Skip-Gram, Mathematically

Training corpus $w_1, w_2, \ldots, w_N$ (with N typically in the billions) from a fixed vocabulary V.

Goal is to maximize the average log probability:

$$\frac{1}{N} \sum_{i=1}^{N} \sum_{-L \le k \le L; k \neq 0} \log p(w_{i+k} \mid w_i)$$

Associate with each $w \in V$ an "input vector" $\mathbf{w} \in \mathbb{R}^d$ and an "output vector" $\tilde{\mathbf{w}} \in \mathbb{R}^d$. Model context probabilities as

$$p(c \mid w) = \frac{\exp(\mathbf{w} \cdot \tilde{\mathbf{c}})}{\sum_{c' \in V} \exp(\mathbf{w} \cdot \tilde{\mathbf{c}}')}.$$

The problem? V is *huge*! $\nabla \log p(c \mid w)$ takes time $O(|V|)$ to compute!