**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 $\nabla_V f = 2(P^T U - VU^T U)$.

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

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