Department of Information and Computer Science
Aalto University School of Science

December 14, 2013

Doctoral Programme Committee
Aalto University School of Science

Dear Members of the Doctoral Programme Committee,

I, Kyunghyun Cho, would like to clarify the corrections made to the draft of the dissertation *Deep Neural Networks: Basic Principles and Recent Advances* (supervisor: Prof. Juha Karhunen, advisors: Dr. Tapani Raiko and Dr. Alexander Ilin) according to the comments from the pre-examiners Prof. Hugo Larochelle and Dr. James Bergstra.

I would like to thank the member of the committee for approving the excellent pre-examiners. Both pre-examiners examined the dissertation throughly and made comments that are invaluable. By carefully studying their comments and revising the dissertation accordingly, I believe, the overall quality of the dissertation has improved.

Please, find enclosed in this letter the detailed list of corrections.

Best regards,

Kyunghyun Cho

encl: Clarification of the Corrections

KyungHyun Cho December 14, 2013

Clarification of the Corrections

1 Prof. Hugo Larochelle

On **the use of additive noise ϵ in Eqs. (2.2) and (2.11)**:

I agree with the comment. As it was already mentioned before both the equations that the observation is noisy, it is unnecessary to have the additive ϵ in the equations. I have removed the ϵ from both the equations and changed the text accordingly.

This was also pointed out by Dr. Bergstra.

On **the constraint of perceptron convergence algorithm on the data representation**:

As Prof. Larochelle pointed out, it was my mistake to state that the original perceptron convergence algorithm worked only with -1/1 representation. I have modified the text accordingly to reflect that the algorithm works also with 0/1 representation.

On **the wrong statement on the fixed-point update rule of the minimum reconstruction error formulation of PCA**:

The minimum reconstruction error formulation without variables specifically defined for hidden variables does not yield fixed-point update rules for PCA, as Prof. Larochelle pointed out. This was my mistake, and this has been removed and rephrased.

This was also pointed out by Dr. Bergstra.

On **the potentially misleading statement on the autoencoder and variational inference**:

Prof. Larochelle correctly pointed out that it may be misleading to simply state that the encoder of an autoencoder performs an approximate inference without explicitly stating that the encoding procedure does not minimize the KL divergence between the variational posterior distribution and the true posterior distribution. I have rephrased the corresponding paragraphs to explicitly state this.

On **the description of the DBM pretraining**:

There were some misleading as well as unclear points in the text. The description of the pretraining algorithm has been strengthened, especially, on the points made by Prof. Larochelle.

On **the presentation of the Metropolis-Hastings algorithm**:

As suggested by Prof. Larochelle, I have put the algorithmic description of the Metropolis-Hastings algorithm in a more standard presentation using a box.

This was also pointed out by Dr. Bergstra.

On **Table 5.1**:

I agree with Prof. Larochelle that it may be misleading to state that the autoencoder is approximate while the deep belief network is more exact. Also, to make the presentation more clear, I have changed the labels of the columns to 'recognition' and 'generation'.

On **mistakes in English**:

Prof. Larochelle has kindly gone through each line of the dissertation to spot small mistakes in English such as typos. I have gone through each and every one of them and made appropriate changes.

2 Dr. James Bergsta

2.1 Technical Comments

On **the potentially misleading statement on the equivalence between PCA and the linear autoencoder**:

I agree with the examiner's concern and have softened the claim by explaining the possible ambiguity in rotation and scaling of solutions found by the linear autoencoder.

On **the relationship between the stochastic gradient descent method and the backpropagation algorithm**:

The use of the term *backpropagation* is often misused to refer to the learning algorithm for a multilayer perceptron. Dr. Bergstra rightly pointed out this misuse in the dissertation, and I have made corrections to make it clear that the backpropagation is an algorithm that computes the gradient efficiently, while the stochastic gradient descent method is an algorithm that utilizes the gradient to estimate the parameters.

On **the two conditions for deep neural networks**:

I agree with Dr. Bergstra that it may be an unnecessary effort trying to precisely define the conditions of deep neural networks. Although the pre-examiner went so far as to drop the whole section, I believe that it is important to clearly discuss these conditions if one is to understand the principles of deep neural networks. I have kept the section but have softened the text as was suggested by Dr. Bergstra.

On **the discussion of recurrent neural networks and their relationship to Boltzmann machines**:

As Dr. Bergstra as well as Prof. Larochelle have pointed out, the recurrent neural network is not directly connected to the publications included in the dissertation. In my opinion, however, it may, to some, not be clear why Boltzmann machines, other than deep Boltzmann machines, are deep neural networks without the explicit connection between Boltzmann machines and recurrent neural networks. Hence, I have left the section discussing the recurrent neural network intact, while I have modified a few other places where I have mentioned recurrent neural networks without any strong motivation, as suggested by the both pre-examiners.

On **the nonstandard definition of classification and regression**:

I agree with Dr. Bergstra as well as Prof. Larochelle that the definitions of classification and regression I originally gave in the dissertation are nonstandard. I have modified them to be more in line with standard terminologies and definitions.

On **the lack of further discussion on regularization**:

I acknowledge that the methods of regularization are worth more than a simple description I gave initially. However, I find the regularization be out of scope for my dissertation. I have modified the single-sentence description of regularization to be more in line with Dr. Bergstra's comment, but have not added any new section on the matter.

On **the XOR problem**:

The XOR problem is itself neither interesting nor motivating, but its property of being separable but not linearly is, in my opinion, more than important to motivate deep neural networks.

On **the Hopfield networks**:

Although Dr. Bergstra suggested removing the section on the Hopfield network, I believe that it is an important neural network model that motivates Boltzmann machines. This is in line with, for instance, MacKay (2003).

On **referring to Haykin (2009) for multilayer perceptrons and other algorithms**:

I agree with Dr. Bergstra that Haykin (2009) is an inappropriate reference unless with chapters or pages specified. I have corrected this mistake in multiple places by, for instance, referring to more original papers (e.g., citing Rosenblatt (1958) instead for multilayer perceptrons).

On **the description of SVM and kernel methods**:

I have changed the potentially misleading statement about the max-margin principle being a good regularizer. I have revised it to say that the max-margin principle provides a well-formed ground on which a model can be selected. Otherwise, I believe a connection from a neural network to SVM is important, so except for softening some text, I have left the section intact.

On **the conjectures on ELM**:

Dr. Bergstra disagreed with my conjecture that ELM will not benefit from having multiple hidden layers. I believe what Dr. Bergstra is referring to differs from what I have mentioned in the text. It has been shown by Dr. Bergstra recently that a so-called null model which has a sophisticated architecture (specifically, convolutional feature detection) but randomly chosen parameters may be used to perform well, but in the case of my argument in the dissertation, ELM does not contain any specialized structure but fully connected layers only. However, I do agree that my wording has been too strong and definite and have toned down and when necessary removed some sentences.

On **denoising autoencoders**:

Dr. Bergstra questioned why the hidden representation of a mid-point between two training points does not collapse to the hidden representation of either of two training point. This happens because of the reconstruction error term and the assumption of a single hidden layer in the text, which I believe Dr. Bergstra mistakenly missed.

On **layer-wise pretraining**:

Dr. Bergstra correctly pointed out that the layer-wise pretraining does not guarantee to disentangle the factors of variations. I fully agree and I have used more commonly agreed term 'encourage'.

On **other minor comments**:

I have revised the text according to the minor comments made by Dr. Bergstra. Most of them were non-standard terminologies and typos in mathematical equations.

2.2 Detailed Comments on Writing

On **moving the section on SGD earlier**:

I agree with Dr. Bergstra that it is usual to use EM or other algorithms for training probabilistic models, not SGD. However, I wanted to make sure that the section explaining probabilistic approaches follow im-

mediately after the sections on linear, shallow models to encourage readers to maintain both perspectives (optimization, probabilistic) of same models. Hence, I have not moved the section.

On **moving the section on explaining-away to Chapter 2**:

I agree that explaining-away effect encourages researchers to move away from PCA towards other overcomplete models such as sparse coding. However, I believe the connection between explaining-away effect and feedforward recognition is more important to emphasize, so I left the section where it was.

On **the possible redundancy in Chapter 5**:

I agree that some readers including Dr. Bergstra may find the content of Chapter 5 be somewhat redundant with that of Chapter 3 and 4. This was done deliberately by myself in order to be precise and thorough about the content of Chapter 5 which constitutes the most important ground for the included publications. Hence, I have some few minor changes in the text but left most of the content intact.

On **informal writing**:

I would like to thank Dr. Bergstra for pointing out many of my rather informal writing styles. According to his suggestions, I have revised the whole text. His suggestions included:

1. not to use 'obvious', 'intuitive', 'clear', 'natural', 'easy', 'basically', straightforward', 'basic' and 'simple' to suggest that something is, e.g., 'obvious' for a reader to understand.
2. not to open a sentence or paragraph with 'however', 'with' and 'because', unless absolutely necessary.
3. not to use a pronoun that refers to a noun in previous paragraphs.

On **using 'references therein'**:

Dr. Bergstra rightly pointed out that the abuse of 'references therein' shows the lack of clear guidance in showing the original reference. I have remove all but one occurrences of 'references therein' by replacing the existing references with more original and precise references.

On **subtle beginning of sections**:

Dr. Bergstra was kind enough to point out many sections that had some subtle beginning. I have fixed many of the sections to read more naturally.

On **other minor comments**:

Dr. Bergstra kindly pointed out many linguistic, grammatically errors in the text, and I have corrected them accordingly.