Appendices

A Note on Adaptive Unigram Table

Algorithm 1 illustrates the efficient implementation of the adaptive unigram table (*c.f.*, Section 3.2.2). In line 8 and 10, F and $\frac{\tau F}{z}$ are not always integers and therefore they are probabilistically converted into integers as explained in the paper.

Time complexity of Algorithm 1 is $\mathcal{O}(1)$ per update in case of $\alpha = 1.0$. When $|T| < \tau$, the update (line 8) takes $\mathcal{O}(1)$ time since we always have $F = 1$. When $\tau \leq |T|$, we have $\tau \leq z$ and consequently $\frac{\tau F}{z} \leq 1$. This means that the update (line 10–13) takes $\mathcal{O}(1)$ time.

Even if $\alpha \neq 1.0$, the value of z becomes sufficiently large in practice, and thus the update becomes efficient as demonstrated in the experiment.

Algorithm 1 Adaptive unigram table.

```
 1: $f(w) \leftarrow 0$ for all $w \in \mathcal{W}$
 2: $z \leftarrow 0$
 3: for $i = 1, \ldots, n$ do
 4:     $f(w_i) \leftarrow f(w_i) + 1$
 5:     $F \leftarrow f(w_i)^\alpha - (f(w_i) - 1)^\alpha$
 6:     $z \leftarrow z + F$
 7:     if $|T| < \tau$ then
 8:         add $F$ copies of $w_i$ to $T$
 9:     else
10:         for $t = 1, \ldots, \frac{\tau F}{z}$ do
11:             $j$ is randomly drawn from $[1, |T|]$
12:             $T[j] \leftarrow w_i$
13:         end for
14:     end if
15: end for
```

B Complete Proofs

This appendix provides complete proofs of Theorems 1, 3, and 5.

B.1 Proof of Theorem 1

Proof. The first order moment of $\Delta\mathcal{L}(\theta)$ can be rewritten as

$$\begin{aligned}
\mathbb{E}[\Delta\mathcal{L}(\theta)] &= \mathbb{E}\Bigg[\frac{2ck}{n}\sum_{w\in\mathcal{W}}\sum_{v\in\mathcal{W}}\sum_{i=1}^{n}\delta_{w_i,w}(q_i(v)-q_n(v))\psi^{-}_{w,v}\Bigg] \\
&= \frac{2ck}{n}\sum_{w\in\mathcal{W}}\sum_{v\in\mathcal{W}}\sum_{i=1}^{n}\mathbb{E}[\delta_{w_i,w}(q_i(v)-q_n(v))\psi^{-}_{w,v}] \\
&= \frac{2ck}{n}\sum_{w\in\mathcal{W}}\sum_{v\in\mathcal{W}}\sum_{i=1}^{n}\mathbb{E}[\mathrm{X}_{i,w}(\mathrm{Y}_{i,v}-\mathrm{Y}_{n,v})\psi^{-}_{w,v}] \\
&= \frac{2ck}{n}\sum_{w\in\mathcal{W}}\sum_{v\in\mathcal{W}}\sum_{i=1}^{n}\Big(\mathbb{E}[\mathrm{X}_{i,w}\mathrm{Y}_{i,v}]-\mathbb{E}[\mathrm{X}_{i,w}\mathrm{Y}_{n,v}]\Big)\psi^{-}_{w,v}.
\end{aligned}$$

Here, for any i and j such that $i \le j$, we have

$$\begin{aligned}
\mathbb{E}[\mathrm{X}_{i,w}\mathrm{Y}_{j,v}] &= \mathbb{E}[\mathrm{X}_{i,w}\frac{1}{j}\sum_{j'=1}^{j}\mathrm{X}_{j',v}] = \frac{1}{j}\sum_{j'=1}^{j}\mathbb{E}[\mathrm{X}_{i,w}\mathrm{X}_{j',v}] \\
&= \frac{1}{j}\sum_{j'=1}^{j}\Big(\mathbb{E}[\mathrm{X}_{i,w}]\mathbb{E}[\mathrm{X}_{j',v}]+\mathbb{V}[\mathrm{X}_{i,w},\mathrm{X}_{j',v}]\Big) \\
&= \mu_w\mu_v + \frac{1}{j}\rho_{w,v}.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
\mathbb{E}[\Delta\mathcal{L}(\theta)] &= \frac{2ck}{n}\sum_{w\in\mathcal{W}}\sum_{v\in\mathcal{W}}\sum_{i=1}^{n}\Big(\mu_w\mu_v+\frac{1}{i}\rho_{w,v}-\mu_w\mu_v-\frac{1}{n}\rho_{w,v}\Big)\psi^{-}_{w,v} \\
&= \frac{2ck(H_n-1)}{n}\sum_{w\in\mathcal{W}}\sum_{v\in\mathcal{W}}\rho_{w,v}\psi^{-}_{w,v}.
\end{aligned}$$

□

B.2 Proof of Theorem 3

To prove Theorem 3, we begin by examining the upper- and lower-bounds of $\mathbb{E}[\mathrm{X}_{i,w}\mathrm{Y}_{j,v}\mathrm{Y}_{k,v}]$ in the following Lemma, and then make use of the bounds to evaluate the order of the second order moment of $\Delta\mathcal{L}(\theta)$.

Lemma. *For any j and k such that $j \le k$, we have*

$$\begin{aligned}
\mathbb{E}[\mathrm{X}_{i,w}\mathrm{Y}_{j,v}\mathrm{Y}_{k,v}] &\le \frac{(jk-2j-k+2)\mu_w\mu_v^2+2j+k-2}{jk}, \\
\mathbb{E}[\mathrm{X}_{i,w}\mathrm{Y}_{j,v}\mathrm{Y}_{k,v}] &\ge \frac{(jk-2j-k+2)\mu_w\mu_v^2}{jk}.
\end{aligned}$$

Proof. We have

$$\begin{aligned}
\mathbb{E}[\mathrm{X}_{i,w}\mathrm{Y}_{j,v}\mathrm{Y}_{k,v}] &= \mathbb{E}[\mathrm{X}_{i,w}\left(\frac{1}{j}\sum_{l=1}^{j}\mathrm{X}_{l,v}\right)\left(\frac{1}{k}\sum_{m=1}^{k}\mathrm{X}_{m,v}\right)] \\
&= \sum_{l=1}^{j}\sum_{m=1}^{k}\frac{\mathbb{E}[\mathrm{X}_{i,w}\mathrm{X}_{l,v}\mathrm{X}_{m,v}]}{jk}.
\end{aligned}$$

To prove the lemma, we rewrite the expression by splitting the set of (l,m) into two subsets. Let $\mathcal{S}_i^{(j,k)}$ $(j \le k)$ be a set of (l,m) such that $\mathrm{X}_{i,w}$, $\mathrm{X}_{l,v}$, and $\mathrm{X}_{m,v}$ are independent from each other (*i.e.*, i, l, and m are all different), and let $\bar{\mathcal{S}}_i^{(j,k)}$ be its complementary set:

$$\begin{aligned}
\mathcal{S}_i^{(j,k)} &= \{(l,m) \in \{1,2,\ldots,j\} \times \{1,2,\ldots,k\} \mid i \neq l \wedge l \neq m \wedge m \neq i\}, \\
\bar{\mathcal{S}}_i^{(j,k)} &= \{1,2,\ldots,j\} \times \{1,2,\ldots,k\} \setminus \mathcal{S}_i^{(j,k)}.
\end{aligned}$$

Then, $\mathbb{E}[\mathrm{X}_{i,w}\mathrm{Y}_{j,v}\mathrm{Y}_{k,v}]$ is upper-bounded as

$$\begin{aligned}
\mathbb{E}[\mathrm{X}_{i,w}\mathrm{Y}_{j,v}\mathrm{Y}_{k,v}] &= \sum_{(l,m)\in\mathcal{S}_i^{(j,k)}}\frac{\mathbb{E}[\mathrm{X}_{i,w}]\mathbb{E}[\mathrm{X}_{l,v}]\mathbb{E}[\mathrm{X}_{m,v}]}{jk} + \sum_{(l,m)\in\bar{\mathcal{S}}_i^{(j,k)}}\frac{\mathbb{E}[\mathrm{X}_{i,w}\mathrm{X}_{l,v}\mathrm{X}_{m,v}]}{jk} \\
&\le \sum_{(l,m)\in\mathcal{S}_i^{(j,k)}}\frac{\mu_w\mu_v^2}{jk} + \sum_{(l,m)\in\bar{\mathcal{S}}_i^{(j,k)}}\frac{1}{jk} \\
&= \frac{|\mathcal{S}_i^{(j,k)}|\mu_w\mu_v^2 + |\bar{\mathcal{S}}_i^{(j,k)}|}{jk},
\end{aligned}$$

where the inequality holds because $\mathrm{X}_{i,w}$, $\mathrm{X}_{l,v}$, and $\mathrm{X}_{m,v}$ are binary random variables and thus $\mathbb{E}[\mathrm{X}_{i,w}\mathrm{X}_{l,v}\mathrm{X}_{m,v}] \le 1$. Here, we have $|\bar{\mathcal{S}}_i^{(j,k)}| = 2j + k - 2$, because $\bar{\mathcal{S}}_i^{(j,k)}$ includes j elements such that $l = m$ and also includes $k-1$ and $j-1$ elements such that $i = l \neq m$ and $i = m \neq l$, respectively. And we consequently have $|\mathcal{S}_i^{(j,k)}| = jk - |\bar{\mathcal{S}}_i^{(j,k)}| = jk - 2j - k + 2$. Therefore, the upper-bound can be rewritten as

$$\mathbb{E}[\mathrm{X}_{i,w}\mathrm{Y}_{j,v}\mathrm{Y}_{k,v}] \le \frac{(jk - 2j - k + 2)\mu_w\mu_v^2 + 2j + k - 2}{jk}.$$

Similarly, by making use of $0 \le \mathbb{E}[\mathrm{X}_{i,w}\mathrm{X}_{l,v}\mathrm{X}_{m,v}]$, the lower-bound can be derived:

$$\begin{aligned}
\mathbb{E}[\mathrm{X}_{i,w}\mathrm{Y}_{j,v}\mathrm{Y}_{k,v}] &= \sum_{(l,m)\in\mathcal{S}_i^{(j,k)}}\frac{\mathbb{E}[\mathrm{X}_{i,w}]\mathbb{E}[\mathrm{X}_{l,v}]\mathbb{E}[\mathrm{X}_{m,v}]}{jk} + \sum_{(l,m)\in\bar{\mathcal{S}}_i^{(j,k)}}\frac{\mathbb{E}[\mathrm{X}_{i,w}\mathrm{X}_{l,v}\mathrm{X}_{m,v}]}{jk} \\
&\ge \sum_{(l,m)\in\mathcal{S}_i^{(j,k)}}\frac{\mu_w\mu_v^2}{jk} + \sum_{(l,m)\in\bar{\mathcal{S}}_i^{(j,k)}}\frac{0}{jk} \\
&= \frac{|\mathcal{S}_i^{(j,k)}|\mu_w\mu_v^2}{jk} = \frac{(jk - 2j - k + 2)\mu_w\mu_v^2}{jk}.
\end{aligned}$$

□

Making use the above Lemma, we can prove Theorem 3.

Proof. The upper-bound of $\mathbb{E}[\Delta\mathcal{L}(\theta)^2]$ is examined to prove the theorem. Let $\Psi_{i,n,w,v} = \delta_{w_i,w}(q_i(v) - q_n(v))\psi^-_{w,v}$. Making use of Jensen's inequality, we have

$$\begin{aligned}
\mathbb{E}[\Delta\mathcal{L}(\theta)^2] &= \mathbb{E}\biggl[\frac{4c^2k^2}{n^2}\biggl(\sum_{w\in\mathcal{W}}\sum_{v\in\mathcal{W}}\sum_{i=1}^{n}\Psi_{i,n,w,v}\biggr)^2\biggr] \\
&= \mathbb{E}\biggl[\frac{4c^2k^2}{n^2}|\mathcal{W}|^4n^2\biggl(\sum_{w\in\mathcal{W}}\sum_{v\in\mathcal{W}}\sum_{i=1}^{n}\frac{1}{|\mathcal{W}|^2n}\Psi_{i,n,w,v}\biggr)^2\biggr] \\
&\le \mathbb{E}\biggl[\frac{4c^2k^2}{n^2}|\mathcal{W}|^4n^2\sum_{w\in\mathcal{W}}\sum_{v\in\mathcal{W}}\sum_{i=1}^{n}\frac{1}{|\mathcal{W}|^2n}\Psi_{i,n,w,v}\biggr] \\
&= \frac{4c^2k^2|\mathcal{W}|^2}{n}\sum_{w\in\mathcal{W}}\sum_{v\in\mathcal{W}}\sum_{i=1}^{n}\mathbb{E}[\Psi^2_{i,n,w,v}].
\end{aligned}$$

Furthermore, the term $\mathbb{E}[\Psi^2_{i,n,w,v}]$ is upper-bounded as

$$\begin{aligned}
\mathbb{E}[\Psi^2_{i,n,w,v}] &= \mathbb{E}[\delta^2_{w_i,v}(q_i(v) - q_n(v))^2(\psi^-_{w,v})^2] \\
&= \mathbb{E}[\delta_{w_i,v}(q_i(v) - q_n(v))^2(\psi^-_{w,v})^2] \\
&= \mathbb{E}[\mathrm{X}_{i,w}(\mathrm{Y}_{i,v} - \mathrm{Y}_{n,v})^2](\psi^-_{w,v})^2 \\
&= (\mathbb{E}[\mathrm{X}_{i,w}\mathrm{Y}^2_{i,v}] - 2\mathbb{E}[\mathrm{X}_{i,w}\mathrm{Y}_{i,v}\mathrm{Y}_{n,v}] + \mathbb{E}[\mathrm{X}_{i,w}\mathrm{Y}^2_{n,v}])(\psi^-_{w,v})^2 \\
&\le \biggl\{\frac{1}{i^2}\biggl((i^2 - 3i + 2)\mu_w\mu_v^2 + 3i - 2\biggr) \\
&\quad - 2\frac{1}{in}(in - 2i - n + 2)\mu_w\mu_v^2 \\
&\quad + \frac{1}{n^2}\biggl((n^2 - 3n + 2)\mu_w\mu_v^2 + 3n - 2\biggr)\biggr\}(\psi^-_{w,v})^2 \\
&= \biggl\{(2\mu_w\mu_v^2 - 2)\frac{1}{i^2} + (-\mu_w\mu_v^2 - \frac{4}{n}\mu_w\mu_v^2 + 3)\frac{1}{i} \\
&\quad + (2\mu_w\mu_v^2 - 2)\frac{1}{n^2} + (\mu_w\mu_v^2 + 3)\frac{1}{n}\biggr\}(\psi^-_{w,v})^2,
\end{aligned}$$

where the above Lemma is used to derive the inequality. Therefore, we have

$$\begin{aligned}\sum_{i=1}^{n}\mathbb{E}[\Psi_{i,n,w,v}^{2}] &\leq \sum_{i=1}^{n}\Bigg\{(2\mu_w\mu_v^2-2)\frac{1}{i^2}+(-\mu_w\mu_v^2-\frac{4}{n}\mu_w\mu_v^2+3)\frac{1}{i}\\ &\quad+(2\mu_w\mu_v^2-2)\frac{1}{n^2}+(\mu_w\mu_v^2+3)\frac{1}{n}\Bigg\}(\psi_{w,v}^{-})^2\\ &=\Bigg\{(2\mu_w\mu_v^2-2)H_{n,2}+(-\mu_w\mu_v^2-\frac{4}{n}\mu_w\mu_v^2+3)H_n\\ &\quad+(2\mu_w\mu_v^2-2)\frac{1}{n}+(\mu_w\mu_v^2+3)\Bigg\}(\psi_{w,v}^{-})^2,\end{aligned}$$

where $H_{n,2}$ represents the generalized harmonic number of order n of 2. Since $H_{n,2} \leq H_n = \mathcal{O}(\log(n))$, we have $\sum_{i=1}^{n}\mathbb{E}[\Psi_{i,n,w,v}^{2}] = \mathcal{O}(\log(n))$ and consequently $\mathbb{E}[\Delta\mathcal{L}(\theta)^2] = \mathcal{O}(\frac{\log(n)}{n})$. □

B.3 Proof of Theorem 5

Proof. The proof is made by the squeeze theorem. Let $l = \mathcal{L}_{\mathrm{B}}(\hat{\theta}) - \mathcal{L}_{\mathrm{B}}(\theta^*)$. Then, Chebyshev's inequality gives, for any $\epsilon_1 > 0$,

$$\begin{aligned}\lim_{n\to\infty}\frac{\mathbb{V}[l]}{\epsilon_1^2} &\geq \lim_{n\to\infty}\Pr\Big[|l-\mathbb{E}[l]|\geq\epsilon_1\Big]\\ &=\lim_{n\to\infty}\Pr\Big[l-\mathbb{E}[l]\leq-\epsilon_1\Big]+\Pr\Big[\epsilon_1\leq l-\mathbb{E}[l]\Big]\\ &=\lim_{n\to\infty}\Pr\Big[l\leq\mathbb{E}[l]-\epsilon_1\Big]+\Pr\Big[\mathbb{E}[l]+\epsilon_1\leq l\Big].\end{aligned}$$

Remind that Eq. (2) in Lemma 4 means that for any $\epsilon_2 > 0$, there exists n' such that if $n' \leq n$ then $|\mathbb{E}[l]| < \epsilon_2$. Therefore we have

$$\begin{aligned}\lim_{n\to\infty}\frac{\mathbb{V}[l]}{\epsilon_1^2} &\geq \lim_{n\to\infty}\Pr\Big[l\leq\mathbb{E}[l]-\epsilon_1\Big]+\Pr\Big[\mathbb{E}[l]+\epsilon_1\leq l\Big]\\ &\geq\lim_{n\to\infty}\Pr\Big[l\leq-\epsilon_2-\epsilon_1\Big]+\Pr\Big[\epsilon_2+\epsilon_1\leq l\Big]\\ &=\lim_{n\to\infty}\Pr\Big[|l|\geq\epsilon_1+\epsilon_2\Big]\geq 0.\end{aligned}$$

The arbitrary property of ϵ_1 and ϵ_2 allows $\epsilon_1+\epsilon_2$ to be rewritten as ϵ. Also, Eq. (3) in Lemma 4 implies that $\lim_{n\to\infty}\frac{\mathbb{V}[l]}{\epsilon_1^2} = 0$. Therefore, the squeeze theorem gives the proof. □

C Theoretical Analysis in Smoothed Case

This appendix investigates the convergence of the first and second order moment of $\Delta\mathcal{L}(\theta)$ in the smoothed case.

C.1 Convergence of the first order moment of $\Delta\mathcal{L}(\theta)$

The first order moment of $\Delta\mathcal{L}(\theta)$ in the smoothed case is given as

$$\mathbb{E}[\Delta\mathcal{L}(\theta)] = \frac{2ck}{n}\sum_{w\in\mathcal{W}}\sum_{v\in\mathcal{W}}\sum_{i=1}^{n}\Big(\mathbb{E}[\mathrm{X}_{i,w}\mathrm{Z}_{i,v}] - \mathbb{E}[\mathrm{X}_{i,w}\mathrm{Z}_{n,v}]\Big)\psi^{-}_{w,v}.$$

Let us investigate $\mathbb{E}[\mathrm{X}_{i,w}\mathrm{Z}_{j,v}]$ as we did $\mathbb{E}[\mathrm{X}_{i,w}\mathrm{Y}_{j,v}]$ in the unsmoothed case. Let $\phi_w = g_w(\mu) - \sum_{v\in\mathcal{W}} M_{w,v}g_v(\mu)$. Then, for any i and j such that $i \le j$, we have

$$\begin{aligned}
\mathbb{E}[\mathrm{X}_{i,w}\mathrm{Z}_{j,v}] &\approx \mathbb{E}[\mathrm{X}_{i,w}\Big(g_v(\mu) + \sum_{v'\in\mathcal{W}} M_{v,v'}(\mathrm{Y}_{j,v'} - g_{v'}(\mu))\Big)] \\
&= \mathbb{E}[\mathrm{X}_{i,w}(\sum_{v'\in\mathcal{W}} M_{v,v'}\mathrm{Y}_{j,v'} + \phi_v)] \\
&= \sum_{v'\in\mathcal{W}} M_{v,v'}E[\mathrm{X}_{i,w}\mathrm{Y}_{j,v'}] + \phi_v E[\mathrm{X}_{i,w}] \\
&= \sum_{v'\in\mathcal{W}} M_{v,v'}(\mu_w\mu_{v'} + \frac{1}{j}\rho_{w,v'}) + \mu_w\phi_v \\
&= \sum_{v'\in\mathcal{W}} M_{v,v'}\mu_w\mu_{v'} + \mu_w\phi_v + \frac{1}{j}\sum_{v'\in\mathcal{W}} M_{v,v'}\rho_{w,v'}.
\end{aligned}$$

Therefore, plugging the above equation into $\mathbb{E}[\Delta\mathcal{L}(\theta)]$ yields $\mathbb{E}[\Delta\mathcal{L}(\theta)] \approx \mathcal{O}(\frac{\log(n)}{n})$.

C.2 Convergence of the second order moment of $\Delta\mathcal{L}(\theta)$

Next, let us examine the convergence of the second order moment of $\Delta\mathcal{L}(\theta)$. This can be confirmed by inspecting $\mathbb{E}[\mathrm{X}_{i,w}\mathrm{Z}_{j,v}\mathrm{Z}_{k,v}]$ and then $\mathbb{E}[\Psi^2_{i,n,w,v}]$ analogously to the unsmoothed case.

For any i, j, and k such that $i \le j \le k$, we have

$$\begin{aligned}
\mathbb{E}[\mathrm{X}_{i,w}\mathrm{Z}_{j,v}\mathrm{Z}_{k,v}] &\approx \mathbb{E}[\mathrm{X}_{i,w}\Big(\sum_{v'\in\mathcal{W}} M_{v,v'}\mathrm{Y}_{j,v'} + \phi_v\Big)\Big(\sum_{v''\in\mathcal{W}} M_{v,v''}\mathrm{Y}_{k,v''} + \phi_v\Big)] \\
&= \sum_{v'\in\mathcal{W}}\sum_{v''\in\mathcal{W}} M_{v,v'}M_{v,v''}\mathbb{E}[\mathrm{X}_{i,w}\mathrm{Y}_{j,v'}\mathrm{Y}_{k,v''}] \\
&\quad + \sum_{v'\in\mathcal{W}} M_{v,v'}\phi_v\mathbb{E}[\mathrm{X}_{i,w}\mathrm{Y}_{j,v'}] + \sum_{v''\in\mathcal{W}} M_{v,v''}\phi_v\mathbb{E}[\mathrm{X}_{i,w}\mathrm{Y}_{k,v''}] + \phi_v^2\mathbb{E}[\mathrm{X}_{i,w}] \\
&= \sum_{v'\in\mathcal{W}}\sum_{v''\in\mathcal{W}} M_{v,v'}M_{v,v''}\mathbb{E}[\mathrm{X}_{i,w}\mathrm{Y}_{j,v'}\mathrm{Y}_{k,v''}] \\
&\quad + \sum_{v'\in\mathcal{W}} M_{v,v'}\phi_v(\mu_w\mu_{v'} + \frac{1}{j}\rho_{w,v'}) \\
&\quad + \sum_{v''\in\mathcal{W}} M_{v,v''}\phi_v(\mu_w\mu_{v''} + \frac{1}{k}\Sigma_{w,v''}) + \mu_w\phi_v^2.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}\mathbb{E}[\Psi_{i,n,w,v}^2] &= \mathbb{E}[\mathrm{X}_{i,w}(\mathrm{Z}_{i,v} - \mathrm{Z}_{n,v})^2]\psi_{w,v}^2 \\ &\approx \sum_{v' \in \mathcal{W}} \sum_{v'' \in \mathcal{W}} M_{v,v'} M_{v,v''} \Big(\mathbb{E}[\mathrm{X}_{i,w}\mathrm{Y}_{i,v'}\mathrm{Y}_{i,v''}] \\ &\quad - 2\mathbb{E}[\mathrm{X}_{i,w}\mathrm{Y}_{i,v'}\mathrm{Y}_{n,v''}] + \mathbb{E}[\mathrm{X}_{i,w}\mathrm{Y}_{n,v'}\mathrm{Y}_{n,v''}] \Big) \psi_{w,v}^2.\end{aligned}$$

Using similar bounds to Lemma 3, we also have $\sum_{i=1}^{n} \mathbb{E}[\Psi_{i,n,w,v}^2] \approx \mathcal{O}(\log(n))$ and consequently $\mathbb{E}[\Delta\mathcal{L}(\theta)^2] \approx \mathcal{O}(\frac{\log(n)}{n})$.

D Theoretical Analysis of Mini-batch SGNS

This appendix demonstrates that Theorems 2 and 3 also hold for the mini-batch SGNS, that is, the first and second order moments of $\Delta\mathcal{L}(\theta)$ are in the order of $\mathcal{O}(\frac{\log(n)}{n})$. We here investigate the mini-batch setting in which M words, as opposed to a single word in the case of incremental SGNS, are processed at a time.

Definition. Let $\mathrm{Y}_{i,w}^{(M)}$ be a random variable that represents $q_i(w)$ when $\alpha = 1.0$ and the mini-batch size is M. Then, it is given as

$$\mathrm{Y}_{i,w}^{(M)} = \mathrm{Y}_{b(i,M),w}$$

where $b(i, M) = \lceil \frac{i}{M} \rceil \times M$. Note that we always have $\mathrm{Y}_{n,w}^{(M)} = \mathrm{Y}_{n,w}$ and $i \leq b(i, M)$.

We first examine the first order moment of $\Delta\mathcal{L}(\theta)$ by taking a similar step as the proof of Theorem 1. The first order moment of $\Delta\mathcal{L}(\theta)$ is given as

$$\begin{aligned}\mathbb{E}[\Delta\mathcal{L}(\theta)] &= \frac{2ck}{n} \sum_{w \in \mathcal{W}} \sum_{v \in \mathcal{W}} \sum_{i=1}^{n} \Big(\mathbb{E}[\mathrm{X}_{i,w}\mathrm{Y}_{j,v}^{(M)}] - \mathbb{E}[\mathrm{X}_{i,w}\mathrm{Y}_{n,v}^{(M)}] \Big) \psi_{w,v}^{-} \\ &= \frac{2ck}{n} \sum_{w \in \mathcal{W}} \sum_{v \in \mathcal{W}} \sum_{i=1}^{n} \Big(\mathbb{E}[\mathrm{X}_{i,w}\mathrm{Y}_{j,v}^{(M)}] - \mathbb{E}[\mathrm{X}_{i,w}\mathrm{Y}_{n,v}] \Big) \psi_{w,v}^{-} \\ &= \frac{2ck}{n} \sum_{w \in \mathcal{W}} \sum_{v \in \mathcal{W}} \rho_{w,v} \psi_{w,v}^{-} \Big(\sum_{i=1}^{n} \frac{1}{b(i, M)} - \sum_{i=1}^{n} \frac{1}{n} \Big).\end{aligned}$$

Because we have

$$\sum_{i=1}^{n} \frac{1}{b(i, M)} \leq \sum_{i=1}^{n} \frac{1}{i} = H_n = \mathcal{O}(\log(n)), \tag{1}$$

we have $\mathbb{E}[\Delta\mathcal{L}(\theta)] = \mathcal{O}(\frac{\log(n)}{n})$.

Next, we investigate the second order moment of $\mathbb{E}[\Delta\mathcal{L}(\theta)]$. Analogously to the last inequality of the proof of Theorem 3, we have

$$\sum_{i=1}^{n} \mathbb{E}[\Psi_{i,n,w,v}^2] \leq \sum_{i=1}^{n} \bigg\{ (2\mu_w\mu_v^2 - 2)\frac{1}{b(i,M)^2} + (-\mu_w\mu_v^2 - \frac{4}{n}\mu_w\mu_v^2 + 3)\frac{1}{b(i,M)} + (2\mu_w\mu_v^2 - 2)\frac{1}{n^2} + (\mu_w\mu_v^2 + 3)\frac{1}{n} \bigg\} (\psi_{w,v}^-)^2.$$

Since we have

$$\sum_{i=1}^{n} \frac{1}{b(i,M)^2} \leq \sum_{i=1}^{n} \frac{1}{i^2} = H_{n,2} = \mathcal{O}(\log(n)), \tag{2}$$

it can be proven that $\mathbb{E}[\Delta\mathcal{L}(\theta)^2] = \mathcal{O}(\frac{\log(n)}{n})$.

E Experimental Configurations

This appendix details the experimental configurations that are not described in the paper.

E.1 Verification of theorems

The vocabulary set in the Gigaword corpus was reduced to 1000 by converting infrequent words into the same special tokens because it is expensive to evaluate the expectation terms in $\Delta\mathcal{L}(\theta)$ for a large vocabulary set.

The parameter θ was set to 100-dimensional vectors each element of which is drawn from $[-0.5, 0.5]$ uniformly at random. In preliminary experiments we confirmed that the result is not sensitive to the choice of the parameter value. Note that the same parameter value was used for all n. We set c and k as $c = 5$ and $k = 5$.

The mean μ_w and covariances $\rho_{w,v}$ are required to compute the theoretical value of the first order moment. They were given as the maximum likelihood estimations from the entire Gigaword corpus.

E.2 Quality of word embeddings

Table 1 summarizes the training configurations. Those parameter values were used for both incremental and batch SGNS. The learning rate was set to 0.1 for **incremental** and **batch**, which use AdaGrad to adjust the learning rate. On the other hand, the learning rate of **w2v**, which uses linear decay function to adjust the learning rate, was set as the default value of 0.025.

In the word similarity and the analogy tasks, we use $\mathbf{t}_w + \mathbf{c}_w$ as an embedding of the word w [2, 1]. The analogy task was performed by using 3CosMul [1].

Parameter	**Value**
Embedding size	400
Number negative samples	10
Subsampling threshold	1.0×10^{-5}
Subsampling method	dirty
Window size	10
Smoothing parameter α	0.75

Table 1: Training configurations. Incremental SGNS used the incrementally-updated frequency for the subsampling.

References

[1] Omer Levy, Yoav Goldberg, and Ido Dagan. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225, 2015.

[2] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of EMNLP*, pages 1532–1543, 2014.