Q. 1 Hyperparameters explored

1. Learning Rate
    a. *learning_rate* controls the size of jumps that we take towards final optimized point. With smaller learning rate, what I found that I needed more *max_num_steps* to achieve the same accuracy on the *dev* set.
    b. Eg. for a smaller learning rate (0.01), I found that the accuracy was generally lower for the same *num_steps* however it increased a little bit and then flattened out after increasing the *num_steps* proportionally.
2. Optimization algorithm – the algorithm that facilitates the gradient updates in each round.
    a. GradientDescentOptimizer – this is the only optimizer that worked on my hardware configuration.
    b. AdamOptimizer – got out-of-memory error did not work
    c. SGDOptimizer – got out-of-memory error did not work
3. *num_skips* and *skip_window* how many words to consider left right and how many times to reuse input to generate label.
    a. doubled – (tried values *skip_window = 8* and *num_skips = 16*)
        i. For this configuration, I found out that for this configuration the Overall accuracy decreased a little bit on the dev set for both the loss functions.
    b. halved – (tried values *skip_window = 4* and *num_skips = 8*)
        i. For this configuration, I found out that for this configuration the Overall accuracy increased a little bit on the dev set for both the loss functions.
4. Without the pretrained model, the performance in terms of accuracy of both the loss functions degraded. (for the default *num_steps*) [Overall accuracy for NCE = 31.8% and for CE = 30.9%]
5. *batch_size* the chuck of data that we process each time
    a. No significant increase in the Overall accuracy was observed after increasing the batch size. On the contrary for higher batch sizes, (512 and above) I observed a drop in the accuracy.

Q. 2 Results on analogy tasks on 5 different configs.

Baseline Performance CE	
Generated by:	score_maxdiff.pl
Mechanical Turk File:	word_analogy_dev_mturk_answers.txt
Test File:	word_analogy_dev_output_cross_entropy.txt
Number of MaxDiff Questions:	914
Number of Least Illustrative Guessed Correctly:	297
Number of Least Illustrative Guessed Incorrectly:	617
Accuracy of Least Illustrative Guesses:	32.50%
Number of Most Illustrative Guessed Correctly:	322
Number of Most Illustrative Guessed Incorrectly:	592
Accuracy of Most Illustrative Guesses:	35.20%
Overall Accuracy:	33.90%

Base Line Performance NCE	
Generated by:	score_maxdiff.pl
Mechanical Turk File:	word_analogy_dev_mturk_answers.txt
Test File:	word_analogy_dev_output_NCE.txt
Number of MaxDiff Questions:	914
Number of Least Illustrative Guessed Correctly:	303
Number of Least Illustrative Guessed Incorrectly:	611
Accuracy of Least Illustrative Guesses:	33.20%
Number of Most Illustrative Guessed Correctly:	320
Number of Most Illustrative Guessed Incorrectly:	594
Accuracy of Most Illustrative Guesses:	35.00%
Overall Accuracy:	34.10%
Skip window = 8 and num_skip = 16 (BEST CE)	
Generated by:	score_maxdiff.pl
Mechanical Turk File:	word_analogy_dev_mturk_answers.txt
Test File:	word_analogy_dev_output_cross_entropy
Number of MaxDiff Questions:	914
Number of Least Illustrative Guessed Correctly:	296
Number of Least Illustrative Guessed Incorrectly:	618
Accuracy of Least Illustrative Guesses:	32.40%
Number of Most Illustrative Guessed Correctly:	317
Number of Most Illustrative Guessed Incorrectly:	597
Accuracy of Most Illustrative Guesses:	34.70%
Overall Accuracy:	33.50%

Skip window = 8 and num_skip = 16 (BEST NCE)	
Generated by:	score_maxdiff.pl
Mechanical Turk File:	word_analogy_dev_mturk_answers.txt
Test File:	word_analogy_dev_output_nce
Number of MaxDiff Questions:	914
Number of Least Illustrative Guessed Correctly:	298
Number of Least Illustrative Guessed Incorrectly:	616
Accuracy of Least Illustrative Guesses:	32.60%
Number of Most Illustrative Guessed Correctly:	313
Number of Most Illustrative Guessed Incorrectly:	601
Accuracy of Most Illustrative Guesses:	34.20%
Overall Accuracy:	33.40%

Skip window = 2 and num_skip = 4 (BEST CE)	
Generated by:	score_maxdiff.pl
Mechanical Turk File:	word_analogy_dev_mturk_answers.txt
Test File:	word_analogy_dev_output_cross_entropy.txt
Number of MaxDiff Questions:	914
Number of Least Illustrative Guessed Correctly:	307
Number of Least Illustrative Guessed Incorrectly:	607
Accuracy of Least Illustrative Guesses:	33.60%
Number of Most Illustrative Guessed Correctly:	325
Number of Most Illustrative Guessed Incorrectly:	589
Accuracy of Most Illustrative Guesses:	35.60%
Overall Accuracy:	34.60%

Skip window = 2 and num_skip = 4 (BEST NCE)	
Generated by:	score_maxdiff.pl
Mechanical Turk File:	word_analogy_dev_mturk_answers.txt
Test File:	word_analogy_dev_output_nce.txt
Number of MaxDiff Questions:	914
Number of Least Illustrative Guessed Correctly:	310
Number of Least Illustrative Guessed Incorrectly:	604
Accuracy of Least Illustrative Guesses:	33.90%
Number of Most Illustrative Guessed Correctly:	331
Number of Most Illustrative Guessed Incorrectly:	583
Accuracy of Most Illustrative Guesses:	36.20%
Overall Accuracy:	35.10%

Learning Rate - 1.5 (CE)	
Generated by:	score_maxdiff.pl
Mechanical Turk File:	word_analogy_dev_mturk_answers.txt
Test File:	word_analogy_dev_output_ce.txt
Number of MaxDiff Questions:	914
Number of Least Illustrative Guessed Correctly:	291
Number of Least Illustrative Guessed Incorrectly:	623
Accuracy of Least Illustrative Guesses:	31.80%
Number of Most Illustrative Guessed Correctly:	317
Number of Most Illustrative Guessed Incorrectly:	597
Accuracy of Most Illustrative Guesses:	34.70%
Overall Accuracy:	33.30%

Learning Rate - 1.5 (NCE)	
Generated by:	score_maxdiff.pl
Mechanical Turk File:	word_analogy_dev_mturk_answers.txt
Test File:	word_analogy_test_output_nce
Number of MaxDiff Questions:	914
Number of Least Illustrative Guessed Correctly:	307
Number of Least Illustrative Guessed Incorrectly:	607
Accuracy of Least Illustrative Guesses:	33.60%
Number of Most Illustrative Guessed Correctly:	314
Number of Most Illustrative Guessed Incorrectly:	600
Accuracy of Most Illustrative Guesses:	34.40%
Overall Accuracy:	34.00%

Learning Rate - 0.5 (CE)	
Generated by:	score_maxdiff.pl
Mechanical Turk File:	word_analogy_dev_mturk_answers.txt
Test File:	word_analogy_test_output_cross_entropy
Number of MaxDiff Questions:	914
Number of Least Illustrative Guessed Correctly:	297
Number of Least Illustrative Guessed Incorrectly:	617
Accuracy of Least Illustrative Guesses:	32.50%
Number of Most Illustrative Guessed Correctly:	322
Number of Most Illustrative Guessed Incorrectly:	592
Accuracy of Most Illustrative Guesses:	35.20%
Overall Accuracy:	33.90%

Learning Rate - 0.5 (NCE)	
Generated by:	score_maxdiff.pl
Mechanical Turk File:	word_analogy_dev_mturk_answers.txt
Test File:	word_analogy_test_output_nce
Number of MaxDiff Questions:	914
Number of Least Illustrative Guessed Correctly:	307
Number of Least Illustrative Guessed Incorrectly:	607
Accuracy of Least Illustrative Guesses:	33.60%
Number of Most Illustrative Guessed Correctly:	314
Number of Most Illustrative Guessed Incorrectly:	600
Accuracy of Most Illustrative Guesses:	34.40%
Overall Accuracy:	34.00%

Learning Rate 0.01 (CE)	
Generated by:	score_maxdiff.pl
Mechanical Turk File:	word_analogy_dev_mturk_answers.txt
Test File:	word_analogy_dev_output_cross_entropy.txt
Number of MaxDiff Questions:	914
Number of Least Illustrative Guessed Correctly:	297
Number of Least Illustrative Guessed Incorrectly:	617
Accuracy of Least Illustrative Guesses:	32.50%
Number of Most Illustrative Guessed Correctly:	317
Number of Most Illustrative Guessed Incorrectly:	597
Accuracy of Most Illustrative Guesses:	34.70%
Overall Accuracy:	33.60%

Learning Rate 0.01 (NCE)	
Generated by:	score_maxdiff.pl
Mechanical Turk File:	word_analogy_dev_mturk_answers.txt
Test File:	word_analogy_dev_output_nce
Number of MaxDiff Questions:	914
Number of Least Illustrative Guessed Correctly:	305
Number of Least Illustrative Guessed Incorrectly:	609
Accuracy of Least Illustrative Guesses:	33.40%
Number of Most Illustrative Guessed Correctly:	315
Number of Most Illustrative Guessed Incorrectly:	599
Accuracy of Most Illustrative Guesses:	34.50%
Overall Accuracy:	33.90%

batch size = 64 (NCE)		batch size = 64 (CE)	
Generated by:	score_maxdiff.pl	Generated by:	score_maxdiff.pl
Mechanical Turk File:	word_analogy_dev_mturk_answers.txt	Mechanical Turk File:	word_analogy_dev_mturk_answers.txt
Test File:	word_analogy_dev_output_ncebatch64.txt	Test File:	word_analogy_dev_output_cross_entropybatch64.txt
Number of MaxDiff Questions:	914	Number of MaxDiff Questions:	914
Number of Least Illustrative Guessed Correctly:	301	Number of Least Illustrative Guessed Correctly:	302
Number of Least Illustrative Guessed Incorrectly:	613	Number of Least Illustrative Guessed Incorrectly:	612
Accuracy of Least Illustrative Guesses:	32.90%	Accuracy of Least Illustrative Guesses:	33.00%
Number of Most Illustrative Guessed Correctly:	324	Number of Most Illustrative Guessed Correctly:	320
Number of Most Illustrative Guessed Incorrectly:	590	Number of Most Illustrative Guessed Incorrectly:	594
Accuracy of Most Illustrative Guesses:	35.40%	Accuracy of Most Illustrative Guesses:	35.00%
Overall Accuracy:	34.20%	Overall Accuracy:	34.00%

Batch Size 512 (NCE)		Batch Size 512 (CE)	
Generated by:	score_maxdiff.pl	Generated by:	score_maxdiff.pl
Mechanical Turk File:	word_analogy_dev_mturk_answers.txt	Mechanical Turk File:	word_analogy_dev_mturk_answers.txt
Test File:	word_analogy_dev_output_nce.txt	Test File:	word_analogy_dev_output_cross_entropy.txt
Number of MaxDiff Questions:	914	Number of MaxDiff Questions:	914
Number of Least Illustrative Guessed Correctly:	301	Number of Least Illustrative Guessed Correctly:	291
Number of Least Illustrative Guessed Incorrectly:	613	Number of Least Illustrative Guessed Incorrectly:	623
Accuracy of Least Illustrative Guesses:	32.90%	Accuracy of Least Illustrative Guesses:	31.80%
Number of Most Illustrative Guessed Correctly:	324	Number of Most Illustrative Guessed Correctly:	316
Number of Most Illustrative Guessed Incorrectly:	590	Number of Most Illustrative Guessed Incorrectly:	598
Accuracy of Most Illustrative Guesses:	35.40%	Accuracy of Most Illustrative Guesses:	34.60%
Overall Accuracy:	34.20%	Overall Accuracy:	33.20%

Q. 3 Top 20 similar words

NCE Loss Function

American		First		Would	
1	british	1	term	1	could
2	german	2	early	2	will
3	french	3	including	3	said
4	italian	4	english	4	so
5	russian	5	abuse	5	must
6	destroy	6	against	6	did
7	leftists	7	class	7	we
8	belief	8	use	8	if
9	detailed	9	working	9	do
10	international	10	word	10	does
11	spirit	11	anarchism	11	india
12	ruler	12	ruler	12	believed
13	advocate	13	used	13	can
14	heterodox	14	although	14	you
15	autres	15	about	15	had
16	united	16	work	16	even
17	follow	17	anarchists	17	should
18	borges	18	anarchy	18	claimed
19	proudhon	19	most	19	only
20	william	20	unnecessary	20	never

Cross Entropy Loss Function

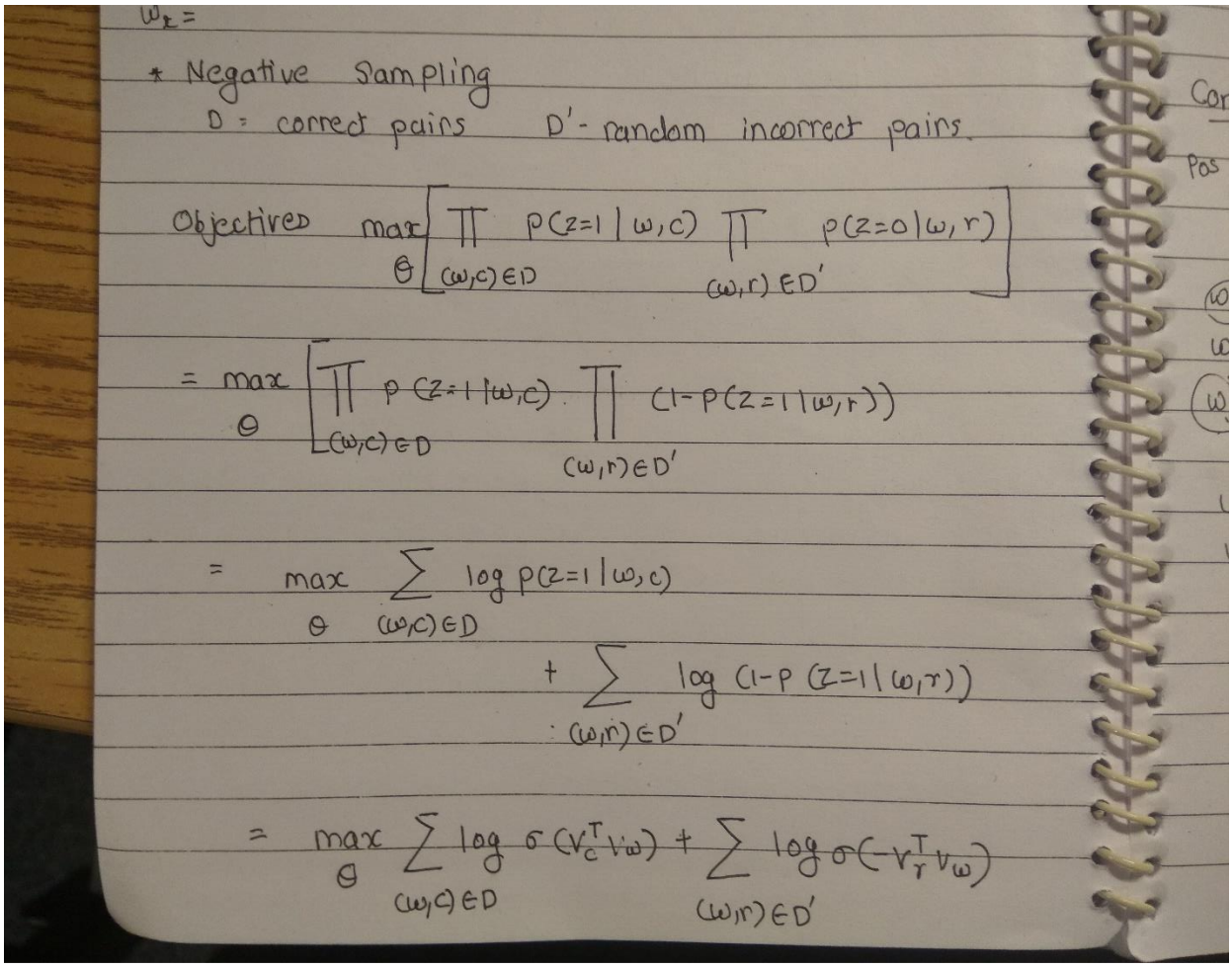
American		Would		First	
1	german	1	could	1	last
2	british	2	will	2	most
3	french	3	must	3	same
4	english	4	did	4	main
5	italian	5	can	5	largest
6	its	6	should	6	original
7	war	7	does	7	best
8	russian	8	india	8	latter
9	european	9	do	9	river
10	understood	10	was	10	west
11	international	11	may	11	diamondbacks
12	of	12	families	12	relocations
13	irish	13	began	13	end
14	canadian	14	we	14	pending
15	borges	15	had	15	following
16	united	16	said	16	current
17	trade	17	is	17	basenjis
18	d	18	appears	18	cell
19	other	19	wanted	19	sacrificing
20	autres	20	once	20	parking

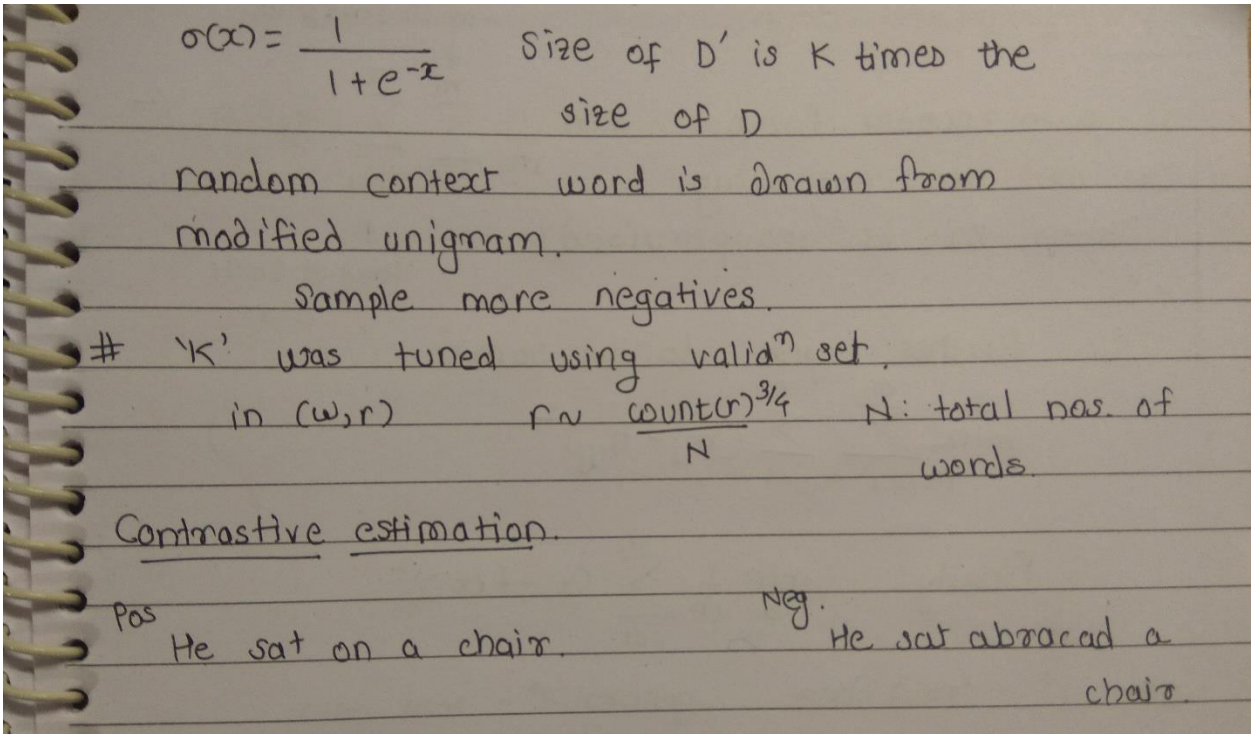
Q. 4 NCE loss summary

The traditional models developed for capturing word representations from the huge chunks of unstructured data don't scale well. They mostly become computationally expensive because of the expensive SoftMax functions that they deploy.

The whole point of Negative Contrastive Estimation is to reduce the training time by transforming a multiclass problem that uses SoftMax function to a binary class problem that uses a relatively inexpensive alternate function. Therefore, they call it a 'log-bilinear' model.

I will try to justify its efficiency and working through ***my notes from Prof. Mitesh Khapra's Lecture.***





As per my notes above, NCE is about reducing the computational overhead of SoftMax function. Models trained by NCE strategy are at par with the models trained by NPLM strategy. The scores in NPLM are converted to probabilities by exponentiation and normalization by following function.

$$P_\theta^h(w) = \frac{\exp(s_\theta(w,h))}{\sum_{w'} \exp(s_\theta(w',h))}.$$

But, in NCE we can drop the denominator in the above equation and just use the numerator. So, to learn the distribution of words for a specific context *h*, we create an auxiliary binary classification problem. For this, we take the positive examples as the pair of *(context, target)* words from the actual corpus and negative examples from a noise distribution. For, the given context, global unigram distribution is used to sample noise.
Therefore, task to identify if our sample came from positive distribution is to just figure out it's probability over (all positive sample + *(k)*no_negative_samples*). It's denoted by the following equation.

$$P^h(D=1|w,\theta) = \frac{P_\theta^h(w)}{P_\theta^h(w) + kP_n(w)} = \sigma\left(\Delta s_\theta(w,h)\right)$$

The code in the assignment was implemented, by studying the following equations.

$$J(\theta, Batch) = \sum_{(w_o,w_c)\in Batch} -\left[\log Pr(D=1,w_o|w_c) + \sum_{x\in V^k} \log(1 - Pr(D=1,w_x|w_c))\right]$$

where,

$$Pr(D=1,w_o|w_c) = \sigma\left(s(w_o,w_c) - \log\left[kPr(w_o)\right]\right)$$
$$Pr(D=1,w_x|w_c) = \sigma\left(s(w_x,w_c) - \log\left[kPr(w_x)\right]\right)$$

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

and

$$s(w_o,w_c) = (u_c^T u_o) + b_o$$

where u_c, and u_o are the context and the target word vectors, and b_o is a bias vector specific to w_o.