Noise Contrastive Estimation and Negative Sampling for Conditional Models: Consistency and Statistical Efficiency

Zhuang Ma
University of Pennsylvania[*]
zhuangma@wharton.upenn.edu

Michael Collins
Google AI Language and Columbia University[†]
mjcollins@google.com

Abstract

Noise Contrastive Estimation (NCE) is a powerful parameter estimation method for log-linear models, which avoids calculation of the partition function or its derivatives at each training step, a computationally demanding step in many cases. It is closely related to *negative sampling methods*, now widely used in NLP. This paper considers NCE-based estimation of *conditional* models. Conditional models are frequently encountered in practice; however there has not been a rigorous theoretical analysis of NCE in this setting, and we will argue there are subtle but important questions when generalizing NCE to the conditional case. In particular, we analyze two variants of NCE for conditional models: one based on a classification objective, the other based on a ranking objective. We show that the ranking-based variant of NCE gives consistent parameter estimates under weaker assumptions than the classification-based method; we analyze the statistical efficiency of the ranking-based and classification-based variants of NCE; finally we describe experiments on synthetic data and language modeling showing the effectiveness and trade-offs of both methods.

1 Introduction

This paper considers parameter estimation in conditional models of the form

$$p(y|x;\theta) = \frac{\exp\left(s(x,y;\theta)\right)}{Z(x;\theta)} \tag{1}$$

where $s(x, y; \theta)$ is the unnormalized score of label y in conjunction with input x under parameters θ, $\mathcal{Y}$ is a finite set of possible labels, and $Z(x;\theta) = \sum_{y\in\mathcal{Y}} \exp\left(s(x,y;\theta)\right)$ is the partition function for input x under parameters θ.

It is hard to overstate the importance of models of this form in NLP. In log-linear models, including both the original work on maximum-entropy models (Berger et al., 1996), and later work on conditional random fields (Lafferty et al., 2001), the scoring function $s(x, y; \theta) = \theta \cdot f(x, y)$ where $f(x, y) \in \mathbb{R}^d$ is a feature vector, and $\theta \in \mathbb{R}^d$ are the parameters of the model. In more recent work on neural networks the function $s(x, y; \theta)$ is a nonlinear function. In Word2Vec the scoring function is $s(x, y; \theta) = \theta_x \cdot \theta'_y$ where y is a word in the context of word x, and $\theta_x \in \mathbb{R}^d$ and $\theta'_y \in \mathbb{R}^d$ are "inside" and "outside" word embeddings x and y.

In many NLP applications the set $\mathcal{Y}$ is large. Maximum likelihood estimation (MLE) of the parameters θ requires calculation of $Z(x;\theta)$ or its derivatives at each training step, thereby requiring a summation over all members of $\mathcal{Y}$, which can be computationally expensive. This has led to many authors considering alternative methods, often referred to as "negative sampling methods", where a modified training objective is used that does not require summation over $\mathcal{Y}$ on each example. Instead negative examples are drawn from some distribution, and a objective function is derived based on binary classification or ranking. Prominent examples are the binary objective used in word2vec ((Mikolov et al., 2013), see also (Levy and Goldberg, 2014)), and the Noise Contrastive Estimation methods of (Mnih and Teh, 2012; Jozefowicz et al., 2016) for estimation of language models.

In spite of the centrality of negative sampling methods, they are arguably not well understood from a theoretical standpoint. There are clear connections to noise contrastive estimation (NCE) (Gutmann and Hyvärinen, 2012), a negative sampling method for parameter estimation in *joint* models of the form

$$p(y) = \frac{\exp\left(s(y;\theta)\right)}{Z(\theta)}; \quad Z(\theta) = \sum_{y\in\mathcal{Y}} \exp\left(s(y;\theta)\right) \tag{2}$$

However there has not been a rigorous theoretical analysis of NCE in the estimation of *conditional* models of the form in Eq. 1, and we will argue there are subtle but important questions when generalizing NCE to the conditional case. In particular, the joint model in Eq 2 has a single partition function $Z(\theta)$ which is estimated as a param-

[*] Part of this work done at Google.

[†] Work done at Google.

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 3698–3707
Brussels, Belgium, October 31 - November 4, 2018. ©2018 Association for Computational Linguistics

eter of the model (Gutmann and Hyvärinen, 2012) whereas the conditional model in Eq 1 has a separate partition function $Z(x;\theta)$ for each value of x. This difference is critical.

We show the following (throughout we define $K \geq 1$ to be the number of negative examples sampled per training example):

- For any $K \geq 1$, a binary classification variant of NCE, as used by (Mnih and Teh, 2012; Mikolov et al., 2013), gives consistent parameter estimates under the assumption that $Z(x;\theta)$ is constant with respect to x (i.e., $Z(x;\theta) = H(\theta)$ for some function H). Equivalently, the method is consistent under the assumption that the function $s(x, y;\theta)$ is powerful enough to incorporate $\log Z(x;\theta)$.
- For any $K \geq 1$, a ranking-based variant of NCE, as used by (Jozefowicz et al., 2016), gives consistent parameter estimates under the much weaker assumption that $Z(x;\theta)$ can vary with x. Equivalently, there is no need for $s(x, y;\theta)$ to be powerful enough to incorporate $\log Z(x;\theta)$.
- We analyze the statistical efficiency of the ranking-based and classification-based NCE variants. Under respective assumptions, both variants achieve Fisher efficiency (the same asymptotic mean square error as the MLE) as $K \rightarrow \infty$.
- We discuss application of our results to approaches of (Mnih and Teh, 2012; Mikolov et al., 2013; Levy and Goldberg, 2014; Jozefowicz et al., 2016) giving a unified account of these methods.
- We describe experiments on synthetic data and language modeling evaluating the effectiveness of the two NCE variants.

2 Basic Assumptions

We assume the following setup throughout:

- We have sets $\mathcal{X}$ and $\mathcal{Y}$, where $\mathcal{X}, \mathcal{Y}$ are finite.
- There is some unknown joint distribution $p_{X,Y}(x, y)$ where $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. We assume that the marginal distributions satisfy $p_X(x) > 0$ for all $x \in \mathcal{X}$ and $p_Y(y) > 0$ for all $y \in \mathcal{Y}$.
- We have training examples $\{x^{(i)}, y^{(i)}\}_{i=1}^n$ drawn I.I.D. from $p_{X,Y}(x, y)$.
- We have a scoring function $s(x, y;\theta)$ where θ are the parameters of the model. For example, $s(x, y;\theta)$ may be defined by a neural network.
- We use Θ to refer to the parameter space. We assume that $\Theta \subseteq \mathbb{R}^d$ for some integer d.
- We use $p_N(y)$ to refer to a distribution from which negative examples are drawn in the NCE approach. We assume that p_N satisfies $p_N(y) > 0$ for all $y \in \mathcal{Y}$.

We will consider estimation under the following two assumptions:

Assumption 2.1 *There exists some parameter value $\theta^* \in \Theta$ such that for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$,*

$$p_{Y|X}(y|x) = \frac{\exp(s(x, y;\theta^*))}{Z(x;\theta^*)} \tag{3}$$

where $Z(x;\theta^*) = \sum_{y \in \mathcal{Y}} \exp(s(x, y;\theta^*))$.

Assumption 2.2 *There exists some parameter value $\theta^* \in \Theta$, and a constant $\gamma^* \in \mathbb{R}$, such that for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$,*

$$p_{Y|X}(y|x) = \exp\left(s(x, y;\theta^*) - \gamma^*\right). \tag{4}$$

Assumption 2.2 is stronger than Assumption 2.1. It requires $\log Z(x;\theta^*) \equiv \gamma^*$ for all $x \in \mathcal{X}$, that is, the conditional distribution is perfectly self-normalized. Under Assumption 2.2, it must be the case that $\forall x \in \mathcal{X}$

$$\sum_y p_{Y|X}(y|x) = \sum_y \exp\{s(x, y;\theta^*) - \gamma^*\} = 1$$

There are $|\mathcal{X}|$ constraints but only $d + 1$ free parameters. Therefore self-normalization is a non-trivial assumption when $|\mathcal{X}| \gg d$. In the case of language modeling, $|\mathcal{X}| = |V|^k \gg d + 1$, where $|V|$ is the vocabulary size and k is the length of the context. The number of constraints grows exponentially fast.

Given a scoring function $s(x, y;\theta)$ that satisfies assumption 2.1, we can derive a scoring function s' that satisfies assumption 2.2 by defining

$$s'(x, y;\theta, \{c_x : x \in \mathcal{X}\}) = s(x, y;\theta) - c_x$$

where $c_x \in \mathbb{R}$ is a parameter for history x. Thus we introduce a new parameter c_x for each possible history x. This is the most straightforward extension of NCE to the conditional case; it is used by (Mnih and Teh, 2012). It has the clear drawback however of introducing a large number of additional parameters to the model.

3 Two Estimation Algorithms

Figure 1 shows two NCE-based parameter estimation algorithms, based respectively on *binary objective* and *ranking objective*. The input to either algorithm is a set of training examples $\{x^{(i)}, y^{(i)}\}_{i=1}^n$, a parameter K specifying the number of negative examples per training example, and a distribution $p_N(\cdot)$ from which negative examples are sampled. The algorithms differ only in the choice of objective function being optimized: L_B^n for binary objective, and L_R^n for ranking objective. Binary objective essentially corresponds to a problem where the scoring function $s(x, y; \theta)$ is used to construct a binary classifier that discriminates between positive and negative examples. Ranking objective corresponds to a problem where the scoring function $s(x, y; \theta)$ is used to rank the true label $y^{(i)}$ above negative examples $y^{(i,1)} \dots y^{(i,K)}$ for the input $x^{(i)}$.

Our main result is as follows:

Theorem 3.1 *(Informal: see section 4 for a formal statement.) For any $K \geq 1$, the binary classification-based algorithm in figure 1 is consistent under Assumption 2.2, but is not always consistent under the weaker Assumption 2.1. For any $K \geq 1$, the ranking-based algorithm in figure 1 is consistent under either Assumption 2.1 or Assumption 2.2. Both algorithms achieve the same statistical efficiency as the maximum-likelihood estimate as $K \to \infty$.*

The remainder of this section gives a sketch of the argument underlying consistency, and discusses use of the two algorithms in previous work.

3.1 A Sketch of the Consistency Argument for the Ranking-Based Algorithm

In this section, in order to develop intuition underlying the ranking algorithm, we give a proof sketch of the following theorem:

Theorem 3.2 *(First part of theorem 4.1 below.) Define $\mathcal{L}_R^\infty(\theta) = \mathbb{E}[\mathcal{L}_R^n(\theta)]$. Under Assumption 2.1, $\bar{\theta} \in \arg\max_\theta \mathcal{L}_R^\infty(\theta)$ if and only if, for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$,*

$$p_{Y|X}(y|x) = \exp(s(x, y; \bar{\theta}))/Z(x, \bar{\theta}).$$

This theorem is key to the consistency argument. Intuitively as n increases $\mathcal{L}_R^n(\theta)$ converges to $\mathcal{L}_R^\infty(\theta)$, and the output to the algorithm converges to θ' such that $p(y|x; \theta') = p_{Y|X}(y|x)$ for all x, y. Section 4 gives a formal argument.

Inputs: Training examples $\{x^{(i)}, y^{(i)}\}_{i=1}^n$, sampling distribution $p_N(\cdot)$ for generating negative examples, an integer K specifying the number of negative examples per training example, a scoring function $s(x, y; \theta)$. Flags {BINARY = true, RANKING = false} if binary classification objective is used, {BINARY = false, RANKING = true} if ranking objective is used.

Definitions: Define $\bar{s}(x, y; \theta) = s(x, y; \theta) - \log p_N(y)$
Algorithm:

- For $i = 1 \dots n$, $k = 1 \dots K$, draw $y^{(i,k)}$ I.I.D. from the distribution $p_N(y)$. For convenience define $y^{(i,0)} = y^{(i)}$.
- If RANKING, define the ranking objective function
$$\mathcal{L}_R^n(\theta) = \frac{1}{n} \sum_{i=1}^n \log \frac{\exp(\bar{s}(x^{(i)}, y^{(i,0)}; \theta))}{\sum_{k=0}^K \exp(\bar{s}(x^{(i)}, y^{(i,k)}; \theta))},$$
and the estimator $\widehat{\theta}_R = \underset{\theta \in \Theta}{\operatorname{argmax}}\, \mathcal{L}_R^n(\theta)$.
- If BINARY, define the binary objective function
$$\mathcal{L}_B^n(\theta, \gamma) = \frac{1}{n} \sum_{i=1}^n \Big\{ \log g(x^{(i)}, y^{(i,0)}; \theta, \gamma) + \sum_{k=1}^K \log \Big(1 - g(x^{(i)}, y^{(i,k)}; \theta, \gamma)\Big) \Big\},$$
and estimator $(\widehat{\theta}_B, \widehat{\gamma}_B) = \underset{\theta \in \Theta, \gamma \in \Gamma}{\operatorname{argmax}}\, \mathcal{L}_B^n(\theta, \gamma)$, where
$$g(x, y; \theta, \gamma) = \frac{\exp(\bar{s}(x, y; \theta) - \gamma)}{\exp(\bar{s}(x, y; \theta) - \gamma) + K}.$$
- Define $\widehat{\theta} = \widehat{\theta}_R$ if RANKING and $\widehat{\theta} = \widehat{\theta}_B$ otherwise. Return $\widehat{\theta}$ and
$$\widehat{p}_{Y|X}(y|x) = \frac{\exp(s(x, y; \widehat{\theta}))}{\sum_{y \in \mathcal{Y}} \exp(s(x, y; \widehat{\theta}))}$$

Figure 1: Two NCE-based estimation algorithms, using ranking objective and binary objective respectively.

We now give a proof sketch for theorem 3.2. Consider the algorithm in figure 1. For convenience define $\bar{y}^{(i)}$ to be the vector $(y^{(i,0)}, y^{(i,1)}, \dots, y^{(i,K)})$. Define $\alpha(x, \bar{y}) = \sum_{k=0}^K p_{X,Y}(x, \bar{y}_k) \prod_{j \neq k} p_N(\bar{y}_j)$, and

$$\begin{aligned} q(k|x, \bar{y}; \theta) &= \frac{\exp(\bar{s}(x, \bar{y}_k; \theta))}{\sum_{k=0}^K \exp(\bar{s}(x, \bar{y}_k; \theta))}, \\ \beta(k|x, \bar{y}) &= \frac{p_{X,Y}(x, \bar{y}_k) \prod_{j \neq k} p_N(\bar{y}_j)}{\alpha(x, \bar{y})} \\ &= \frac{p_{Y|X}(\bar{y}_k|x)/p_N(\bar{y}_k)}{\sum_{k=0}^N p_{Y|X}(\bar{y}_k|x)/p_N(\bar{y}_k)} \\ \mathrm{C}(x, \bar{y}; \theta) &= -\sum_{k=0}^K \beta(k|x, \bar{y}) \log q(k|x, \bar{y}; \theta) \end{aligned}$$

Intuitively, $q(\cdot|x, \bar{y}; \theta)$ and $\beta(\cdot|x, \bar{y})$ are posterior distributions over the true label $k \in \{0 \ldots K\}$ given an input $x, \bar{y}$, under the parameters θ and the true distributions $p_{X,\bar{Y}}(x, \bar{y})$ respectively; $\mathrm{C}(x, \bar{y}; \theta)$ is the negative cross-entropy between these two distributions.

The proof of theorem 3.2 rests on two identities. The first identity states that the objective function is the expectation of the negative cross-entropy w.r.t. the density function $\frac{1}{K+1}\alpha(x, \bar{y})$ (see Section B.1.1 of the supplementary material for derivation):

$$\mathcal{L}_R^\infty(\theta) = \sum_x \sum_{\bar{y}} \frac{1}{K+1}\alpha(x, \bar{y})\mathrm{C}(x, \bar{y}; \theta). \quad (5)$$

The second identity concerns the relationship between $q(\cdot|x, \bar{y}; \theta)$ and $\beta(\cdot|x, \bar{y})$. Under assumption 2.1, for all $x, \bar{y}$, $k \in \{0 \ldots K\}$,

$$\begin{aligned} & q(k|x, \bar{y}; \theta^*) \\ = & \frac{p_{Y|X}(\bar{y}_k|x)Z(x; \theta^*)/p_N(y_k)}{\sum_{k=0}^K p_{Y|X}(\bar{y}_k|x)Z(x; \theta^*)/p_N(y_k)} \\ = & \beta(k|x, \bar{y}) \end{aligned} \quad (6)$$

It follows immediately through the properties of negative cross entropy that

$$\forall x, \bar{y}, \quad \theta^* \in \underset{\theta}{\operatorname{argmax}}\, \mathrm{C}(x, \bar{y}; \theta) \quad (7)$$

The remainder of the argument is as follows:

- Eqs. 7 and 5 imply that $\theta^* \in \operatorname{argmax}_\theta \mathcal{L}_R^\infty(\theta)$.
- Assumption 2.1 implies that $\alpha(x, \bar{y}) > 0$ for all $x, \bar{y}$. It follows that any $\theta' \in \arg\max_\theta \mathcal{L}_R^\infty(\theta)$ satisfies

$$\begin{aligned} & \text{for all } x, \bar{y}, k, \\ & q(k|x, \bar{y}; \theta') = q(k|x, \bar{y}; \theta^*) = \beta(k|x, \bar{y}) \end{aligned} \quad (8)$$

Otherwise there would be some $x, \bar{y}$ such that $\mathrm{C}(x, \bar{y}; \theta') < \mathrm{C}(x, \bar{y}; \theta^*)$.

- Eq. 8 implies that $\forall x, y$, $p(y|x; \theta') = p(y|x; \theta^*)$. See the proof of lemma B.3 in the supplementary material.

In summary, the identity in Eq. 5 is key: the objective function in the limit, $\mathcal{L}_R^\infty(\theta)$, is related to a negative cross-entropy between the underlying distribution $\beta(\cdot|x, \bar{y})$ and a distribution under the parameters, $q(\cdot|x, \bar{y}; \theta)$. The parameters θ^* maximize this negative cross entropy over the space of all distributions $\{q(\cdot|x, \bar{y}; \theta), \theta \in \Theta\}$.

3.2 The Algorithms in Previous Work

To motivate the importance of the two algorithms, we now discuss their application in previous work.

Mnih and Teh (2012) consider language modeling, where $x = w_1 w_2 \ldots w_{n-1}$ is a history consisting of the previous $n-1$ words, and y is a word. The scoring function is defined as

$$s(x, y; \theta) = (\sum_{i=1}^{n-1} C_i r_{w_i}) \cdot q_y + b_y - c_x$$

where r_{w_i} is an embedding (vector of parameters) for history word w_i, q_y is an embedding (vector of parameters) for word y, each C_i for $i = 1 \ldots n-1$ is a matrix of parameters specifying the contribution of r_{w_i} to the history representation, b_y is a bias term for word y, and c_x is a parameter corresponding to the log normalization term for history x. Thus each history x has its own parameter c_x. The binary objective function is used in the NCE algorithm. The noise distribution $p_N(y)$ is set to be the unigram distribution over words in the vocabulary.

This method is a direct application of the original NCE method to conditional estimation, through introduction of the parameters c_x corresponding to normalization terms for each history. Interestingly, Mnih and Teh (2012) acknowledge the difficulties in maintaining a separate parameter c_x for each history, and set $c_x = 0$ for all x, noting that empirically this works well, but without giving justification.

Mikolov et al. (2013) consider an NCE-based method using the binary objective function for estimation of word embeddings. The skip-gram method described in the paper corresponds to a model where x is a word, and y is a word in the context. The vector v_x is the embedding for word x, and the vector v'_y is an embedding for word y (separate embeddings are used for x and y). The method they describe uses

$$\bar{s}(x, y; \theta) = v'_y \cdot v_x$$

or equivalently

$$s(x, y; \theta) = v'_y \cdot v_x + \log p_N(y)$$

The negative-sampling distribution $p_N(y)$ was chosen as the unigram distribution $p_Y(y)$ raised to the power $3/4$. The end goal of the method was to learn useful embeddings v_w and v'_w for each word in the vocabulary; however the method gives a consistent estimate for a model of the form

$$
\begin{aligned}
p(y|x) &= \frac{\exp\left(v'_y \cdot v_x + \log p_N(y)\right)}{\sum_y \exp\left(v'_y \cdot v_x + \log p_N(y)\right)} \\
&= \frac{p_N(y)\exp\left(v'_y \cdot v_x\right)}{Z(x;\theta)}
\end{aligned}
$$

assuming that Assumption 2.2 holds, i.e. $Z(x;\theta) = \sum_y p_N(y)\exp\left(v'_y \cdot v_x\right) \equiv H(\theta)$ which does not vary with x.

Levy and Goldberg (2014) make a connection between the NCE-based method of (Mikolov et al., 2013), and factorization of a matrix of pointwise mutual information (PMI) values of (x, y) pairs. Consistency of the NCE-based method under assumption 2.2 implies a similar result, specifically: if we define $p_N(y) = p_Y(y)$, and define $s(x, y; \theta) = v'_y \cdot v_x + \log p_N(y)$ implying $\bar{s}(x, y; \theta) = v'_y \cdot v_x$, then parameters v'_y and v_x converge to values such that

$$
p(y|x) = \frac{p_Y(y)\exp\left(v'_y \cdot v_x\right)}{H(\theta)}
$$

or equivalently

$$
\text{PMI}(x, y) = \log \frac{p(y|x)}{p(y)} = v'_y \cdot v_x - \log H(\theta)
$$

That is, following (Levy and Goldberg, 2014), the inner product $v'_y \cdot v_x$ is an estimate of the PMI up to a constant offset $H(\theta)$.

Finally, Jozefowicz et al. (2016) introduce the ranking-based variant of NCE for the language modeling problem. This is the same as the ranking-based algorithm in figure 1. They do not, however, make the connection to assumptions 2.2 and 2.1, or derive the consistency or efficiency results in the current paper. Jozefowicz et al. (2016) partially motivate the ranking-based variant throught the *importance sampling* viewpoint of Bengio and Senécal (2008). However there are two critical differences: 1) the algorithm of Bengio and Senécal (2008) does not lead to the same objective L^n_R in the ranking-based variant of NCE; instead it uses importance sampling to derive an objective that is similar but not identical; 2) the importance sampling method leads to a biased estimate of the gradients of the log-likelihood function, with the bias going to zero only as $K \to \infty$. In contrast the theorems in the current paper show that the NCE-based methods are *consistent for any value of K*. In summary, while it is tempting to view the ranking variant of NCE as an importance sampling method, the NCE-based view gives stronger guarantees for finite values of K.

4 Theory

This section states the main theorems. The supplementary material contains proofs. Throughout the paper, we use $\mathbb{E}_X[\,\cdot\,], \mathbb{E}_Y[\,\cdot\,], \mathbb{E}_{X,Y}[\,\cdot\,], \mathbb{E}_{Y|X=x}[\,\cdot\,]$ to represent the expectation w.r.t. $p_X(\cdot), p_Y(\cdot), p_{X,Y}(\cdot,\cdot), p_{Y|X}(\cdot|x)$. We use $\|\cdot\|$ to denote either the l_2 norm when the operand is a vector or the spectral norm when the operand is a matrix. Finally, we use $\Rightarrow$ to represent converge in distribution. Recall that we have defined

$$
\bar{s}(x, y; \theta) = s(x, y; \theta) - \log p_N(y).
$$

4.1 Ranking

In this section, we study noise contrastive estimation with ranking objective under Assumption 2.1. First consider the following function:

$$
\begin{aligned}
L^\infty_R(\theta) = & \sum_{x, y_0, \cdots, y_K} p_{X,Y}(x, y_0) \prod_{i=1}^{K} p_N(y_i) \\
& \times \log\left(\frac{\exp(\bar{s}(x, y_0; \theta))}{\sum_{k=0}^{K} \exp(\bar{s}(x, y_k; \theta))}\right).
\end{aligned}
$$

By straightforward calculation, one can find that

$$
L^\infty_R(\theta) = \mathbb{E}\left[L^n_R(\theta)\right].
$$

Under mild conditions, $L^n_R(\theta)$ converges to $L^\infty_R(\theta)$ as $n \to \infty$. Denote the set of maximizers of $L^\infty_R(\theta)$ by Θ^*_R, that is

$$
\Theta^*_R = \arg\max_{\theta \in \Theta} L^\infty_R(\theta).
$$

The following theorem shows that any parameter vector $\bar{\theta} \in \Theta^*_R$ if and only if it gives the correct conditional distribution $p_{Y|X}(y|x)$.

Assumption 4.1 *(Identifiability). For any $\theta \in \Theta$, if there exists a function $c(x)$ such that $s(x, y; \theta) - s(x, y; \theta^*) \equiv c(x)$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$, then $\theta = \theta^*$ and thus $c(x) = 0$ for all x.*

Theorem 4.1 *Under Assumption 2.1, $\bar{\theta} \in \Theta^*_R$ if and only if, for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$,*

$$
p_{Y|X}(y|x) = \exp(s(x, y; \bar{\theta}))/Z(x, \bar{\theta}).
$$

In addition, Θ^*_R is a singleton if and only if Assumption 4.1 holds.

Next we consider consistency of the estimation algorithm based on the ranking objective under the following regularity assumptions:

Assumption 4.2 *(Continuity). $s(x, y; \theta)$ is continuous w.r.t. θ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$.*

Assumption 4.3 *Θ_R^* is contained in the interior of a compact set $\Theta \subset \mathbb{R}^d$.*

For a given estimate $\widehat{p}_{Y|X}$ of the conditional distribution $p_{Y|X}$, define the error metric $d(\cdot, \cdot)$ by

$$d\left(\widehat{p}_{Y|X}, p_{Y|X}\right) = \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p_{X,Y}(x, y) \times \left(\widehat{p}_{Y|X}(y|x) - p_{Y|X}(y|x)\right)^2.$$

For a sequence of IID observations $(x^{(1)}, y^{(1)})$, $(x^{(2)}, y^{(2)}), \ldots,$ define the sequences of estimates $(\widehat{\theta}_R^1,\ \widehat{p}_{Y|X}^1)$, $(\widehat{\theta}_R^2,\ \widehat{p}_{Y|X}^2), \ldots$ where the n^{th} estimate $(\widehat{\theta}_R^n,\ \widehat{p}_{Y|X}^n)$ is obtained by optimizing the ranking objective of figure 1 on $(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(n)}, y^{(n)})$.

Theorem 4.2 *(Consistency) Under Assumptions 2.1, 4.2, 4.3, the estimates based on the ranking objective are strongly consistent in the sense that for any fixed $K \geq 1$,*

$$\mathbb{P}\Big\{\lim_{n\to\infty} \min_{\theta^* \in \Theta_R^*} \|\widehat{\theta}_R^n - \theta^*\| = 0\Big\} = \mathbb{P}\Big\{\lim_{n\to\infty} d\left(\widehat{p}_{Y|X}^n,\ p_{Y|X}\right) = 0\Big\} = 1$$

Further, if Assumption 4.1 holds,

$$\mathbb{P}\left\{\lim_{n\to\infty} \widehat{\theta}_R^n = \theta^*\right\} = 1.$$

Remark 4.1 *Thoughout the paper, all NCE estimators are defined for some fixed K. We suppress the dependence on K to simplify notation (e.g. $\widehat{\theta}_R^n$ should be interpreted as $\widehat{\theta}_R^{n,K}$).*

4.2 Classification

Now we turn to the analysis of NCE with binary objective under Assumption 2.2. First consider the following function,

$$L_B^\infty(\theta, \gamma) = \sum_{x,y} \Big\{p_{X,Y}(x, y) \log\left(g(x, y; \theta, \gamma)\right) + K p_X(x) p_N(y) \log\left(1 - g(x, y; \theta, \gamma)\right)\Big\}$$

One can find that

$$L_B^\infty(\theta, \gamma) = \mathbb{E}\left[L_B^n(\theta, \gamma)\right].$$

Denote the set of maximizers of $L_B^\infty(\theta, \gamma)$ by Ω_B^* :

$$\Omega_B^* = \arg \max_{\theta\in\Theta, \gamma\in\Gamma} L_B^\infty(\theta, \gamma).$$

Parallel results of Theorem 4.1, 4.2 are established as follows.

Assumption 4.4 *(Identifiability). For any $\theta \in \Theta$, if there exists some constant c such that $s(x, y; \theta) - s(x, y; \theta^*) \equiv c$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$, then $\theta = \theta^*$ and thus $c = 0$.*

Assumption 4.5 *Ω_B^* is in the interior of $\Theta \times \Gamma$ where $\Theta \subset \mathbb{R}^d, \Gamma \subset \mathbb{R}$ are compact sets.*

Theorem 4.3 *Under Assumption 2.2, $(\bar{\theta}, \bar{\gamma}) \in \Omega_B^*$ if and only if, for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$,*

$$p_{Y|X}(y|x) = \exp(s(x, y; \bar{\theta}) - \bar{\gamma})$$

for all (x, y). Ω_B^* is a singleton if and only if Assumption 4.4 holds.

Similarly we can define the sequence of estimates $(\widehat{\theta}_B^1,\ \widehat{\gamma}_B^1,\ \widehat{p}_{Y|X}^1)$, $(\widehat{\theta}_B^2,\ \widehat{\gamma}_B^2,\ \widehat{p}_{Y|X}^2), \ldots$ based on the binary objective.

Theorem 4.4 *(Consistency) Under Assumption 2.2, 4.2, 4.5, the estimates defined by the binary objective are strongly consistent in the sense that for any $K \geq 1$,*

$$\mathbb{P}\Big\{\lim_{n\to\infty} \min_{(\theta^*, \gamma^*)\in\Omega_B^*} \|(\widehat{\theta}_B^n, \widehat{\gamma}_B^n) - (\theta^*, \gamma^*)\| = 0\Big\} = \mathbb{P}\Big\{\lim_{n\to\infty} d\left(\widehat{p}_{Y|X}^n,\ p_{Y|X}\right) = 0\Big\} = 1$$

If further Assumption 4.4 holds,

$$\mathbb{P}\left\{\lim_{n\to\infty} (\widehat{\theta}_B^n, \widehat{\gamma}_B^n) = (\theta^*, \gamma^*)\right\} = 1.$$

4.3 Counterexample

In this section, we give a simple example to demonstrate that the binary classification approach fails to be consistent when assumption 2.1 holds but assumption 2.2 fails (i.e. the partition function depends on the input).

Consider $X \in \mathcal{X} = \{x_1, x_2\}$ with marginal distribution

$$p_X(x_1) = p_X(x_2) = 1/2,$$

and $Y \in \mathcal{Y} = \{y_1, y_2\}$ generated by the conditional model specified in assumption 2.1 with the score function parametrized by $\theta = (\theta_1, \theta_2)$ and

$$s(x_1, y_1; \theta) = \log\theta_1,$$

$s(x_1, y_2; \theta) = s(x_2, y_1; \theta) = s(x_2, y_2; \theta) = \log \theta_2$.

Assume the true parameter is $\theta^* = (\theta_1^*, \theta_2^*) = (1, 3)$. By simple calculation,

$$Z(\theta^*; x_1) = 4, \; Z(\theta^*; x_2) = 6,$$

$$p_{X,Y}(x_1, y_1) = 1/8, p_{X,Y}(x_1, y_2) = 3/8,$$
$$p_{X,Y}(x_2, y_1) = p_{X,Y}(x_2, y_2) = 1/4.$$

Suppose we choose the negative sampling distribution $p_N(y_1) = p_N(y_2) = 1/2$. For any $K \geq 1$, by the Law of Large Numbers, as n goes to infinity, $L_B^n(\theta, \gamma)$ will converge to $L_B^\infty(\theta, \gamma)$. Substitute in the parameters above. One can show that

$$\begin{aligned} L_B^\infty(\theta, \gamma) = & \frac{1}{8} \log \frac{2\theta_1}{2\theta_1 + K \exp(\gamma)} \\ & + \frac{K}{4} \log \frac{K \exp(\gamma)}{2\theta_1 + K \exp(\gamma)} \\ & + \frac{7}{8} \log \frac{2\theta_2}{2\theta_2 + K \exp(\gamma)} \\ & + \frac{3K}{4} \log \frac{K \exp(\gamma)}{2\theta_2 + K \exp(\gamma)}. \end{aligned}$$

Setting the derivatives w.r.t. θ_1, θ_2 to zero, one will obtain

$$\theta_1 = \frac{1}{4} \exp(\gamma), \;\; \theta_2 = \frac{7}{12} \exp(\gamma).$$

So for any $(\widetilde{\theta}_1, \widetilde{\theta}_2, \widetilde{\gamma}) \in \mathrm{argmax}_{\theta, \gamma} L_B^\infty(\theta, \gamma)$, $(\widetilde{\theta}_1, \widetilde{\theta}_2, \widetilde{\gamma})$ will satisfy the equalities above. Then the estimated distribution $\widetilde{p}_{Y|X}$ will satisfy

$$\frac{\widetilde{p}_{Y|X}(y_1|x_1)}{\widetilde{p}_{Y|X}(y_2|x_1)} = \frac{\widetilde{\theta}_1}{\widetilde{\theta}_2} = \frac{1/4}{7/12} = \frac{3}{7},$$

which contradicts the fact that

$$\frac{p_{Y|X}(y_1|x_1)}{p_{Y|X}(y_2|x_1)} = \frac{p_{X,Y}(x_1, y_1)}{p_{X,Y}(x_1, y_2)} = \frac{1}{3}.$$

So the binary objective does not give consistent estimation of the conditional distribution.

4.4 Asymptotic Normality and Statistical Efficiency

Noise Contrastive Estimation significantly reduces the computational complexity, especially when the label space $|\mathcal{Y}|$ is large. It is natural to ask: does such scalability come at a cost? Classical likelihood theory tells us, under mild conditions, the maximum likelihood estimator (MLE) has nice properties like asymptotic normality and Fisher efficiency. More specifically, as the sample size goes to infinity, the distribution of the MLE will converge to a multivariate normal distribution, and the mean square error of the MLE will achieve the Cramer-Rao lower bound (Ferguson, 1996).

We have shown the consistency of the NCE estimators in Theorem 4.2 and Theorem 4.4. In this part of the paper, we derive their asymptotic distribution and quantify their statistical efficiency. To this end, we restrict ourselves to the case where θ^* is identifiable (i.e. Assumptions 4.1 or 4.4 hold) and the scoring function $s(x, y; \theta)$ satisfies the following smoothness condition:

Assumption 4.6 *(Smoothness). The scoring function $s(x, y; \theta)$ is twice continuous differentiable w.r.t. θ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$.*

We first introduce the following maximum-likelihood estimator.

$$\begin{aligned} \widehat{\theta}^{\mathrm{MLE}} &= \arg\min_{\theta} \; L_{\mathrm{MLE}}^n(\theta) \\ &:= \arg\min_{\theta} \sum_{i=1}^n \log \left(\frac{\exp(s(x^{(i)}, y^{(i)}; \theta))}{\sum_{y \in \mathcal{Y}} \exp(s(x^{(i)}, y; \theta))} \right). \end{aligned}$$

Define the matrix

$$\mathcal{I}_{\theta^*} = \mathbb{E}_X \left[\mathrm{Var}_{Y|X=x} \left[\nabla_\theta s(x, y; \theta^*) \right] \right].$$

As shown below, $\mathcal{I}_{\theta^*}$ is essentially the Fisher information matrix under the conditional model.

Theorem 4.5 *Under Assumption 2.1, 4.1, 4.3, and 4.6, if $\mathcal{I}_{\theta^*}$ is non-singular, as $n \to \infty$*

$$\sqrt{n}(\widehat{\theta}^{MLE} - \theta^*) \;\Rightarrow\; \mathcal{N}(0, \mathcal{I}_{\theta^*}^{-1}).$$

For any given estimator $\widehat{\theta}$, define the scaled asymptotic mean square error by

$$\mathrm{MSE}_\infty(\widehat{\theta}) = \lim_{n \to \infty} \mathbb{E} \left[\left\| \sqrt{\frac{n}{d}} \left(\widehat{\theta} - \theta^* \right) \right\|^2 \right],$$

where d is the dimension of the parameter θ^*. Theorem 4.5 implies that,

$$\mathrm{MSE}_\infty(\widehat{\theta}^{\mathrm{MLE}}) = \mathrm{Tr}(\mathcal{I}_{\theta^*}^{-1})/d.$$

where $\mathrm{Tr}(\cdot)$ denotes the trace of a matrix. According to classical MLE theory (Ferguson, 1996), under certain regularity conditions, this is the best achievable mean square error. So the next question to answer is: can these NCE estimators approach this limit?

Assumption 4.7 *There exist positive constants c, C such that $\sigma_{\min}(\mathcal{I}_{\theta^*}) \geq c$ and*

$$\max_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \Big\{ |s(x,y;\theta^*)|, \|\nabla_\theta s(x,y;\theta^*)\|, \\ \left\|\nabla^2_\theta s(x,y;\theta^*)\right\| \Big\} \leq C.$$

where $\sigma_{\min}(\cdot)$ denotes the smallest singular value.

Theorem 4.6 (Ranking) *Under Assumption 2.1, 4.1, 4.3, 4.6, 4.7, there exists an integer K_0 such that for all $K \geq K_0$, as $n \to \infty$*

$$\sqrt{n}\left(\widehat{\theta}_R - \theta^*\right) \Rightarrow \mathcal{N}(0, \mathcal{I}^{-1}_{R,K}), \qquad (9)$$

for some matrix $\mathcal{I}_{R,K}$. There exists a constant C such that for all $K \geq K_0$,

$$|\operatorname{MSE}_\infty(\widehat{\theta}_R) - \operatorname{MSE}_\infty(\widehat{\theta}^{MLE})| \leq C/\sqrt{K} \\ \|\mathcal{I}^{-1}_{R,K} - \mathcal{I}^{-1}_{\theta^*}\| \leq C/\sqrt{K}$$

Theorem 4.7 (Binary) *Under Assumption 2.2, 4.4, 4.5, 4.6, 4.7, there exists an integer K_0 such that, for any $K \geq K_0$, as $n \to \infty$*

$$\sqrt{n}\left(\widehat{\theta}_B - \theta^*\right) \Rightarrow \mathcal{N}(0, \mathcal{I}^{-1}_{B,K}), \qquad (10)$$

for some matrix $\mathcal{I}_{B,K}$. There exists a constant C such that for all $K \geq K_0$,

$$|\operatorname{MSE}_\infty(\widehat{\theta}_B) - \operatorname{MSE}_\infty(\widehat{\theta}^{MLE})| \leq C/K \\ \|\mathcal{I}^{-1}_{B,K} - \mathcal{I}^{-1}_{\theta^*}\| \leq C/K.$$

Remark 4.2 *Theorem 4.6 and 4.7 reveal that under respective model assumptions, for any given $K \geq K_0$ both NCE estimators are asymptotically normal and $\sqrt{n}$-consistent. Moreover, both NCE estimators approach Fisher efficiency (statistical optimality) as K grows.*

5 Experiments

5.1 Simulations

Suppose we have a feature space $\mathcal{X} \subset \mathbb{R}^d$ with $|\mathcal{X}| = m_x$, label space $\mathcal{Y} = \{1, \cdots, m_y\}$, and parameter $\theta = (\theta_1, \cdots, \theta_{m_y}) \in \mathbb{R}^{m_y \times d}$. Then for any given sample size n, we can generate observations $(x^{(i)}, y^{(i)})$ by first sampling $x^{(i)}$ uniformly from $\mathcal{X}$ and then sampling $y^{(i)} \in \mathcal{Y}$ by the condional model

$$p(y|x;\theta) = \exp(x'\theta_y) / \sum_{y=1}^{m_y} \exp(x'\theta_y).$$

We first consider the estimation of θ by MLE and NCE-ranking. We fix $d = 4, m_x = 200, m_y = 100$ and generate $\mathcal{X}$ and the parameter θ from separate mixtures of Gaussians. We try different configurations of (n, K) and report the KL divergence between the estimated distribution and true distribution, as summarized in the left panel of figure 2. The observations are:

- The NCE estimators are consistent for any fixed K. For a fixed sample size, the NCE estimators become comparable to MLE as K increases.
- The larger the sample size, the less sensitive are the NCE estimators to K. A very small value of K seems to suffice for large sample size.

Apparently, under the parametrization above, the model is not self-normalized. To use NCE-binary, we add an extra x-dependent bias parameter b_x to the score function (i.e. $s(x,y;\theta) = x'\theta_y + b_x$) to make the model self-normalized or else the algorithm will not be consistent. Similar patterns to figure 2 are observed when varying sample size and K (see Section A.1 of the supplementary material). However this makes NCE-binary not directly comparable to NCE-ranking/MLE since its performance will be compromised by estimating extra parameters and the number of extra parameters depends on the richness of the feature space $\mathcal{X}$. To make this clear, we fix $n = 16000, d = 4, m_y = 100, K = 32$ and experiment with $m_x = 100, 200, 300, 400$. The results are summarized on the right panel of figure 2. As $|\mathcal{X}|$ increases, the KL divergence will grow while the performance of NCE-ranking/MLE is independent of $|\mathcal{X}|$. Without the x-dependent bias term for NCE-binary, the KL divergence will be much higher due to lack of consistency (0.19, 0.21, 0.24, 0.26 respectively).

5.2 Language Modeling

We evaluate the performance of the two NCE algorithms on a language modeling problem, using the Penn Treebank (PTB) dataset (Marcus et al., 1993). We choose (Zaremba et al., 2014) as the benchmark where the conditional distribution is modeled by two-layer LSTMs and the parameters are estimated by MLE (note that the current state-of-the-art is (Yang et al., 2018)). Zaremba et al. (2014) implemented 3 model configurations: “Small” , “Medium” and “Large”, which have 200, 650 and 1500 units per layer respectively. We follow their setup (model size, unrolled steps, dropout ratio, etc) but train the model by maximiz-

	Small		Medium		Large	
MLE	111.5		82.7		78.4	
NCE	Ranking	Binary	Ranking	Binary	Ranking	Binary
$K = 200$	113.8	106.8	83.2	82.1	79.3	76.0
$K = 400$	112.9	105.6	82.3	81.5	77.9	75.6
$K = 800$	111.9	105.3	81.4	81.6	77.8	75.7
$K = 1600$	110.6	**104.8**	81.7	81.5	77.5	75.9
reg-MLE	105.4		79.9		77.0	
reg-Ranking ($K = 1600$)	105.4		**79.8**		**75.0**	
reg-Binary ($K = 1600$)	**104.8**		82.5		75.7	

Table 1: Perplexity on the test set of Penn Treebank. We show performance for the ranking v.s. binary loss algorithms, with different values for K, and with/without regularization.

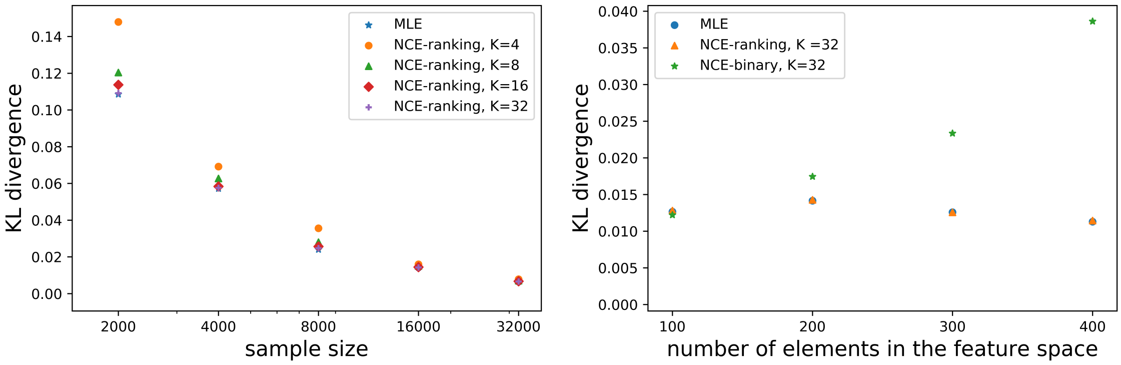

Figure 2: KL divergence between the true distribution and the estimated distribution.

ing the two NCE objectives. We use the unigram distribution as the negative sampling distribution and consider $K = 200, 400, 800, 1600$.

The results on the test set are summarized in table 1. Similar patterns are observed on the validation set (see Section A.2 of the supplementary material). As shown in the table, the performance of NCE-ranking and NCE-binary improves as the number of negative examples increases, and finally outperforms the MLE.

An interesting observation is, without regularization, the binary classification approach outperforms both ranking and MLE. This suggests the model space (two-layer LSTMs) is rich enough as to approximately incorporate the x-dependent partition function $Z(\theta; x)$, thus making the model approximately self-normalized. This motivates us to modify the ranking and MLE objectives by adding the following regularization term:

$$\frac{\alpha}{n}\sum_{i=1}^{n}\left(\log\left(\frac{1}{m}\sum_{j=1}^{m}\exp\left(\bar{s}(x^{(i)},\widetilde{y}^{(i,j)};\theta)\right)\right)\right)^{2} \approx \alpha\,\mathbb{E}_X\left[(\log Z(x;\theta))^2\right],$$

where $\widetilde{y}^{(i,j)}, 1 \leq j \leq m$ are sampled from the noise distribution $p_N(\cdot)$. This regularization term promotes a constant partition function, that is $Z(x;\theta) \approx 1$ for all $x \in \mathcal{X}$. In our experiments, we fix m to be 1/10 of the vocabulary size, $K = 1600$ and tune the regularization parameter α. As shown in the last three rows of the table, regularization significantly improves the performance of both the ranking approach and the MLE.

6 Conclusions

In this paper we have analyzed binary and ranking variants of NCE for estimation of conditional models $p(y|x;\theta)$. The ranking-based variant is consistent for a broader class of models than the binary-based algorithm. Both algorithms achieve Fisher efficiency as the number of negative examples increases. Experiments show that both algorithms outperform MLE on a language modeling task. The ranking-based variant of NCE outperforms the binary-based variant once a regularizer is introduced that encourages self-normalization.

Acknowledgments

The authors thank Emily Pitler and Ali Elkahky for many useful conversations about the work, and David Weiss for comments on an earlier draft of the paper.

References

Yoshua Bengio and Jean-Sébastien Senécal. 2008. Adaptive importance sampling to accelerate training of a neural probabilistic language model. *IEEE Transactions on Neural Networks*, 19(4):713–722.

Adam L. Berger, Vincent J. Della Pietra, and Stephen A. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Comput. Linguist.*, 22(1):39–71.

Thomas Shelburne Ferguson. 1996. *A course in large sample theory*, volume 49. Chapman & Hall London.

Michael U Gutmann and Aapo Hyvärinen. 2012. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research*, 13(Feb):307–361.

Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. 2016. Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 2177–2185, Cambridge, MA, USA. MIT Press.

Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119, USA. Curran Associates Inc.

Andriy Mnih and Yee W Teh. 2012. A fast and simple algorithm for training neural probabilistic language models. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 1751–1758.

Zhilin Yang, Zihang Dai, Ruslan Salakhutdinov, and William W. Cohen. 2018. Breaking the softmax bottleneck: A high-rank RNN language model. In *International Conference on Learning Representations*.

Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. 2014. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*.