word2vec gradients Charlie Colley 9/29/17

$\sigma(x) = \frac{1}{1+e^x}$

$$\frac{\partial \sigma}{\partial x}(x) = \frac{\partial}{\partial x}(\frac{e^x}{1+e^x}) = \frac{(1+e^x)e^x - e^x(e^x)}{(1+e^x)^2} = (\frac{e^x}{1+e^x})(\frac{1-e^x+e^x}{(1+e^x)})$$

$$= \sigma(x)(\frac{1+e^x}{1+e^x} - \frac{e^x}{(1+e^x)}) = \sigma(x)(1-\sigma(x))$$

Thus $\frac{\partial \sigma(x)}{\partial x}(x) = \sigma(x)(1-\sigma(x))$.

Objective function:
$\arg\max_{V_c, V_w} = \Sigma_{(w,c)\in D} \log(\sigma(\langle v_c, v_w \rangle)) + \Sigma_{(w,c)v\in D'} \log(\sigma(-\langle v_c, v_w \rangle))$

$$\nabla_{v_c}(\langle v_c, v_w \rangle)_k = \frac{\partial}{\partial (v_c)_k} \sum_i (v_c)_i (v_w)_i = \frac{\partial}{\partial (v_c)_k} (\sum_{i\neq k} (v_c)_i (v_w)_i + (v_c)_k (v_w)_k) = (v_w)_k$$

by symmetry, the gradient $\nabla_{v_w}\langle v_c, v_w \rangle$ is v_c. Let $B \in \mathbb{R}^{n\times m}$.

$$\nabla_B(\langle v_c, v_w \rangle_B)_{kl} = \frac{\partial}{\partial B_{kl}} \sum_{ij} B_{ij} (v_c)_i (v_w)_i =$$

$$\frac{\partial}{\partial B_{kl}} (\sum_{i,j\neq k} (v_c)_i (v_w)_i + B_{kl}(v_c)_k (v_w)_l) = (v_c)_l (v_w)_k = (v_c \otimes v_w)_{kl}$$

Then the gradient of $\log(\sigma(\langle v_c, v_w \rangle))$ with respect to v_c is

$$\frac{\partial}{\partial (v_c)_k}(\log(\sigma(\langle v_c, v_w \rangle))) = (\frac{1}{\sigma(\langle v_c, v_w \rangle)})(\sigma(\langle v_c, v_w \rangle)(1-\sigma(\langle v_c, v_w \rangle)))(v_w)_k$$

$$= (1-\sigma(\langle v_c, v_w \rangle)))(v_w)_k$$

again by symmetry we can show that the gradient with respect to v_w is similarly $\nabla_{v_w}(log(\sigma(\langle v_c, v_w \rangle))) = (1-\sigma(\langle v_c, v_w \rangle))v_c$

Let $P \in \mathbb{R}^{n\times n}$ and $U, V \in \mathbb{R}^{n\times d}$ and let $f = \|P - UV^T\|_F^2$

$$
\begin{aligned}
(\nabla_U f)_{ij} =& \frac{\partial}{\partial U_{ij}} \|P - UV^T\|_F^2 = \frac{\partial}{\partial U_{ij}} (\sum_{i_1,i_2=1}^{n} (P_{i_1,i_2} - \sum_{i_3}^{d} U_{i_1,i_3} V^T_{i_3,i_2})^2) \\
=& \frac{\partial}{\partial U_{ij}} ([\sum_{i_1\neq i, i_2=1}^{n} (P_{i_1,i_2} - \sum_{i_3}^{d} U_{i_1,i_3} V^T_{i_3,i_2})^2] + (P_{i,i_2} - \sum_{i_3}^{d} U_{i,i_3} V^T_{i_3,i_2})^2) \\
=& \frac{\partial}{\partial U_{ij}} \sum_{i_2}^{n} (P_{i,i_2} - \sum_{i_3}^{d} U_{i,i_3} V^T_{i_3,i_2})^2 \\
=& \sum_{i_2}^{n} 2(P_{i,i_2} - \sum_{i_3}^{d} U_{i,i_3} V^T_{i_3,i_2}) \frac{\partial}{\partial U_{ij}} (P_{i,i_2} - \sum_{i_3}^{d} U_{i,i_3} V^T_{i_3,i_2}) \\
=& \sum_{i_2} 2(P_{i,i_2} - \sum_{i_3}^{d} U_{i,i_3} V^T_{i_3,i_2})(V^T_{j,i_2}) \\
=& \sum_{i_2} 2(P_{i,i_2} - (UV^T)_{i,i_2})(V_{i_2,j}) = 2(PV - UV^TV)_{ij}
\end{aligned}
$$

$$
(\nabla_V f)_{ij} = \frac{\partial}{\partial V_{ij}} \|P - UV^T\|_F^2 = \frac{\partial}{\partial V_{ij}} \|P^T - VU^T\|_F^2
$$

Now note that by the generality of the last gradient computation we can show that $\nabla_V f = 2(P^TU - VU^TU)$.

Consider the function $g = \lambda \|U\|_F^2$ for $\lambda \in \mathbb{R}$, then the gradient with respect to U is

$$
(\nabla_U g)_{ij} = \lambda \frac{\partial}{\partial U_{ij}} (\sum_{i_1,i_2=1}^{n,d} U^2_{i_1,i_2}) = \lambda \frac{\partial}{\partial U_{ij}} ([\sum_{i_1\neq i, i_2\neq j}^{n,d} U^2_{i_1,i_2}] + U^2_{ij}) = 2\lambda U_{ij}
$$

Let $P \in \mathbb{R}^{n\times n}$ and $U \in \mathbb{R}^{n\times d}$ and let $f = \|P - UU^T\|_F^2$

$$
\begin{aligned}
(\nabla_U f)_{ij} =& \frac{\partial}{\partial U_{ij}} \|P - UU^T\|_F^2 = \frac{\partial}{\partial U_{ij}} \sum_{i_1,i_2}^{n,n} (P_{i_1,i_2} - \sum_{i_3} U_{i_1,i_3} U^T_{i_3,i_2})^2 \\
=& \frac{\partial}{\partial U_{ij}} (\sum_{i_1 \neq i, i_2}^{n,n} (P_{i_1,i_2} - \sum_{i_3} U_{i_1,i_3} U_{i_2,i_3})^2 + \sum_{i_2} (P_{i,i_2} - \sum_{i_3} U_{i,i_3} U_{i_2,i_3})^2) \\
=& \frac{\partial}{\partial U_{ij}} (\sum_{i_1, i_2 \neq i}^{n,n} (P_{i_1,i_2} - \sum_{i_3} U_{i_1,i_3} U_{i_2,i_3})^2 + \sum_{i_1 \neq i} (P_{i_1,i} - \sum_{i_3} U_{i_1,i_3} U_{i,i_3})^2 \\
&+ \sum_{i_2 \neq i} (P_{i,i_2} - \sum_{i_3} U_{i,i_3} U_{i_2,i_3})^2) + (P_{i,i_2} - \sum_{i_3} U_{i,i_3} U_{i,i_3})^2) \\
=& \frac{\partial}{\partial U_{ij}} (\sum_{i_1 \neq i} (P_{i_1,i} - \sum_{i_3} U_{i_1,i_3} U_{i,i_3})^2 + \sum_{i_2 \neq i} (P_{i,i_2} - \sum_{i_3} U_{i,i_3} U_{i_2,i_3})^2 + (P_{i,i_2} - \sum_{i_3} U_{i,i_3} U_{i,i_3})^2) \\
=& \sum_{i_1 \neq i} 2(P_{i_1,i} - \sum_{i_3} U_{i_1,i_3} U_{i,i_3})(-U_{i_1,j}) + \sum_{i_2 \neq i} 2(P_{i,i_2} - \sum_{i_3} U_{i,i_3} U_{i_2,i_3})(-U_{i_2,j}) \\
&+ 2(P_{i,i_2} - \sum_{i_3} U_{i,i_3} U_{i,i_3})(-2U_{ij}) \\
=& \sum_{i_1} 2(P_{i_1,i} - \sum_{i_3} U_{i_1,i_3} U_{i,i_3})(-U_{i_1,j}) + \sum_{i_2} 2(P_{i,i_2} - \sum_{i_3} U_{i,i_3} U_{i_2,i_3})(-U_{i_2,j})) \\
=& \sum_{i_1} 2(P - UU^T)_{i_1,i}(-U_{i_1,j}) + \sum_{i_1} 2(P - UU^T)_{i,i_1}(-U_{i_1,j})) \\
(\nabla_U f) =& -2((P - UU^T)^T(U) + (P - UU^T)U) = -2((P - UU^T)^T + (P - UU^T))U
\end{aligned}
$$

if $P = P^T$ then we get $\nabla_U f = -4(P - UU^T)U$

Let $P \in \mathbb{R}^{n\times n}$, $U \in \mathbb{R}^{n\times d}$, and $B \in \mathbb{R}^{d\times d}$. Consider the function $f : \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times d} \times \mathbb{R}^{n\times n} \mapsto \mathbb{R}$ such that

$$f(P, U, B) = \|P - UBB^TU^T\|_F^2$$

useful gradients

$$\begin{aligned}
&\frac{\partial}{\partial U_{ij}}(-\sum_{i_3,i_4}^{d,d} U_{i,i_3}C_{i_3,i_4}U_{i,i_4}) \\
&= -\frac{\partial}{\partial U_{ij}}(\sum_{i_3\neq j,i_4}^{d,d} U_{i,i_3}C_{i_3,i_4}U_{i,i_4} + \sum_{i_4}^{d} U_{i,j}C_{j,i_4}U_{i,i_4}) \\
&= -\frac{\partial}{\partial U_{ij}}(\sum_{i_3\neq j,i_4\neq j}^{d,d} U_{i,i_3}C_{i_3,i_4}U_{i,i_4} + \sum_{i_3\neq j}^{d} U_{i,i_3}C_{i_3,j}U_{i,j} + \sum_{i_4}^{d} U_{i,j}C_{j,i_4}U_{i,i_4}) \\
&= -\frac{\partial}{\partial U_{ij}}(\sum_{i_3\neq j,i_4\neq j}^{d,d} U_{i,i_3}C_{i_3,i_4}U_{i,i_4} + \sum_{i_3\neq j}^{d} U_{i,i_3}C_{i_3,j}U_{i,j} + \sum_{i_4\neq j}^{d} U_{i,j}C_{j,i_4}U_{i,i_4} + U_{i,j}C_{j,j}U_{i,j}) \\
&= -\frac{\partial}{\partial U_{ij}}(\sum_{i_3\neq j}^{d} U_{i,i_3}C_{i_3,j}U_{i,j} + \sum_{i_4\neq j}^{d} U_{i,j}C_{j,i_4}U_{i,i_4} + U_{i,j}C_{j,j}U_{i,j}) \\
&= -(\sum_{i_3\neq j}^{d} U_{i,i_3}C_{i_3,j} + \sum_{i_4\neq j}^{d} C_{j,i_4}U_{i,i_4} + 2U_{i,j}C_{j,j}) \\
&= -(\sum_{i_3}^{d} U_{i,i_3}C_{i_3,j} + \sum_{i_4}^{d} C_{j,i_4}U_{i,i_4}) = UC_{i,j} + CU^T_{ji} = UC_{i,j} + UC^T_{i,j}
\end{aligned}$$

Letting $C = BB^T$

$$
\begin{aligned}
(\nabla_U f)_{ij} &= \frac{\partial}{\partial U_{ij}} \|P - UBB^TU^T\|_F^2 = \frac{\partial}{\partial U_{ij}} \|P - UCU^T\|_F^2 = \frac{\partial}{\partial U_{ij}} \sum_{i_1,i_2}^{n,n} (P_{i_1,i_2} - (UCU^T)_{i_1,i_2})^2 \\
&= \frac{\partial}{\partial U_{ij}} \sum_{i_1,i_2}^{n,n} (P_{i_1,i_2} - \sum_{i_3}^{d} (U_{i_1,i_3}(CU^T)_{i_3,i_2})^2 = \frac{\partial}{\partial U_{ij}} \sum_{i_1,i_2}^{n,n} (P_{i_1,i_2} - \sum_{i_3,i_4}^{d,d} U_{i_1,i_3} C_{i_3,i_4} U^T_{i_4,i_2})^2 \\
&= \frac{\partial}{\partial U_{ij}} \sum_{i_1,i_2}^{n,n} (P_{i_1,i_2} - \sum_{i_3,i_4}^{d,d} U_{i_1,i_3} C_{i_3,i_4} U_{i_2,i_4})^2 \\
&= \frac{\partial}{\partial U_{ij}} ([\sum_{\substack{i_1,i_2 \\ \{i_1,i_2\} \neq i}}^{n,n} (P_{i_1,i_2} \sum_{i_3,i_4}^{d,d} U_{i_1,i_3} C_{i_3,i_4} U_{i_2,i_4})^2] \\
&+ [\sum_{i_2}^{n} (P_{i,i_2} - \sum_{i_3,i_4}^{d,d} U_{i,i_3} C_{i_3,i_4} U_{i_2,i_4})^2)] + [\sum_{i_1}^{n} (P_{i_1,i} - \sum_{i_3,i_4}^{d,d} U_{i_1,i_3} C_{i_3,i_4} U_{i,i_4})^2]) \\
&= \frac{\partial}{\partial U_{ij}} ([\sum_{i_2}^{n} (P_{i,i_2} - \sum_{i_3,i_4}^{d,d} U_{i,i_3} C_{i_3,i_4} U_{i_2,i_4})^2)] + [\sum_{i_1}^{n} (P_{i_1,i} - \sum_{i_3,i_4}^{d,d} U_{i_1,i_3} C_{i_3,i_4} U_{i,i_4})^2])
\end{aligned}
$$

Note that each terms of the sum can be computed as

$$
\begin{aligned}
&\frac{\partial}{\partial U_{ij}} \sum_{i_2}^{n} (P_{i,i_2} - \sum_{i_3,i_4}^{d,d} U_{i,i_3} C_{i_3,i_4} U_{i_2,i_4})^2 \\
&= \frac{\partial}{\partial U_{ij}} (\sum_{i_2 \neq i}^{n} (P_{i,i_2} - \sum_{i_3,i_4}^{d,d} U_{i,i_3} C_{i_3,i_4} U_{i_2,i_4})^2 + (P_{i,i} - \sum_{i_3,i_4}^{d,d} U_{i,i_3} C_{i_3,i_4} U_{i,i_4})^2) \\
&= \sum_{i_2 \neq i}^{n} 2(P_{i,i_2} - \sum_{i_3,i_4}^{d,d} U_{i,i_3} C_{i_3,i_4} U_{i_2,i_4})(-\sum_{i_4}^{d} C_{j,i_4} U_{i_2,i_4}) \\
&+ 2(P_{i,i} - \sum_{i_3,i_4}^{d,d} U_{i,i_3} C_{i_3,i_4} U_{i,i_4})(-\sum_{i_3}^{d} U_{i,i_3} C_{i_3,j} - \sum_{i_4}^{d} C_{j,i_4} U_{i,i_4}) \\
&= \sum_{i_2 \neq i}^{n} \left[2(P_{i,i_2} - UCU^T_{i,i_2})(-CU^T_{j,i_2})\right] + 2(P_{i,i} - UCU^T_{i,i})(-UC_{i,j} - UC^T_{i,j}) \\
&= \sum_{i_2}^{n} \left[2(P_{i,i_2} - UCU^T_{i,i_2})(-UC^T_{i_2,j})\right] + 2(P_{i,i} - UCU^T_{i,i})(-UC_{i,j})
\end{aligned}
$$

$$\frac{\partial}{\partial U_{ij}} \sum_{i_1}^{n} (P_{i_1,i} - \sum_{i_3,i_4}^{d,d} U_{i_1,i_3} C_{i_3,i_4} U_{i,i_4})^2$$

$$= \sum_{i_1 \neq i}^{n} \big[\frac{\partial}{\partial U_{ij}} (P_{i_1,i} - \sum_{i_3,i_4}^{d,d} U_{i_1,i_3} C_{i_3,i_4} U_{i,i_4})^2 \big] + \frac{\partial}{\partial U_{ij}} (P_{i,i} - \sum_{i_3,i_4}^{d,d} U_{i,i_3} C_{i_3,i_4} U_{i,i_4})^2$$

$$= \sum_{i_1 \neq i}^{n} 2(P_{i_1,i} - \sum_{i_3,i_4}^{d,d} U_{i_1,i_3} C_{i_3,i_4} U_{i,i_4})(- \sum_{i_3}^{d} U_{i_1,i_3} C_{i_3,j})$$

$$+ 2(P_{i,i} - \sum_{i_3,i_4}^{d,d} U_{i,i_3} C_{i_3,i_4} U_{i,i_4})(-(\sum_{i_3}^{d} U_{i,i_3} C_{i_3,j} + \sum_{i_4}^{d} C_{j,i_4} U_{i,i_4}))$$

$$= \sum_{i_1 \neq i}^{n} [2(P_{i_1,i} - UCU^T_{i_i,i})(-UC_{i_1,j})] + 2(P_{i,i} - UCU^T_{i,i})(-UC^T_{i,j} - CU^T_{j,i})$$

$$= \sum_{i_1}^{n} [2(P_{i_1,i} - UCU^T_{i_i,i})(-UC_{i_1,j})] + 2(P_{i,i} - UCU^T_{i,i})(-CU^T_{j,i})$$

plugging these back into the original computation we get

$$= \sum_{i_1}^{n} [2(P_{i_1,i} - UCU^T_{i_i,i})(-UC_{i_1,j})] + \sum_{i_2}^{n} \big[2(P_{i,i_2} - UCU^T_{i,i_2})(-UC^T_{i_2,j}) \big]$$

$$+ 2(P_{i,i} - UCU^T_{i,i})(-(UC_{i,j} + CU^T_{j,i}))$$

$$= \sum_{i_1}^{n} [2(P_{i_1,i} - UCU^T_{i_i,i})(-UC_{i_1,j})] - 2(5(P - UCU^T)(UC^T))_{i,j}$$

$$+ 2(P_{i,i} - UCU^T_{i,i})(-(UC_{i,j} + UC^T_{i,j}))$$

$$
\begin{aligned}
&Tr(\partial U, \frac{\partial \|P - UBB^TU^T\|_F^2}{\partial U} = \lim_{\epsilon \to 0} \frac{\|P - (U + \partial U)BB^T(U + \partial U)^T\|_F^2 - \|P - UBB^TU^T\|_F^2}{\epsilon} \\
&= \frac{Tr((P - (U + \epsilon\partial U)BB^T(U + \epsilon\partial U)^T)^T(P - (U + \epsilon\partial U)BB^T(U + \epsilon\partial U)^T)) - \|P - UBB^TU^T\|_F^2}{\epsilon} \\
&= \frac{Tr((P - (U + \epsilon\partial U)BB^T(U + \epsilon\partial U)^T)^T(P - (U + \epsilon\partial U)BB^T(U + \epsilon\partial U)^T)) - \|P - UBB^TU^T\|_F^2}{\epsilon}
\end{aligned}
$$

Note that

$$
\begin{aligned}
&(P - (U + \epsilon\partial U)BB^T(U + \epsilon\partial U)^T)^T(P - (U + \epsilon\partial U)BB^T(U + \epsilon\partial U)) \\
&= (P - UBB^TU^T + \epsilon\partial UBB^TU^T + \epsilon UBB^T\partial U^T + \epsilon^2\partial UBB^T\partial U^T)^T \\
&(P - (UBB^TU^T + \epsilon\partial UBB^TU^T + \epsilon UBB^T\partial U^T + \epsilon^2\partial UBB^T\partial U^T) \\
&= P^TP -
\end{aligned}
$$