

in the next position. Every incorrect word should have low probability under the model. It can be computationally costly to enumerate all of these words. Instead, it is possible to sample only a subset of the words. Using the notation introduced in equation 12.8, the gradient can be written as follows:

$$\frac{\partial \log P(y | C)}{\partial \theta} = \frac{\partial \log \text{softmax}_y(\mathbf{a})}{\partial \theta} \quad (12.13)$$

$$= \frac{\partial}{\partial \theta} \log \frac{e^{a_y}}{\sum_i e^{a_i}} \quad (12.14)$$

$$= \frac{\partial}{\partial \theta} (a_y - \log \sum_i e^{a_i}) \quad (12.15)$$

$$= \frac{\partial a_y}{\partial \theta} - \sum_i P(y = i | C) \frac{\partial a_i}{\partial \theta} \quad (12.16)$$

where  $\mathbf{a}$  is the vector of pre-softmax activations (or scores), with one element per word. The first term is the **positive phase** term (pushing  $a_y$  up) while the second term is the **negative phase** term (pushing  $a_i$  down for all  $i$ , with weight  $P(i | C)$ ). Since the negative phase term is an expectation, we can estimate it with a Monte Carlo sample. However, that would require sampling from the model itself. Sampling from the model requires computing  $P(i | C)$  for all  $i$  in the vocabulary, which is precisely what we are trying to avoid.

Instead of sampling from the model, one can sample from another distribution, called the proposal distribution (denoted  $q$ ), and use appropriate weights to correct for the bias introduced by sampling from the wrong distribution (Bengio and S  n  cal, 2003; Bengio and S  n  cal, 2008). This is an application of a more general technique called **importance sampling**, which will be described in more detail in section 17.2. Unfortunately, even exact importance sampling is not efficient because it requires computing weights  $p_i/q_i$ , where  $p_i = P(i | C)$ , which can only be computed if all the scores  $a_i$  are computed. The solution adopted for this application is called **biased importance sampling**, where the importance weights are normalized to sum to 1. When negative word  $n_i$  is sampled, the associated gradient is weighted by

$$w_i = \frac{p_{n_i}/q_{n_i}}{\sum_{j=1}^N p_{n_j}/q_{n_j}}. \quad (12.17)$$

These weights are used to give the appropriate importance to the  $m$  negative samples from  $q$  used to form the estimated negative phase contribution to the