1 Guidelines for Pre-Examiner

Aalto University aims at the highest possible quality of its dissertations.

- A dissertation must contain new scientific findings in its area of research.
- Research methods and new research findings are to be presented clearly and they should withstand the scrutiny appropriate for scientific research.
- The candidates own contribution to the research shall be sufficient and clearly stated.
- A dissertation must follow good scientific practice and ethically sustainable principles.

The manuscript of the dissertation submitted for pre-examination must be complete and its language must be faultless.

Article dissertation An article dissertation may comprise scientific articles and/or conference publications which should be peer-reviewed and published or accepted for publication in a scientific series or some other work.

Examination: An article dissertation is to be regarded as a whole even though individual articles have been approved in peer-reviewed series or other works. In such a case, the examination is directed to the summary and the dissertation as a whole. While preparing a statement of a dissertation manuscript, the pre-examiner shall evaluate whether the candidates contribution to the thesis is sufficient.

The purpose of the pre-examination is to establish whether the manuscript fulfills general quality requirements. Therefore, particular attention should be paid to the following aspects:

- The scientific correctness of the dissertation, and the clarity of the hypotheses or the objectives of the research, particularly when examining parts that have not been peer-reviewed yet
- The sufficiency of the doctoral candidates own research input (particularly if the dissertation includes co-authored publications)
- The significance and status of the dissertation in the field
- The scope of the dissertation and the sufficiency of the material
- The doctoral candidates ability to obtain results from the material examined in the dissertation
- The logic of the dissertations structure
- The knowledge and use of literature in the field and the ability to use references correctly
- Language

In his/her statement, the examiner should give a summation of the most important results and merits of the dissertation along with discovered shortcomings. The examiner is also asked to compare the dissertation manuscript to other dissertations of his/her university or to other dissertations that he/she has examined, and evaluate whether the work belongs to the top 10%. The examiner is asked to mention this in his/her statement.

If a pre-examiner does not suggest corrections, 1-2 pages will suffice as the length of the statement

2 Pre-Examination Report

Kyunghyun Cho's doctoral work investigates what have come to be called "deep learning" algorithms. These algorithms use statistical learning techniques to discover neural networks that satisfy engineering objectives such as function approximation and density modeling. Learning techniques for searching this promising model class are both computationally and statistically slow, almost to the point of making search inviable. Cho's publications contribute two novel techniques for estimating the error gradient used for learning Boltzmann Machines, and an extension of the DBM model family to facilitate certain applications (image modeling, image in-painting, de-noising, and missing data imputation). His doctoral work provides more-efficient strategies for model search, and evidence that models thus discovered have potentially valuable properties. This work both lowers the technical barrier of entry to the field for researchers and increasing the motivation to explore deep learning algorithms for engineering ends.

Scientific correctness of experiments:
Quite sufficient.

Sufficiency: The doctoral candidate's own research is quite sufficient. Ten publications is an extraordinary output from a single PhD degree.

Significance and status in the field:
Quite sufficient. Google Scholar counts approx 70 citations to the publications that make up this dissertation. Despite working outside of the dominant "deep learning" labs, Cho has tackled widely recognized problems and earned a place in the literature of the field.

Scope of dissertation and sufficiency of material:
Quite sufficient.

Candidate's ability to obtain results:
Quite sufficient. The publication of so many papers in competitive conferences is compelling objective evidence that Cho can obtain results.

Logic of dissertation's structure:
Acceptable; however, I struggled at times with the structure of the non-peer-reviewed portion of the thesis (see below.)

Knowledge and use of literature of the field:
Sufficient. Cho demonstrates a great deal of practical and theoretical knowledge of the field.

Language:
Borderline. In terms of style and tone, some passages of the non-peer-reviewed portion of the dissertation are insufficiently formal for academic writing (see below).

I have organized my detailed comments into two sections: Technical, and Writing. These comments all relate to the non-peer-reviewed portion of the dissertation. The peer-reviewed portion is polished, easy to understand, and I do not have any issues with the content. I would say that the work belongs among the top 10% of dissertations, despite shortcomings in the writing of the current manuscript.

Dr. James Bergstra, Ph.D.

3 Detailed Technical Comments

In this section, I have enumerated several more-or-less minor technical comments related to the non-peer-reviewed part of the dissertation. None of them is very serious, but they are either worth thinking about, or else they should be easy fixes to the text.

- Pg 20: "corresponds exactly" is too strong, because a linear autoencoder learns a rotated and possibly scaled PCA basis.
- Pg 21: over-precise definition of "Shallow" models here (only input and output units) runs contrary to standard usage and the author's own development later on.
- Pg 21: subtle point re: use of term "forward computations". Author wrote "From this probabilistic perspective, forward computations in neural networks are interpreted as computing the conditional probability..." But I think the author meant that the computations in neural networks can be interpreted as forward computations in a graphical model.
- Pg 22: "universal learning algorithm" is an unusual descriptor for gradient descent (with its local minimum problems), and anyway has been used to mean other things.
- Pg 22: The two "conditions" that allow the author to classify Boltzmann Machines as deep have not yet been presented. Regardless, I would suggest the author to drop the attempt to set out conditions for the "deep" label because the dissertation does not use the conditions to establish new formal or technical results. Why bother being precise about a definition that is not used for anything? Some intuition for why some things are called "deep" is appropriate, but the several passages in the dissertation that try to prove that a commonly-accepted label is also consistent with some definition seem unnecessary to me.
- Consider dropping discussion of recurrent neural networks: what do the algorithms, models, and applications associated with recurrent networks serve to motivate publications? I do not see the connection. Recurrent networks pop up in a couple of places, and I think that in each place they do more harm to the clarity of the dissertation than good.
- Pg 28: Nonstandard definition of classification and regression. Would ten thousand classes become regression?
- Pg 29: Gaussian noise ϵ gets ahead into probabilistic interpretation, and anyway regularization is written a few sentences later in terms of another symbol λ.
- Pg 29: Regularization is a subtle but crucial concept in understanding most machine learning research. To simply say it is "an often-used method for preventing an estimated model from overfitting to training samples" is inadequate and not even entirely accurate. I would suggest replacing this sentence with an entire section.
- Pg 31: I have never heard the term "nonlinearly separable" and I do not understand it. I wonder if the author simply meant "not linearly separable."
- Pg 32: Solving XOR is an unusual motivation for what is now called *deep learning* (!)
- Pg 32: Section 2.2 is notably missing a sentence of the form "Unsupervised learning is ...", and the text preceding 2.2.1 is awkward. Consider drawing a distinction between density estimation and feature extraction as two defining and not-necessarily-identical goals of unsupervised learning. Clarity here is especially important in the context of the dissertation because these objectives are the ones to which the author is appealing in several of the publications.

- Pg 34: In the context of explaining a linear autoencoder, it is not obvious why q having to be smaller than p is a problem or why anyone would want to "work around" it. If there had been a distinction between density modeling and feature extraction, it might be easier to explain. Also, this might be a good place to get into spare coding and explaining away.
- Pg 35: Consider deleting the section on Hopfield networks. (Does it motivate publications?) If the author decides to keep it, make sure to be clear that it is meant to be a recurrent network, and that it operates on binary data (unlike the previous models discussed).
- Pg 42: why are we re-deriving the probability-free version of PCA? Probabilistic PCA was introduced by reference to non-probabilistic version.
- Pg 45: Eq 2.34 appears to be missing a division by N.
- Pg 47: I can appreciate why the author would say that exotic methods such as TONGA [capitalize it] and Hessian-Free are out of scope of this thesis because the publications only use SGD, but there is more to it than that. Both the author's work and these other methods are, broadly speaking, responses to the need for better training algorithms for deep networks. What are the issues with SGD that these other methods are trying to address? Are these algo issues with the e.g. enhanced gradient? How might one combine the enhanced gradient with these other methods? Would that be a good idea? In the interest of motivating the author's work and situating it in the landscape of research on optimization and learning for deep models, I would suggest going into more detail.
- Pg 49: Haykin 2009 is unusual references for "multi-layer perceptron", it is a huge textbook with broad scope and hundreds of references. Consider a citation to the PDP books that introduced backpropagation or maybe Minsky's book that introduced the perceptron without learning.
- Pg 50: The top un-numbered equation should be numbered, and the ϕ symbols should have subscripts indicating that they may be different for different layers no?
- Pg 51 (and later in section 3.4) Be careful not to call backpropagation a learning algorithm. Gradient descent (of whatever variety) is the learning algorithm; backpropagation is an implementation of the chain rule, which computes the gradient.
- Pg 51: The passage that begins "Another way to look at what each intermediate layer does is to consider it as a *feature detector*" would be much clearer if the author began by defining what he had in mind by a "feature" and then proceded to explain how it is is that a neural network detects it. This point is important, and the text of page 51 is not clear.
- Pg 52: Writing "The maximum-margin principle provides a principled way of regularizing the parameters of a model." is confusing in two ways. First, it suggests that the regularization methods seen earlier are not principled (I do not think this is true, did the author mean this?) Second, under my definition of regularization, max-margin is a loss function that has nothing to do with regularization. The author did not define regularization so perhaps there is a misunderstanding on that front.
- Pg 52: Introducing kernel methods by writing "A kernel method employed by an SVM does a similar thing to what the intermediate hidden layers of an MLP do." Is both stylistically and technically weak. Kernel methods are also not obviously related to the author's work. Consider cutting this whole section on SVMs?
- Pg 52-53: This passage includes discussion of whether SVMs and MLPs with a single hidden layer are shallow or deep. Why does it matter? Consider cutting the discussion. The author appears to revert to the common terminology rather than the two criteria he set out earlier, but I would rather see this passage removed than corrected in this respect. Similar comments for the discussion of the ELM on pages 53-54. The last paragraph on ELMs contains unsubstantiated conjecture that I happen to disagree with, and which anyway has no obvious role in the introduction of the author's work, so I would suggest deleting that paragraph in particular more strongly.

- Pg 59: Text is missing a description of the inference algorithm for Sigmoid Belief Nets?
- Pg 60: Comparison to Deep Belief Network should wait for later, after DBN has been introduced.
- Pg 66: The suggestion that regularization in general operates by affecting the range $\mathbb{Q}$ seems misleading because most regularization strategies work by adjusting densities over $\mathbb{R}$. It is important to situate and motivate the author's sparsification technique, but I think the key point (that train-time density-based regularization strategies are insufficient; hard, inference-time range constraints work better) is not made clearly enough here. What is is missing from PSD, that the author added/corrected in his contributions?
- Pg 68: The passage about disentangling factors of variation should (in my opinion) be tightened up, in order to not overstate the advances of deep learning. For example, although the author writes "we cannot expect that this type of transformation that captures the data manifold happens when we estimate the parameters of an MLP," I believe that this is exactly what happens when we train an MLP. If MLP training works, how else could it work except by disentangling the requisite factors of variation? The interest in layer-wise pre-training mainly pertains to cases where MLP training fails. The author writes that MLP training is "not likely nor guaranteed" to capture the data manifold, but "Instead, there is another approach that gradually builds up a sequence of transforms... capture the data manifold." The writing here suggests that layerwise pretraining is guaranteed to disentangle factors of variation. While disentangling has not been defined very precisely here and there is consequently room for misinterpretation, I am fairly confident that disentangling is not guaranteed by current algorithms.
- Pg 69: Why does "ignoring any direction of variation injected by κ" reveal the data manifold? It appears that κ injects variation in every direction. The explanation through pages 70 and 71 did not help me. If x and x' are nearby, why do they not map to the same h for the same reason that $x + \epsilon$ map to the same h. The intuition given would make sense for Neighbourhood Components Analysis or Locally Linear Embedding (which should arguably be cited as part of a brief introduction to manifold learning algorithms), but I do not understand it in the context of the sigmoidal denoising autoencoder.
- Pg 75: The term "stochastic backpropagataion" is non-standard, usually this is called SGD.
- Pg 76: What does it mean to "poorly estimate" some parameters within a non-identifiable model? I am not sure this can be fixed by rephrasing, the entire section seems to hinge on the idea.
- Pg 76: I do not understand the "hypothesis" stated in para 2. If the deep network works like a shallow network to reduce training error, then why does it not work as well as a shallow network at test time? Think about how this section might be re-written to be a clear motivation for the author's own work on learning in deep models.
- Pg 81: the word "equivalent" appears right over top of an $\approx$ symbol, why not an $=$ symbol?
- Pg 93: I did not follow the explanation of parallel tempered MCMC. Consider rewriting to be clearer regarding: What data structures are involved, how are they updated in the course of the algorithm, in what cases would this work better than a non-tempered chain and in what cases would it work worse?
- Pg 107: Eqn 4.31 is missing a $\lim_{t \to \infty}$
- Pg 109: Define "score" in the technical sense in which it is used here. Score has an informal meaning, which could be confusing.
- Pg 111: Eqn 4.40 is not look like the energy function for a Boltzmann machine, beyond also being "an energy function". The text introducing Eqn 4.40 is confusing.

As you'll note, the density of technical comments decreases from the beginning of the dissertation to the end. Although reviewer fatigue may have played a role (sorry!), I also feel that the text was much stronger toward the end, as the subject matter approached the algorithms and models that are the focus of the publications.

4 Detailed Comments on the Writing

The pre-examiner's guide specifically demanded that dissertation writing be "flawless" but the current version of the non-peer-reviewed portion of the dissertation has left some room for improvement. I have organized my detailed comments on the writing into two areas: Organization and Style. The "Organization" section makes some recommendations for the overall line of introduction and motivation. The "Style" section enumerates some examples of imprecise writing.

4.1 Organization

The principal (and significant) contribution of this doctoral work is the set publications starting on page 160. These publications cover:

1. Training algorithms for RBMs (pubs I, II, III, IV, V, VII)
2. Expanded model class for DBMs (pubs V, VI, IX, X)
3. Applications: image classification, energy-based density estimation of image patches, image de-noising, inpainting and data imputation,.

Given that these publications represent the bulk of the doctoral work, the role of the first part of the dissertation is to introduce and motivate that work. The first part of the dissertation must help the reader to understand:

- What is the state of affairs in the field that motivated the author to do what he did?
- Relative to that introduction, what did the author do?
- What were the outcomes of the author's investigations, in terms of what others may now do?

The author might say that the state of affairs was, for example, that deep models were emerging as successful models for various applications (as he does in the first paragraph of the abstract), but that the most promising models remain too difficult to train and mathematically inappropriate for potential new applications. The author's contribution was to develop new better training algorithms, and to adapt existing models and algorithms to work in settings where they had not been used before.

My feeling is that Part I of the dissertation is not sufficiently focused on motivating and introducing the publications. I do not know to what extent my role as a pre-examiner entitles me to make broad editorial remarks, but I would recommend that the author start with a skeletal outline of the answers to the questions I wrote just above, and then look at each section and paragraph of the first part of the dissertation with a view to ensuring that every single passage makes a necessary contribution to developing those answers.

As the author works through the document, he may find it useful to make notes to himself regarding why various sections cannot be cut, or cannot be moved. Such notes would be the basis for introductory and summary sections throughout the dissertation.

Section introductions and summaries are critically important places for signposts to help a reader through a 160-page technical document. Imagine the reader as wanting to learn some thing Z, but not yet ready. At every new section the reader mentally wonders "Are we there yet? Am I going to learn about Z now?" The role of a chapter-introduction or section-introduction is to explain "No, we have seen X and Y, but before we get to Z we need to learn about e.g. furlijiggs..." Then the role of the summary is to remind the reader why they were even reading this section: "Remember you wanted to learn Z and I told you we needed to learn about fulijiggs? Well I have given you a basic intro: I told you ..." and optionally letting them know how to go on making their way through your document: "The quickest way to Z is to take a quick intro to jagoofars (Chapter 3) and then go on to Z in Chapter 4. The remainder of this Chapter is about how to make fulijiggs (Section 2.2) and how to use them to build things (Section 2.3). " Without this DAG in mind (which can only be maintained by constant reminders), a reader too easily becomes lost in a long document, starts paying attention to the wrong things, gets confused, and generally loses sight of the big picture.

I have made a list of some specific opportunities for re-organization that came to mind as I read, but they should not be taken as a necessary or sufficient set of changes.

1. Intro - consider addressing additional questions: What is relationship between the neural networks that achieve the application successes and the author's own work? Do they motivate his research? Where it has not been shown directly, is there reason to believe that SoA neural networks would benefit from including this research? How does this research open up future research fronts in deep learning?

2. Pg 29: The description of which methods for linear regression require analytic solutions, which require optimization algorithms, and how those algorithms work seems unclear (possibly unnecessary?).

3. Consider cutting Hopfield network. I don't think it motivates the publications or helps the reader to understand their contributions.

4. Pg 44: Consider cutting the section "What makes neural networks deep". It is out of place between probabilistic PCA and stochastic gradient descent. The two criteria from Bengio and LeCun 2007 should appear earlier.

5. Pg 45: Consider moving SGD earlier, even as early as the non-probabilistic linear auto-encoder. It is strange to introduce SGD after Probabilistic PCA because PPCA is usually handled by the EM algorithm.

6. Pg 62: Consider discussion of explaining-away to much earlier, alongside PCA for example, near the passage that discusses why it is problematic that the number of latent variables must be smaller than the number of observed variables.

7. Pg 81: Consider delaying discussion of CD until later.

8. 4.2: Consider deleting this section. Why is it important to establish that Boltzmann machines are deep, or that they bear some relation to recurrent neural networks?

9. 4.3: Consider moving this section on how to estimate the parameters of Boltzmann Machines *before* the authors' own contribution regarding the enhanced gradient.

10. 4.5.2: describes training procedure for DBN, but then section 5 begins by suggesting to the reader: Would it not be a good idea to do layer-wise unsupervised pretraining? Consider reversing the order of these sections. Also, section 5.1 includes a lengthy description of how a feed-forward MLP works, and repeats the disentangling story from the manifold learning section. Consider combining this with the previous location that discussed disentangling.

11. Several sections in part 5 seem somewhat redundant with earlier material ("Auto-encoders and Boltzmann Machines", "Pretraining generative models") consider shortening this section and focusing on what the reader needs to know to appreciate why the author's contributions are important.

12. Part 6 (Discussion): This section is missing a reminder of what problems the author observed in the field of deep learning, how the author addressed them (now, unlike the introduction, it is OK to use technical terminology and state the contributions in full force), the degree to which the problems have been addressed, and what new areas of research have been opened up. An enumeration of "Matters Which Have Not Been Discussed" seems comparatively out of place; if they weren't important enough to put before the summary, why bother?

4.2 Style and Language

There are some mundane technical writing issues that occur consistently throughout the dissertation, which are easy to fix:

- technical terms should be *italicized* when introduced, and defined immediately,
- footnotes should be written *after* a sentence-ending period,
- the author's voice should be fixed to either "we/us"-style (e.g. pg 28) or "the author"-style (e.g. pages 15, 21).

The author should search the text for case-insensitive matches to "obvious", "intuitive", "clear", "natural", "easy", "basically", "straightforward", and "simple"; and wherever one of these words is used to tell the reader how e.g. "obvious" something is, it should be removed. For example: "It is obvious that there exists more than once computational path" should be written: "There is more than one computational path", or perhaps "Since there is more than one computational path". "It basically states" should be rewritten: "It states" and so on.

Many passages in the dissertation could be condensed and written in a more formal style. For example in the abstract we find the sentence "In the first part of this thesis, the author aims at investigating these models and finding a common set of basic principles for deep neural networks." Syntactically this could be tightened up: a reference to "these models" in a paragraph-opener is bad form, the phrase "aims at investigating" is awkward, and the word "basic" is not adding anything. Making these modifications might yield e.g. "The first part of this thesis investigates several deep models in search of common principles for deep neural networks" which is shorter, and conveys more information. The shorter form also makes it clearer that there is a semantic problem: a "principle" is a general, simple law that explains multiple phenomena, but "deep neural networks" are not phenomena in need of explanation. The author might clarify the situation by writing e.g. "The first part of this thesis investigates shallow and deep neural networks in search of principles that explain why deep neural networks work so well across a range of applications."

Here are some other sentences from throughout the dissertation that represent opportunities for clearer, more formal writing (not an exhaustive list).

- Opening second paragraph of 1.1: "*Although the recent surge of popularity stems from the breakthrough* [what breakthrough?], *involving a* [cut 'a'] *layer-wise pretraining* [terms have not yet been defined], *in 2006, research on artificial neural networks in general* [cut 'in general'] *has begun* [wrong tense] *as early as* [unnecessarily informal language] *1958 when Rosenblatt (1958) proposed a perceptron learning algorithm* [terms not yet defined]." Consider splitting this into shorter sentences.
- Opening third paragraph of 1.1: "*These types* [antecedent in previous paragraph?] *of artificial neural networks are interesting not only on their own* [the author frequently describes things as interesting, it is important to explain to the reader **why** something is interesting], *but by connections among themselves* [watch out for ambiguity: neural networks are "connectionist" models] *and with other machine learning approaches.*" I do not happen to agree that the existence of connections between statistical models makes those models more interesting, and I wonder, as a reader, why I should either know or care whether the author finds something interesting.
- Opening para 3 of 1.2.5: "*Although deep neural networks have shown extremely competitive* [what does that mean?] *performance in various machine learning tasks, the theoretical motivation* [strange term] *for using them is still debated* [by whom?]." If "extremely competitive" means "the best" then the reason to use them seems clear enough. If the author means to point out "although they work, we do not really know why" then rephrase.
- Pg 28: Is author using "neural network" to refer to the *graph* of a function? "*With this linear unit as an output unit, we can construct a simple neural network that can simulate the unknown function f*"
- Pg 30: "*A perceptron can perfectly simulate the unknown function f when the training damples are linearly separable. It means* [Author does not mean "Consequently" so write "linear separability means"] *that there exists a linear hyperplane that separates...*" The definition here could be tightened up too.

- Pg 31: "*Note that it is not necessary* [for what?] *to use a Heaviside step function. It is possible to use any other nonlinear saturating function... One possibility is the sigmoid... Another possible choice is the hyperbolic tangent... The set of weights can be estimated in another way... However, in this case the cross-entropy cost function can be used.*" As a reader, I wonder why the author is enumerating all these possibilities. One good reason would be that the publications later do in fact make these choices, so the reader should pay attention. This would be a good opportunity to also remark on what may well be a possibility but would not, in fact, be a good choice.

- Section 2.2.1 begins: "*Firstly* [of how many? what list are we in?], *we look at the case where hidden variables are assumed to have generated training samples* [linearly, right? section titles are not considered part of the text.]. *In this case, it is desirable* [by whom? for what?] *to learn* [for the reader to learn? or the model?] *not only an unknown function f, but also another function g which* [that] *is an inverse function of f. Opposite to f, the inverse function* [previous text is redundant] *g recognizes a given sample by finding a corresponding state of the hidden variables* [is this introducing a technical definition for "recognizing"? What if f is not technically invertible? (It usually is not.)]. *Let us start* [start what?] *by constructing a neural network with* [exclusively?] *linear units...*"

- Pg 47 has a paragraph that begins "Hence, ". The word "hence" should not even be used to begin a sentence, much less a paragraph.

- Pg 49: A citation directs the reader to "*Haykin, 2009, and references therein.*" This textbook as almost one thousand citations covering a huge variety of topics. As a reader, I feel like the author is telling me to "google it", and dodging his responsibility to provide more precise guidance. If a statement is so broad that a citation to an entire textbook is appropriate, consider attaching citations to more precise nearby sentences instead. If a reference to a textbook really is the right option, be precise about a chapter, a page, or some particular passage. There are a few other occurrences of references to a textbook and "references therein" which should also be removed.

- Pg 51: Paragraph begins with "*This model is a typical example*". As I have pointed out earlier, it is bad form to start a paragraph with a pronoun. This occurrence is doubly bad because the nearest model in the preceding text is not at all a "typical example", so if I were not chalking the writing up to being confusing I would say it was factually wrong.

- Pg 76: Section 3.4.1 starts: "*However, it has been noticed by many researchers that it is not easy to train deep neural networks using the plain stochastic backpropagation (see, e.g., Bengio and LeCun, 2007, and references therein). Especially, neural networks with more than two intermediate layers of hidden units usually result in a worse generalization performance than those with only one or two intermediate layers (Bengio et al., 2007).*" Opening a section with the word "*However*" is bad form, and there are some questionable word choices here, but more importantly, the sentences here are in the wrong order. Compare to e.g. "It has been observed that neural networks with more than two intermediate layers of hidden units usually exhibit lower accuracy than networks with only one or two intermediate layers (Bengio et al., 2007). [Begin by introducing relevant evidence, the motivating phenomenon that requires discussion.] This may seem surprising because a multilayer network can approximate a linear one, and so the bottom layers of a deep network should be able to immitate the linear first layer of a shallow one. [Spell out the assumption.] However, Bengio and LeCun (2007) point out that this is not actually what happens... [Set up the story.]"

- Similar to the previous item, and related to some of my comments on organization, I found that many early sections began too abruptly. For example, one early section begins "Consider a case where a set D of N input/output pairs is given". Such an opening makes me wonder: Why? Where are we going? What should I pay attention to? A better opening might be "Several of the author's publications (I, II, ...) deal with what is known as supervised learning. *Supervised learning* is the general problem of function approximation from data. The problem of supervised learning can be stated mathematically in terms of a data set D of N input/output pairs..."

- Pg 90: The reuse of the text typeface for writing pseudocode is nonstandard in the field. Consider e.g. LaTeX's algorithm package.
- Pg 91: Paras begin "With this proposal distribution", "Because the conditional probability...", then "However, this sampling-based approach...", then "The latter one is especially true...". Every one of these clauses breaks standard English style-guide recommendations. No sentences (much less paragraph) should begin with these uses of "Because" or "However" and pronouns that refer to nouns in previous paragraphs should be repeated for clarity.