}, ..., w^0_{\mathcal{G}(K)}]$;
Initial word embedding $V_i^0 = V_o^0$;
Stepsize: γ; Small constant $\epsilon = 10^{-8}$;
Exponential decay: $\beta_1 = 0.9$, $\beta_2 = 0.999$;
1_{st} moment vector: $m^0_{\mathcal{G}(k)} = 0, \forall k$, $m_V^0 = 0$;
2_{nd} moment vector: $v^0_{\mathcal{G}(k)} = 0, \forall k$, $v_V^0 = 0$;

for $t = 0, 1, 2, \ldots, T-1$ **do**
  for $k = 1, \ldots, K$ **in parallel do**
    Compute delayed gradient g_k^t for module k following (8):
    Compute mixed gradient g_V^t for embedding layer following (9):
    Update biased first moment estimate:
      $m^{t+1}_{\mathcal{G}(k)} = \beta_1 \cdot m^t_{\mathcal{G}(k)} + (1-\beta_1) \cdot g_k^t$;
      $m_V^{t+1} = \beta_1 \cdot m_V^t + (1-\beta_1) \cdot g_V^t$;
    Update biased second moment estimate:
      $v^{t+1}_{\mathcal{G}(k)} = \beta_2 \cdot v^t_{\mathcal{G}(k)} + (1-\beta_2) \cdot (g_k^t)^2$;
      $v_V^{t+1} = \beta_2 \cdot v_V^t + (1-\beta_2) \cdot (g_V^t)^2$;
    Compute bias-correct first moment estimate:
      $\hat{m}^{t+1}_{\mathcal{G}(k)} = m^{t+1}_{\mathcal{G}(k)} / (1-\beta_1^{t+1})$;
      $\hat{m}_V^{t+1} = m_V^{t+1} / (1-\beta_1^{t+1})$;
    Compute bias-correct second moment estimate:
      $\hat{v}^{t+1}_{\mathcal{G}(k)} = v^{t+1}_{\mathcal{G}(k)} / (1-\beta_2^{t+1})$;
      $\hat{v}_V^{t+1} = v_V^{t+1} / (1-\beta_2^{t+1})$;
    Update weights and embedding layer following Adam:
      $w^{t+1}_{\mathcal{G}(k)} = w^t_{\mathcal{G}(k)} - \gamma \cdot \frac{\hat{m}^{t+1}_{\mathcal{G}(k)}}{\left(\sqrt{\hat{v}^{t+1}_{\mathcal{G}(k)}} + \epsilon\right)}$;
      $V_i^{t+1} = V_o^{t+1} = V_i^t - \gamma \cdot \frac{\hat{m}_V^{t+1}}{\left(\sqrt{\hat{v}_V^{t+1}} + \epsilon\right)}$;
  end for
end for
Output w^s, V_i^s and V_o^s randomly from $\{w^t\}_{t=0}^{T-1}$, $\{V_i^t\}_{t=0}^{T-1}$ and $\{V_o^t\}_{t=0}^{T-1}$.

B Proof

Proof to Lemma 1

Proof: Let $\tilde{w} = [V, w]$, it is satisfied that:

$$\nabla f(\tilde{w}^t) \quad = \quad \sum_{k=1}^{K} \nabla f_{\mathcal{G}(k)}\left(\tilde{w}^t\right) + \nabla f_{V_i}\left(\tilde{w}^t\right) + \nabla f_{V_o}\left(\tilde{w}^t\right) \tag{18}$$

According to Assumption 1, the following inequality holds that:

$$f(\tilde{w}^{t+1}) \quad \leq \quad f(\tilde{w}^t) + \nabla f(\tilde{w}^t)^T \left(\tilde{w}^{t+1} - \tilde{w}^t\right) + \frac{L}{2} \left\|\tilde{w}^{t+1} - \tilde{w}^t\right\|_2^2. \tag{19}$$

From the update rule in Algorithm 1, we take expectation on both sides and obtain: small

$$
\begin{aligned}
&\mathbb{E}\left[f(\tilde{w}^{t+1})\right] \\
\leq\ & f(\tilde{w}^t) - \gamma_t \nabla f(\tilde{w}^t)^T \Bigg(\sum_{k=1}^{K} \nabla f_{\mathcal{G}(k)}\left(\tilde{w}^{t-K+k}\right) + \frac{1}{2}\nabla f_V\left(\tilde{w}^{t-K+1}\right) + \frac{1}{2}\nabla f_V\left(\tilde{w}^t\right) \\
& + \nabla f\left(\tilde{w}^t\right) - \nabla f\left(\tilde{w}^t\right) \Bigg) + \frac{L\gamma_t^2}{2}\mathbb{E}\Bigg\| \sum_{k=1}^{K} \nabla f_{\mathcal{G}(k),x_{i(t-K+k)}}\left(\tilde{w}^{t-K+k}\right) \\
& + \frac{1}{2}\nabla f_{V,x_{i(t-K+1)}}\left(\tilde{w}^{t-K+1}\right) + \frac{1}{2}\nabla f_{V,x_{i(t)}}\left(\tilde{w}^t\right) - \nabla f(\tilde{w}^t) + \nabla f(\tilde{w}^t) \Bigg\|_2^2 \\
=\ & f(\tilde{w}^t) - \left(\gamma_t - \frac{L\gamma_t^2}{2}\right)\left\|\nabla f(\tilde{w}^t)\right\|_2^2 + \frac{L\gamma_t^2}{2}\mathbb{E}\Bigg\| \sum_{k=1}^{K} \nabla f_{\mathcal{G}(k),x_{i(t-K+k)}}\left(\tilde{w}^{t-K+k}\right) \\
& + \frac{1}{2}\nabla f_{V,x_{i(t-K+1)}}\left(\tilde{w}^{t-K+1}\right) + \frac{1}{2}\nabla f_{V,x_{i(t)}}\left(\tilde{w}^t\right) - \nabla f(\tilde{w}^t) \Bigg\|_2^2 \\
& - \left(\gamma_t - L\gamma_t^2\right)\nabla f(\tilde{w}^t)^T \Bigg(\sum_{k=1}^{K} \nabla f_{\mathcal{G}(k)}\left(\tilde{w}^{t-K+k}\right) + \frac{1}{2}\nabla f_V\left(\tilde{w}^{t-K+1}\right) + \frac{1}{2}\nabla f_V\left(\tilde{w}^t\right) - \nabla f(\tilde{w}^t) \Bigg) \\
=\ & f(\tilde{w}^t) - \left(\gamma_t - \frac{L\gamma_t^2}{2}\right)\left\|\nabla f(\tilde{w}^t)\right\|_2^2 + \underbrace{\frac{L\gamma_t^2}{2}\mathbb{E}\Bigg\| \sum_{k=1}^{K} \nabla f_{\mathcal{G}(k),x_{i(t-K+k)}}\left(\tilde{w}^{t-K+k}\right) - \sum_{k=1}^{K} \nabla f_{\mathcal{G}(k)}(\tilde{w}^t) \Bigg\|_2^2}_{Q_1} \\
& \underbrace{+ \frac{L\gamma_t^2}{2}\mathbb{E}\Bigg\| \frac{1}{2}\nabla f_{V,x_{i(t-K+1)}}\left(\tilde{w}^{t-K+1}\right) + \frac{1}{2}\nabla f_{V,x_{i(t)}}\left(\tilde{w}^t\right) - \nabla f_V(\tilde{w}^t) \Bigg\|_2^2}_{Q_2} \\
& \underbrace{- \left(\gamma_t - L\gamma_t^2\right)\nabla f(\tilde{w}^t)^T \Bigg(\sum_{k=1}^{K} \nabla f_{\mathcal{G}(k)}\left(\tilde{w}^{t-K+k}\right) + \frac{1}{2}\nabla f_V\left(\tilde{w}^{t-K+1}\right) + \frac{1}{2}\nabla f_V\left(\tilde{w}^t\right) - \nabla f(\tilde{w}^t) \Bigg)}_{Q_3}
\end{aligned} \tag{20}
$$

where the equalities follow from the unbiased gradient $\mathbb{E}\left[\nabla f_{x_i}(w)\right] = \nabla f(w)$ and $\left[\nabla f_{\mathcal{G}(k)}(w)\right]_j = 0,\ \forall j \notin \mathcal{G}(k)$. Because of $\|x+y\|_2^2 \leq 2\|x\|_2^2 + 2\|y\|_2^2$, we have the upper bound of Q_1 as follows:

$$
\begin{aligned}
& Q_1 \\
=\ & \frac{L\gamma_t^2}{2}\mathbb{E}\left\| \sum_{k=1}^{K} \nabla f_{\mathcal{G}(k),x_{i(t-K+k)}}(\tilde{w}^{t-K+k}) - \sum_{k=1}^{K} \nabla f_{\mathcal{G}(k)}(\tilde{w}^t) - \sum_{k=1}^{K} \nabla f_{\mathcal{G}(k)}(\tilde{w}^{t-K+k}) + \sum_{k=1}^{K} \nabla f_{\mathcal{G}(k)}(\tilde{w}^{t-K+k}) \right\|_2^2 \\
\leq\ & L\gamma_t^2\, \underbrace{\mathbb{E}\left\| \sum_{k=1}^{K} \nabla f_{\mathcal{G}(k),x_{i(t-K+k)}}(\tilde{w}^{t-K+k}) - \sum_{k=1}^{K} \nabla f_{\mathcal{G}(k)}(\tilde{w}^{t-K+k}) \right\|_2^2}_{Q_4} \\
& + L\gamma_t^2 \underbrace{\left\| \sum_{k=1}^{K} \nabla f_{\mathcal{G}(k)}(\tilde{w}^{t-K+k}) - \sum_{k=1}^{K} \nabla f_{\mathcal{G}(k)}(\tilde{w}^t) \right\|_2^2}_{Q_5}.
\end{aligned} \tag{21}
$$

Similarly, we get the upper bound of Q_2 as follows:

$$
\begin{aligned}
Q_2 \leq\ & \frac{L\gamma_t^2}{4}\mathbb{E}\left\|\nabla f_{V,x_{i(t-K+k)}}(\tilde{w}^{t-K+1}) - \nabla f_V(\tilde{w}^t)\right\|_2^2 + \frac{L\gamma_t^2}{4}\left\|\nabla f_{V,x_{i(t)}}(\tilde{w}^t) - \nabla f_V(\tilde{w}^t)\right\|_2^2 \\
\leq\ & \frac{L\gamma_t^2}{2}\underbrace{\mathbb{E}\left\|\nabla f_{V,x_{i(t-K+k)}}(\tilde{w}^{t-K+1}) - \nabla f_V(\tilde{w}^{t-K+1})\right\|_2^2}_{Q_6} + \frac{L\gamma_t^2}{2}\underbrace{\mathbb{E}\left\|\nabla f_V(\tilde{w}^{t-K+1}) - \nabla f_V(\tilde{w}^t)\right\|_2^2}_{Q_7} \\
& + \frac{L\gamma_t^2}{4}\left\|\nabla f_{V,x_{i(t)}}(\tilde{w}^t) - \nabla f_V(\tilde{w}^t)\right\|_2^2.
\end{aligned} \tag{22}
$$

Because of $xy \leq \frac{1}{2}\|x\|_2^2 + \frac{1}{2}\|y\|_2^2$, we have:

$$\begin{aligned} Q_3 \quad \leq \quad & \frac{\gamma_t - L\gamma_t^2}{2} \left\|\nabla f(\tilde{w}^t)\right\|_2^2 + \frac{\gamma_t - L\gamma_t^2}{2} \left\| \sum_{k=1}^{K} \nabla f_{\mathcal{G}(k)} \left(\tilde{w}^{t-K+k}\right) - \sum_{k=1}^{K} \nabla f_{\mathcal{G}(k)} \left(\tilde{w}^{t}\right) \right\|_2^2 \\ & + \frac{\gamma_t - L\gamma_t^2}{8} \left\|\nabla f_V \left(\tilde{w}^{t-K+1}\right) - \nabla f_V(\tilde{w}^t)\right\|_2^2. \end{aligned} \tag{23}$$

According to Assumption 2, we can bound Q_4 as follows:

$$\begin{aligned} Q_4 \quad &= \quad \sum_{k=1}^{K} \mathbb{E} \left\| \nabla f_{\mathcal{G}(k), x_{i(t-K+k)}} (\tilde{w}^{t-K+k}) - \nabla f_{\mathcal{G}(k)} (\tilde{w}^{t-K+k}) \right\|_2^2 \\ &\leq \quad \sum_{k=1}^{K} \mathbb{E} \left\| \nabla f_{\mathcal{G}(k), x_{i(t-K+k)}} (\tilde{w}^{t-K+k}) \right\|_2^2 \\ &\leq \quad KM, \end{aligned} \tag{24}$$

where the first equality follows from the definition of $\nabla f_{\mathcal{G}(k)}(w)$ so that $[\nabla f_{\mathcal{G}(k)}(w)]_j = 0,\ \forall j \notin \mathcal{G}(k)$ and the last inequality is from Assumption 2. Similarly, we can also bound Q_6 as follows:

$$Q_6 \quad \leq \quad M. \tag{25}$$

We can get the upper bound of Q_5:

$$\begin{aligned} Q_5 \quad &= \quad \sum_{k=1}^{K} \left\| \nabla f_{\mathcal{G}(k)} (\tilde{w}^{t-K+k}) - \nabla f_{\mathcal{G}(k)} (\tilde{w}^{t}) \right\|_2^2 \\ &\leq \quad \sum_{k=1}^{K} \left\| \nabla f (\tilde{w}^{t-K+k}) - \nabla f (\tilde{w}^{t}) \right\|_2^2 \\ &\leq \quad L^2 \sum_{k=1}^{K} \left\| \sum_{j=\max\{0,t-K+k\}}^{t-1} \left(\tilde{w}^{j+1} - \tilde{w}^{j}\right) \right\|^2 \\ &\leq \quad L^2 \gamma_{\max\{0,t-K+1\}}^2 K \sum_{k=1}^{K} \sum_{j=\max\{0,t-K+k\}}^{t-1} \left\| \sum_{k=1}^{K} \nabla f_{\mathcal{G}(k), x_{(j-K+k)}} \left(\tilde{w}^{j-K+k}\right) \right. \\ &\quad \left. + \nabla f_{V, x_{(j-K+k)}} \left(\tilde{w}^{j-K+k}\right) + \nabla f_{V, x_{(j)}} \left(\tilde{w}^{j}\right) \right\|_2^2 \\ &\leq \quad KL\gamma_t \frac{\gamma_{\max\{0,t-K+1\}}}{\gamma_t} \sum_{k=1}^{K} \sum_{j=\max\{0,t-K+k\}}^{t-1} \left\| \sum_{k=1}^{K} \nabla f_{\mathcal{G}(k), x_{(j-K+k)}} \left(\tilde{w}^{j-K+k}\right) \right. \\ &\quad \left. + \nabla f_{V, x_{(j-K+k)}} \left(\tilde{w}^{j-K+k}\right) + \nabla f_{V, x_{(j)}} \left(\tilde{w}^{j}\right) \right\|_2^2 \\ &\leq \quad L\gamma_t \sigma K^3 (K+4) M, \end{aligned} \tag{26}$$

where the second inequality is from Assumption 1, the fourth inequality follows from that $L\gamma_t \leq 1$ and the last inequality follows from $\|z_1 + ... + z_r\|_2^2 \leq r(\|z_1\|_2^2 + ... + \|z_r\|_2^2)$, Assumption 2 and $\sigma := \max_t \frac{\gamma_{\max\{0,t-K+1\}}}{\gamma_t}$. Similarly, we can bound Q_7:

$$Q_7 \quad \leq \quad L\gamma_t \sigma K^2 (K+4) M. \tag{27}$$

Integrating the upper bound of Q_1 to Q_7 in (21), we have:

$$\mathbb{E}\left[f(\tilde{w}^{t+1})\right] - f(\tilde{w}^t) \quad \leq \quad -\frac{\gamma_t}{2} \left\|\nabla f(\tilde{w}^t)\right\|^2 + \gamma_t^2 L M_K, \tag{28}$$

where we let $M_K = (K + \frac{3}{4})M + \sigma(\frac{K^2}{2} + K^3)(K+4)M$.

□

Proof to Theorem 1

Proof: When γ_t is constant and $\gamma_t = \gamma$, we have $\sigma = 1$. Because of the definition of M_K and taking total expectation of (15) in Lemma 1, we obtain:

$$\mathbb{E}\left[f(w^{t+1})\right] - f(w^t) \leq -\frac{\gamma}{2}\mathbb{E}\left\|\nabla f(w^t)\right\|_2^2 + \gamma^2 L M_K. \tag{29}$$

Summing (29) from $t = 0$ to $T - 1$, we have:

$$\begin{aligned}\mathbb{E}\left[f(w^T)\right] - f(w^0) \quad \leq \quad & -\frac{\gamma}{2}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(w^t)\right\|_2^2 \\ & + T\gamma^2 L M_K. \end{aligned} \tag{30}$$

Supposing w^* is the optimal solution for $f(w)$, it holds that $f(w^*) - f(w^0) \leq \mathbb{E}\left[f(w^T)\right] - f(w^0)$. Rearranging inequality (30) and dividing both sides by $\frac{2T}{\gamma}$, we complete the proof.

□

Proof to Theorem 2

Proof: $\{\gamma_t\}$ is a diminishing sequence and $\gamma_t = \frac{\gamma_0}{1+t}$, such that $\sigma \leq K$ and $M_K = (K + \frac{3}{4})M + (\frac{K^3}{2} + K^4)(K + 4)M$. Taking total expectation of (15) in Lemma 1 and summing it from $t = 0$ to $T - 1$, we obtain:

$$\begin{aligned}\mathbb{E}\left[f(w^T)\right] - f(w^0) \quad \leq \quad & -\frac{1}{2}\sum_{t=0}^{T-1}\gamma_t\mathbb{E}\left\|\nabla f(w^t)\right\|_2^2 \\ & + \sum_{t=0}^{T-1}\gamma_t^2 L M_K. \end{aligned} \tag{31}$$

Suppose w^* is the optimal solution for $f(w)$; therefore $f(w^*) - f(w^0) \leq \mathbb{E}\left[f(w^T)\right] - f(w^0)$. Rearranging inequality (31) and dividing both sides by T, we complete the proof.

□