ACADEMIC WRITING IN ARTIFICIAL INTELLIGENCE: AN EXPLORATORY ANALYSIS ON CONVENTIONS

Alicia Wang
College of Engineering
University of California, Berkeley
Berkeley, CA 94720, USA
aliciawang@berkeley.edu

ABSTRACT

As a subfield of software engineering, artificial intelligence (AI) is still new compared to other established disciplines such as mechanical engineering or physics. Because of this, writing conventions used within academic research for AI are not as established, meaning that guides are sparse and it is more difficult for incoming members to learn how to properly write a well-structured paper in AI. This paper presents an exploratory study on what the AI community values in research by open coding a corpus of papers from the 2018 International Conference on Learning Representations. We present our findings and provide our analysis on the sections that a good AI research paper should emphasize: related literature, robust experimentation, and specific future implementations.

Keywords: research design, research paradigms, artificial intelligence, academic writing

1 INTRODUCTION

With the recent developments in software engineering, artificial intelligence (AI) has grown in the past decade as a research field. Every year, more and more AI papers are published in various workshops and conferences as academic discourse within the field continues to expand. This increasing trend in academic papers begs the question: what determines a successful paper in this field?

In academic writing, the conventions used are often reflective of the specific field. For example, there are well-known research paradigms in other fields of science and engineering such as the experimental models of physics and the double-blind methods of medicine (Shaw, 2003). However, since software engineering has yet to develop its defining research paradigm, there have been no significant research paradigms or models; the most prominent research is a detailed overview and categorization of recent design trends and approaches in software engineering research (Shaw, 2003; Theisen et al., 2017). For artificial intelligence in particular, only a few guides have been published online but, even then, they are typically based on personal experiences and judgements; for example, a researcher has shared a popular AI research guide based on his time spent doing research at MIT (Silver). The most concrete guide online is written by the top AI conference, NeurIPS, yet is too brief to give substantial advice (Welling & Ghahramani).

The goal of this paper is to explore what aspects of academic papers are considered indispensable by the artificial intelligence community. Specifically, what are some of the criteria for a good artificial intelligence paper, and what are some examples that can better illustrate these points?

Using a database of papers from the 2018 International Conference for Learning Representations (an annual conference that is ranked second for artificial intelligence research by GoogleScholar), we examine what makes a paper successful or, in this case, admitted to the conference. We do so by analyzing the revisions researchers have made to their papers between the first submitted draft and the finalized admitted paper.

The results show that there are two significant sections in academic writing that the artificial intelligence research community favors: related works and experimentation. Through this research, this paper hopes to provide some insight on what the current community values to guide starting writers in this field.

1.1 Related Works

Very little has been published about discourse conventions of academic writing in software engineering. A particular work published in the 2003 IEEE conference, however, analyzes the trends and statistics of research papers submitted to the 2002 ICSE conference (Shaw, 2003). This work is formatted as both a guide and a collection of data, categorizing the different research question types, result types, and validation types of software engineering papers. It then offers advice to other researchers and introduces the most common writing techniques in this field. Although this work provides software engineers with substantial evidence and helpful advice, it is outdated and its corpus is too broad, as it uses papers from all fields of software engineering.

This study instead focuses on more current research pertaining to the field of AI. As mentioned previously, the conference committee for NeurIPS, the Conference on Neural Information Processing Systems, has published an informal guide on how to design a good research paper for artificial intelligence conferences (Welling & Ghahramani). The committee walks through the different types of AI research papers and the criteria for each. However, this guide is brief, lacks examples, and is targeted toward experienced researchers. In comparison, this research is more exploratory and will present more qualitative and quantitative evidence in terms of AI writing conventions.

2 Methodology

This research analyzes 15 oral papers from the 6th International Conference on Learning Representations (ICLR 2018). These papers are accessible through the OpenReview database, which allows users to view both poster and oral papers accepted to the conference. This corpus was chosen because of its accessibility to the public: its revisions, reviews, and ratings can be accessed by any curious user. In addition, we have limited our immense corpus to only include the oral papers instead of the poster papers, as oral papers are often suited for topics that attract wider audiences and whose main ideas can be grasped within the time limits of a presentation (ACL, 2017). The selection of 15 papers is randomly chosen from an initial database of 23 papers.

In these conferences, there is often a reviewing period after a paper is first submitted. After this review period, authors are able to make changes to their submission up until the final conference. For this research in particular, we will be analyzing the differences between the first submitted paper and the final paper presented at the conference. We find that these paper revisions will be beneficial in reflecting what the artificial intelligence community values because of the iterative process of peer review.

3 Research Process

3.1 Data

In "Guidelines for Writing a Good NIPS Paper" from the NIPS 2006 Program Committee, we find that our corpus of papers extracted from the ICLR 2018 conference are categorized as "Control and Reinforcement Learning" papers. These papers can be further organized by their purpose:

- Proposes new algorithm or model
- Proposes an algorithm with improved analysis
- Proves certain task is difficult or impossible

It was observed that these papers often combine some of the points listed above to craft their research. The most common of these papers will find some error within a current algorithm or design through an experimental model then propose a newer and better implementation. We can therefore presume that academic research in AI—especially in this conference—has a very specific writing procedure that takes from one or multiple of the above research styles.

3.2 Open Coding

For every paper, both the original text and revised text are placed side-by-side and compared. Text that was deleted from the original paper is highlighted in red. Text that was added in the revised paper is highlighted in green, as shown in Figure 1.

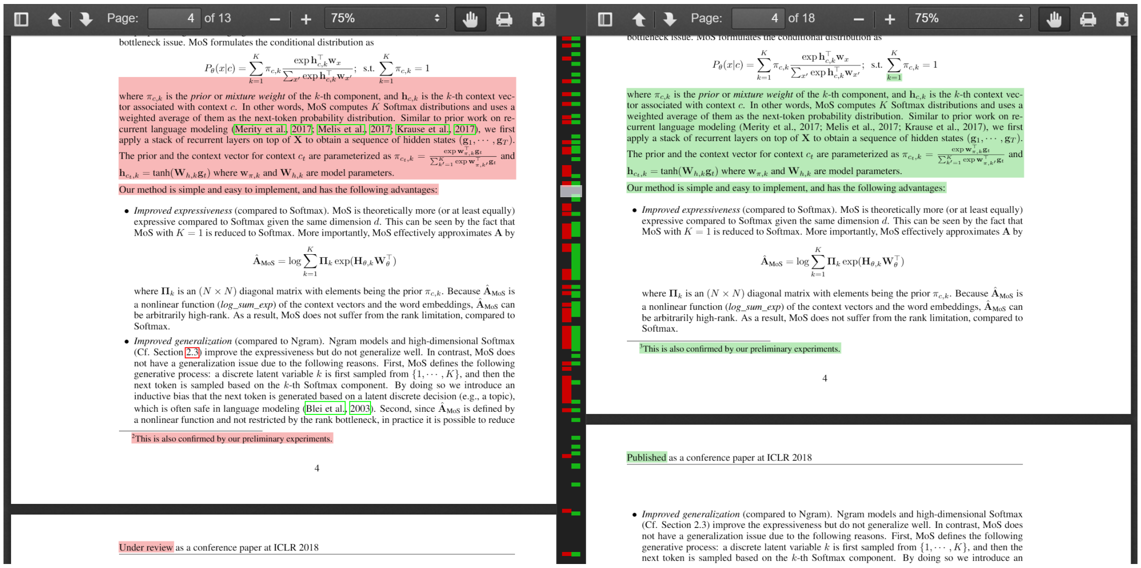


Figure 1: Using pdfdiff to analyze the texts

Here, a qualitative research method called open coding is used (Gibbs, 2007; Saldaña, 2009). Every significant change is labeled, and this is repeated for every paper. Minor revisions involving grammatical structure and spelling are omitted from the labelling process. After this first step, the labels are grouped into various categories. Note that this process is best done with a group rather than an individual researcher, as the collected labels may possibly be biased (Khandkar).

Five major categories have fit the labelled revisions of the open coded papers. The characteristics and statistics of every revision are summarized below (more detailed data can be found in the *Appendix*):

- **Related Work** ***(12 out of 15 of papers analyzed)***: This is every revision where the authors have added at least two lines of additional background to the paper, often in their "Related Works" section. In this case, they failed to properly credit or mention another related piece of work, so they have corrected this by adding additional literature.
- **Additional Experimentation** ***(8 out of 15 papers analyzed)***: For this type of revision, the authors have added additional findings to their experiments mostly in the "Experiments" and "Data" sections. They either appended entire rows or columns to their data or introduced entirely new tests with different parameters and metrics. There is a noticeable change in their presented data and subsequent analysis.
- **Future Applications** ***(7 out of 15 papers analyzed)***: In this case, the authors have added at least two lines of additional future work to their paper, often in the "Future Work", "Discussion", or "Conclusion" sections. The authors did not thoroughly state how their work will have an impact, in addition to any future implementations or experimentations for their model.
- **Mathematical Clarity** ***(5 out of 15 of papers analyzed)***: In this revision, the authors were not mathematically clear enough in their proofs or theorems, often adding more explanation or changing variables. Since many papers make some minor edits on their proofs, this category has been limited to only including papers that have changed entire sections or paragraphs of their theorems or proof.
- **Writing Clarity** ***(4 out of 15 of papers analyzed)***: In this revision, authors were not clear enough about what they were trying to achieve in their research. Most papers are well-

written, but many papers do have minor edits in language. Thus, this category has been limited to only including papers that have changed entire sections or paragraphs to explain the purpose of their research better.

4 ANALYSIS

In our analysis, we further explore the top three revisions we found through analyzing our corpus: “Related Work,” “Additional Experimentation,” and “Future Applications”. A few examples of each instance are provided and explained on how they might apply to the values in the AI research field.

4.1 REVISION: RELATED WORKS

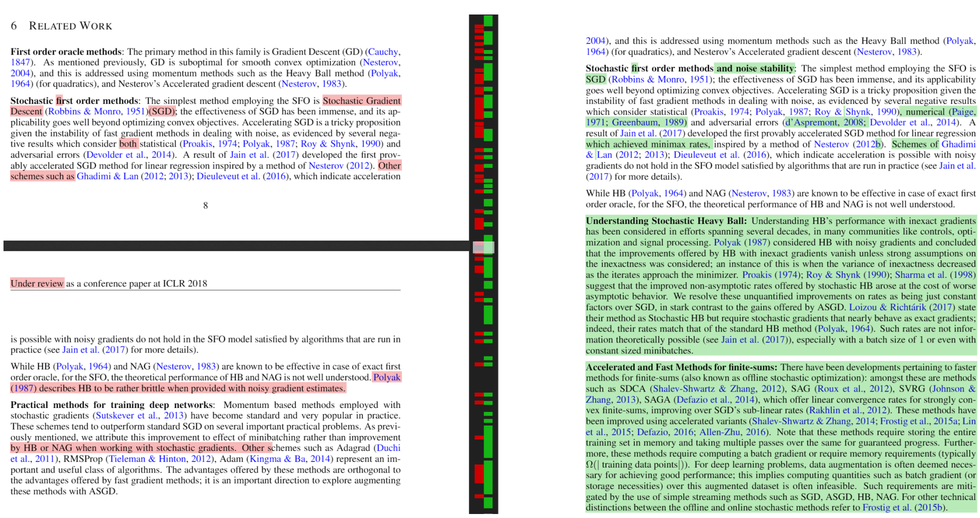
From our corpus, we have found that many papers made major revisions to their ”Related Works” sections. This is often pointed out in the community through peer reviews, where other researchers will mention how the authors have forgotten to mention a crucial piece of work in their paper. This revision is often a result of not being aware of related literature.

4.1.1 EXAMPLE

In the paper ”On the insufficiency of existing momentum schemes for Stochastic Optimization” by Kidambi et al. (2018), we can see that the authors have made some heavy revisions in the ”Related Work” section. In Figure 2, they have added two entire sections named “Understanding Stochastic Heavy Ball” and “Accelerated and Fast Methods for finite-sums.” These sections were added because of the community response to the brevity of the “Stochastic first order methods” section in the original draft. By further elaborating on other stochastic methods in the field of AI, the authors can establish precisely how their work impacts the community, an explanation that was not as apparent in the first draft. This would greatly benefit other researchers by directing them to similar works or seeing the relationship between this experiment and others.

4.1.2 APPLICATION

Although many other fields also require the research to mention other related literature, it seems that there is a heavy emphasis on crafting a detailed and specific ”Related Works” section in the AI community. Most authors have continuously made adjustments and edits to their ”Related Works”, making sure to be as detailed as possible. By mentioning other works and how their research does not conflict with past research, authors can better position their work and make their paper contribution more significant.



6 RELATED WORK

First order oracle methods: The primary method in this family is Gradient Descent (GD) (Cauchy, 1847). As mentioned previously, GD is suboptimal for smooth convex optimization (Nesterov, 2004), and this is addressed using momentum methods such as the Heavy Ball method (Polyak, 1964) (for quadratics), and Nesterov's Accelerated gradient descent (Nesterov, 1983).

Stochastic first order methods: The simplest method employing the SFO is Stochastic Gradient Descent (Robbins & Monro, 1951)(SGD); the effectiveness of SGD has been immense, and its applicability goes well beyond optimizing convex objectives. Accelerating SGD is a tricky proposition given the instability of fast gradient methods in dealing with noise, as evidenced by several negative results which consider both statistical (Proakis, 1974; Polyak, 1987; Roy & Shynk, 1990) and adversarial errors (Devolder et al., 2014). A result of Jain et al. (2017) developed the first provably accelerated SGD method for linear regression inspired by a method of Nesterov (2012). Other schemes such as Ghadimi & Lan (2012; 2013); Dieuleveut et al. (2016), which indicate acceleration

8

Under review as a conference paper at ICLR 2018

is possible with noisy gradients do not hold in the SFO model satisfied by algorithms that are run in practice (see Jain et al. (2017) for more details).

While HB (Polyak, 1964) and NAG (Nesterov, 1983) are known to be effective in case of exact first order oracle, for the SFO, the theoretical performance of HB and NAG is not well understood. Polyak (1987) describes HB to be rather brittle when provided with noisy gradient estimates.

Practical methods for training deep networks: Momentum based methods employed with stochastic gradients (Sutskever et al., 2013) have become standard and very popular in practice. These schemes tend to outperform standard SGD on several important practical problems. As previously mentioned, we attribute this improvement to effect of minibatching rather than improvement by HB or NAG when working with stochastic gradients. Other schemes such as Adagrad (Duchi et al., 2011), RMSProp (Tieleman & Hinton, 2012), Adam (Kingma & Ba, 2014) represent an important and useful class of algorithms. The advantages offered by these methods are orthogonal to the advantages offered by fast gradient methods; it is an important direction to explore augmenting these methods with ASGD.

2004), and this is addressed using momentum methods such as the Heavy Ball method (Polyak, 1964) (for quadratics), and Nesterov's Accelerated gradient descent (Nesterov, 1983).

Stochastic first order methods and noise stability: The simplest method employing the SFO is SGD (Robbins & Monro, 1951); the effectiveness of SGD has been immense, and its applicability goes well beyond optimizing convex objectives. Accelerating SGD is a tricky proposition given the instability of fast gradient methods in dealing with noise, as evidenced by several negative results which consider statistical (Proakis, 1974; Polyak, 1987; Roy & Shynk, 1990), numerical (Paige, 1971; Greenbaum, 1989) and adversarial errors (d'Aspremont, 2008; Devolder et al., 2014). A result of Jain et al. (2017) developed the first provably accelerated SGD method for linear regression which achieved minimax rates, inspired by a method of Nesterov (2012b). Schemes of Ghadimi & Lan (2012; 2013); Dieuleveut et al. (2016), which indicate acceleration is possible with noisy gradients do not hold in the SFO model satisfied by algorithms that are run in practice (see Jain et al. (2017) for more details).

While HB (Polyak, 1964) and NAG (Nesterov, 1983) are known to be effective in case of exact first order oracle, for the SFO, the theoretical performance of HB and NAG is not well understood.

Understanding Stochastic Heavy Ball: Understanding HB's performance with inexact gradients has been considered in efforts spanning several decades, in many communities like controls, optimization and signal processing. Polyak (1987) considered HB with noisy gradients and concluded that the improvements offered by HB with inexact gradients vanish unless strong assumptions on the inexactness was considered; an instance of this is when the variance of inexactness decreased as the iterates approach the minimizer. Proakis (1974); Roy & Shynk (1990); Sharma et al. (1998) suggest that the improved non-asymptotic rates offered by stochastic HB arose at the cost of worse asymptotic behavior. We resolve these unquantified improvements on rates as being just constant factors over SGD, in stark contrast to the gains offered by ASGD. Loizou & Richtárik (2017) state their method as Stochastic HB but require stochastic gradients that nearly behave as exact gradients; indeed, their rates match that of the standard HB method (Polyak, 1964). Such rates are not information theoretically possible (see Jain et al. (2017)), especially with a batch size of 1 or even with constant sized minibatches.

Accelerated and Fast Methods for finite-sums: There have been developments pertaining to faster methods for finite-sums (also known as offline stochastic optimization): amongst these are methods such as SDCA (Shalev-Shwartz & Zhang, 2012), SAG (Roux et al., 2012), SVRG (Johnson & Zhang, 2013), SAGA (Defazio et al., 2014), which offer linear convergence rates for strongly convex finite-sums, improving over SGD's sub-linear rates (Rakhlin et al., 2012). These methods have been improved using accelerated variants (Shalev-Shwartz & Zhang, 2014; Frostig et al., 2015a; Lin et al., 2015; Defazio, 2016; Allen-Zhu, 2016). Note that these methods require storing the entire training set in memory and taking multiple passes over the same for guaranteed progress. Furthermore, these methods require computing a batch gradient or require memory requirements (typically $\Omega(|$ training data points$|)$). For deep learning problems, data augmentation is often deemed necessary for achieving good performance; this implies computing quantities such as batch gradient (or storage necessities) over this augmented dataset is often infeasible. Such requirements are mitigated by the use of simple streaming methods such as SGD, ASGD, HB, NAG. For other technical distinctions between the offline and online stochastic methods refer to Frostig et al. (2015b).

Figure 2: Comparison of related works between original and final draft

4.2 Revision: Additional Experimentation

For this revision, the author has changed some parameters of their experiment or tested new models and has incorporated new data into the paper. These data often appear as a new row in tables, or a new labelling of the graph.

The significance of this revision is reflected in the community. The authors have ultimately re-conducted their experiment during the revision process, most likely because of incomplete or incorrect data that had been pointed out by other reviewers. It could also be a part of the experiment that the authors have simply overlooked, such as using different parameters or a more complex system altogether.

4.2.1 Example

The following example is pulled from the paper ”Synthetic and Natural Noise Both Break Neural Machine Translation” by Belinkov & Bisk (2018). Here, we can see additional experimentation being performed, as shown in Figures 3 and 4. Figure 4 represents material from the original paper and shows how the performance of a neural machine translation (NMT) system degrades when translating German to English as a function of the percent of German words modified (Belinkov & Bisk, 2018). The two types of noise are shown, generated when the translated words are random permutations or when adjacent letters are swapped. In the final paper, shown in Figure 4, the authors have revised the chart. They added a second NMT system named Char2Char as comparison, with yet another type of noise to be analyzed: human error. These are extra experiments that the authors have conducted after their first draft, finding the first set of experimental data insufficient for their research.

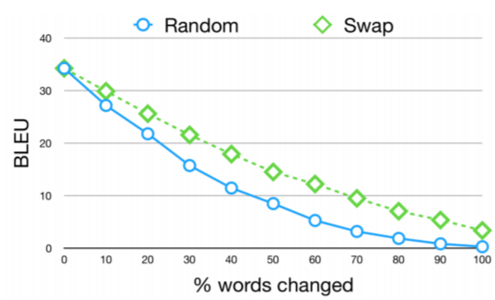


Figure 1: Degradation of Nematus (Sennrich et al., 2017) performance as noise increases.

Figure 3: Graph from first draft

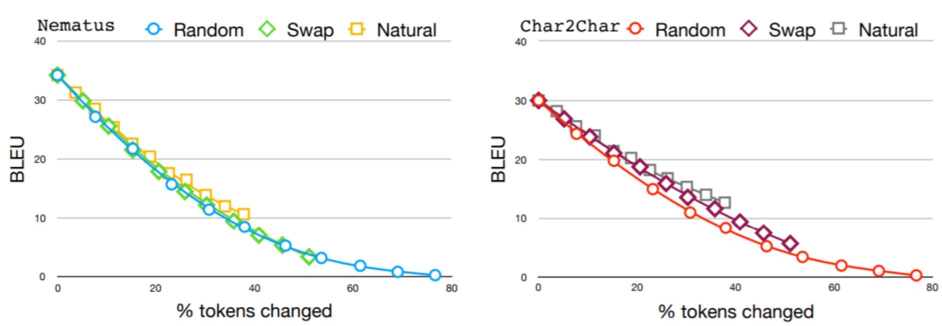


Figure 1: Degradation of Nematus (Sennrich et al., 2017) and char2char (Lee et al., 2017) performance as noise increases.

Figure 4: Graph from final paper

4.2.2 Application

There seems to be a shared sentiment in the AI community for more robust experimentation, as experiments in artificial intelligence can often be retested faster than experiments in other disciplines. In addition, AI models are often open-source, so experiments must often be replicable. Additional experimentation will improve the model and results as a whole.

4.3 Revision: Future Work

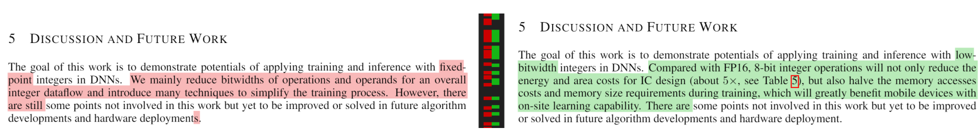
The community also values future work. This could mean mentioning other implementations of the models they propose, or how their model can further improve AI research; proposed future work should be specific and applicable to the field today.

4.3.1 Example

This example is from the paper ”Training and Inference with Integers in Deep Neural Networks” by Wu et al. (2018). An example of a revision in the “Future Works” section is shown in Figure 5. At first, the authors do not mention specific ways their work influences the current field, saying they “...mainly reduce bitwidths of operations and operands for an overall integer dataflow and introduce many techniques to simplify the training process” (Wu et al., 2018). After the revision, however, they include more specifics, indicating how their work will not only “reduce the energy and area costs,” but also how in the long run it will “greatly benefit mobile devices with on-site learning capability” (Wu et al., 2018). To further support this statement, the authors have also included an additional table showing the lower costs of their research, as seen in Figure 6. We can understand the importance of authors explaining how their research specifically fits into their target field and any possible long-term improvements that can be made.

4.3.2 Application

This revision shows that researchers are continuously building on what is published year by year. If the paper does not clearly state how its research can be applied or implemented in current technology, then it would be difficult for the community to appreciate its use. This is especially important in AI research, as artificial intelligence models are constantly being shared and built upon one another.



5 Discussion and Future Work

The goal of this work is to demonstrate potentials of applying training and inference with fixed-point integers in DNNs. We mainly reduce bitwidths of operations and operands for an overall integer dataflow and introduce many techniques to simplify the training process. However, there are still some points not involved in this work but yet to be improved or solved in future algorithm developments and hardware deployments.

5 Discussion and Future Work

The goal of this work is to demonstrate potentials of applying training and inference with low-bitwidth integers in DNNs. Compared with FP16, 8-bit integer operations will not only reduce the energy and area costs for IC design (about 5×, see Table 5), but also halve the memory accesses costs and memory size requirements during training, which will greatly benefit mobile devices with on-site learning capability. There are some points not involved in this work but yet to be improved or solved in future algorithm developments and hardware deployment.

Figure 5: Comparison of original draft to final draft

Table 5: Rough relative costs in 45nm 0.9V from Sze et al. (2017).

Operation	Energy(pJ)		Area(μm^2)	
	MUL	ADD	MUL	ADD
8-bit INT	0.2 pJ	0.03 pJ	282	36
16-bit FP	1.1 pJ	0.40 pJ	1640	1360
32-bit FP	3.7 pJ	0.90 pJ	7700	4184

Figure 6: Table added in final draft

5 Discussion and Future Work

Although this is an exploratory study on the importance of certain sections within an AI research paper, we have come up with a few recommendations based on the findings:

- Read as much surrounding literature as possible. Background research is extremely important; make sure the contribution of your work is clear.

- State and keep track of any parameters for your research. Constantly run experiments and tweak your results until your research questions are answered completely.
- Explicitly state what improvement your research would make in the community. Try to quantify your improvements if you can; for example, showing that your model has higher accuracy or a lower cost.
- Make sure your theorems and mathematical proofs are clear. Label all steps.
- As always, be as clear and concise as possible.

This study did have a few limitations due to the size of the corpus and research group. In the future, we hope to analyze more artificial intelligence research outside of the scope of the ICLR conference. We also hope to go through the open coding process with a research group to minimize individual error and bias. With this, we believe we can make an in-depth guide for beginner researchers in AI, or at least outline some of the conventions that are appearing in this field. We hope that our study can act as a step in the right direction to formalize the academic writing processes in AI.

ACKNOWLEDGMENTS

I would like to thank Professor Caroline Cole for her guidance throughout the research process and making it less terrifying. She has taught us a lot this year about the different conventions used in various disciplines and has really opened our eyes about writing in research as a whole.

REFERENCES

Yonatan Belinkov and Yonatan Bisk. Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=BJ8vJebC-`.

Graham Gibbs. *Analyzing Qualitative Data*. 2007. doi: 10.4135/9781849208574. URL `https://methods.sagepub.com/book/analyzing-qualitative-data`.

Shahedul Huq Khandkar. Open coding. URL `http://pages.cpsc.ucalgary.ca/~saul/wiki/uploads/CPSC681/open-coding.pdf`.

Rahul Kidambi, Praneeth Netrapalli, Prateek Jain, and Sham M. Kakade. On the insufficiency of existing momentum schemes for stochastic optimization. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=rJTutzbA-`.

Johnny Saldaña. The coding manual for qualitative researchers. 2009.

Mary Shaw. Writing good software engineering research papers: Minitutorial. In *Proceedings of the 25th International Conference on Software Engineering*, ICSE '03, pp. 726–736, Washington, DC, USA, 2003. IEEE Computer Society. ISBN 0-7695-1877-X. URL `http://dl.acm.org/citation.cfm?id=776816.776925`.

Tom Silver. Lessons from my first two years of ai research. URL `https://web.mit.edu/tslvr/www/lessons_two_years.html`.

Christopher Theisen, Marcel Dunaiski, Laurie Williams, and Willem Visser. Writing good software engineering research papers: Revisited. pp. 402–402, 05 2017. doi: 10.1109/ICSE-C.2017.51.

Max Welling and Zoubin Ghahramani.

Shuang Wu, Guoqi Li, Feng Chen, and Luping Shi. Training and inference with integers in deep neural networks. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=HJGXzmspb`.

A APPENDIX

	Paper Revisions: Open Coding								
	*P = Present, NP = Not present								
	Paper Name	Final Reviewer Ratings			Related Work	Additional Experimentation	Future Work	Mathematical clarity	Writing clarity
1	On the Convergence of Adam and Beyond	8	8	9	P	NP	NP	P	NP
2	Synthetic and Natural Noise Both Break Neural Machine Translation	8	7	7	P	P	P	NP	NP
3	Multi-Scale Dense Networks for Resource Efficient Image Classification	10	7	8	P	P	NP	NP	NP
4	Training and Inference with Integers in Deep Neural Networks: 8/4, 7/3, 7/4, Design new method	8	7	7	NP	P	P	NP	NP
5	Spherical CNNs	9	7	8	P	P	P	NP	P
6	Ask the Right Questions: Active Question Reformulation with Reinforcement Learning	6	8	7	P	P	P	NP	NP
7	Wasserstein Auto-Encoders	8	8	8	P	P	NP	NP	NP
8	Spectral Normalization for GANs	7	8	7	P	NP	P	NP	NP
9	Progressive Growing of GANs for Improved Quality, Variability, and Stability	8	8	8	P	NP	NP	NP	NP
10	AmbientGAN: Generative models from lossy measurements	8	7	7	NP	NP	NP	NP	P
11	On the insufficiency of existing momentum schemes for Stochastic Optimization	8	7	7	P	NP	NP	P	NP
12	Certifying Some Distributional Robustness with Principled Adversarial Training	9	9	9	P	P	P	P	NP
13	Learning Deep Mean Field Games for Modeling Large Population Behavior	10	8	8	P	NP	NP	P	P
14	Breaking the Softmax Bottleneck: A High-Rank RNN Language Model	7	7	8	P	P	P	P	NP
15	Continuous Adaptation via Meta-Learning in Nonstationary and Competitive Environments	9	7	8	NP	NP	NP	NP	P
	Total Number of Papers with Revision				12	8	7	5	4
	% of Papers with Revision				0.8	0.533	0.467	0.333	0.267

Figure 7: Open Coding data

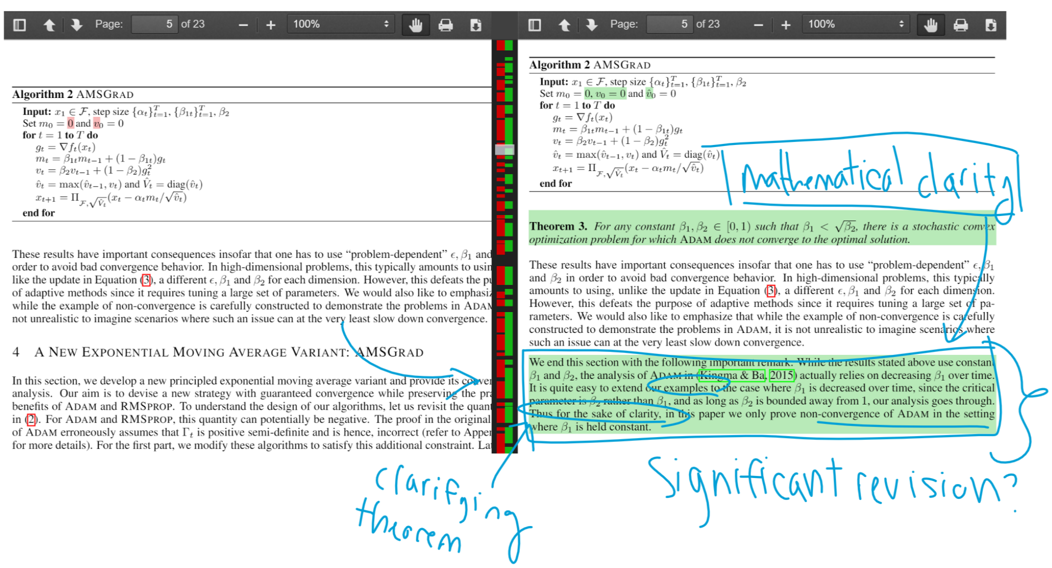


Figure 8: Example of Open Coding Notes for "On the Convergence of Adam and Beyond"