Approximately Correct
Technical and Social Perspectives on Machine Learning

Heuristics for Scientific Writing (a Machine Learning Perspective)

It's January 28th and I should be working on my paper submissions. So should you! But why write when we can **meta-write**? ICML deadlines loom only twelve days away. And KDD follows shortly after. The schedule hardly lets up there, with ACL, COLT, ECML, UAI, and NIPS all approaching before the summer break. Thousands of papers will be submitted to each.

The tremendous surge of interest in machine learning along with ML's democratization due to open source software, YouTube coursework, and the availability of preprint articles are all exciting happenings. But every rose has a thorn. Of the thousands of papers that hit the arXiv in the coming month, many will be unreadable. Poor writing will damn some to rejection while others will fail to reach their potential impact. Even among accepted and influential papers, careless writing will sow confusion and damn some papers to later criticism for sloppy scholarship (you better hope Ali Rahimi and Ben Recht don't win another test of time award!).

But wait, there's hope! Your technical writing doesn't have to stink. Over the course of my academic career, I've formed strong opinions about how to write a paper (as with all opinions, you may disagree). While one-liners can be trite, I learned early in my PhD from Charles Elkan that many important heuristics for scientific paper writing can be summed up in snappy maxims. These days, as I work with younger students, teaching them how to write clear scientific prose, I find myself repeating these one-liners, and occasionally inventing new ones.

The following list consists of easy-to-memorize dictates, each with a short explanation. Some address language, some address positioning, and others address aesthetics. **Most are just heuristics** so take each with a grain of salt, especially when they come into conflict. But if you're going to violate one of them, have a good reason. **This can be a living document, if you have some gems, please leave a comment.**

The Introduction

KEEP YOUR ABSTRACT SHORT

You can't get it all out in the abstract. Don't even try. **Think of the abstract as the 2-minute spotlight talk advertising your paper**. The points should feel like bullets. Here's one (among many) tried-and-true formula:

1. Contextualize the problem in either one sentence or one phrase
2. Identify what's wrong with existing approaches
3. Go big: clearly state your major contribution (can also lead with this).
4. Two or three sentences to sell the details, major quantitative result, etc.

Here's the first brilliant abstract I ever read in a machine learning paper

"Mixtures of Gaussians are among the most fundamental and widely used statistical models. Current techniques for learning such mixtures from data are local search heuristics with weak performance guarantees. We present the first provably correct algorithm for learning a mixture of Gaussians. The algorithm is very simple and returns the true centers of the Gaussians to within the precision specified by the user, with high probability. It runs in time only linear in the dimension of the data and polynomial in the number of Gaussians."

– Sanjoy Dasgupta in "Learning Mixtures of Gaussians"

Note that Sanjoy might have made it even more compact by combining the first two sentences: ***"Current techniques for learning mixtures of Gaussians from data are local search heuristics with weak performance guarantees."***

Pro: Even terser. **Con:** leading with the key phrase "*Mixtures of Gaussians*" catches the eye in a way this version does not.

DON'T TEASE THE READER

Follow-up to above: If you have a great quantitative result, stick the number right in the abstract and the introduction. If your paper yields a single equation that can be operationalized, stick it right in the introduction. People should read on because they are interested, not because you are teasing them by withholding information.

DELETE GENERIC OPENINGS

"The last 10 years have witnessed tremendous growth in data and computers." "Deep learning has had many successes at many things". If the first sentence for your paper can be pre-pended to any paper in all of ML/big data, delete it. First impressions matter. The first sentence is the most precious real estate in your introduction. Don't squander it.

Q BEFORE A

It's difficult to get excited about a solution if you don't believe there is a problem. If your paper is completely abstract and has no bearing on the real world, then it should be evaluated as a work of pure mathematics. It probably won't fare well in that theater. If possible: lead with a compelling real-world example, formalize it as an abstract problem, and then close the loop with experiments that address the motivating case.

FOCUS ON WHAT YOUR METHOD *DOES*, NOT WHAT IT *DOESN'T* DO

Sometimes you need to set up a contrast. But don't get bogged down in describing ideas in the negative, especially your own. When all else is equal (semantically), it's much more readable to ditch the indirection and just say precisely what something is, not what it isn't. This is especially true for your own methods.

Organization

WORDS ARE NOT SENTENCES. SENTENCES ARE NOT PARAGRAPHS. PARAGRAPHS ARE NOT SUBSECTIONS. SECTIONS CONTAIN MORE THAN ONE (OR ZERO) SUBSECTIONS. PAPERS CONTAIN MORE THAN ONE SECTION.

One immediate tell that you are engaging with a lousy writer is that the paper doesn't look right, before you read a single word. Sections, like bullets on slides should be balanced. If you just list the Section titles, they should make sense as belonging to the same scope. Same concept applies all the way down. Occasionally, a paragraph can have 2 sentences. **But the safe heuristic is paragraphs have 3 sentences minimum**.

A READER SHOULD UNDERSTAND YOUR PAPER JUST FROM LOOKING AT THE FIGURES, OR *WITHOUT* LOOKING AT THE FIGURES

A blind reader should understand precisely what you do, even if they miss a couple granular bits of data captured in figures. Any critical observation or technical details must appear in the paper's main text, which can reference figures for visual corroboration.

Similarly, the figures should tell a coherent story. If your reader skips to the figures (**reviewers will**), they should be able to see roughly what's going on and understand the significance of the findings. If it's not obvious whether higher or lower scores on y axis are better, the caption ought to say this.

But don't go overboard. Captions should not be giant paragraphs. A good caption should be between 1 and 3 lines. **Note:** the computer vision community understandably has a very different relationship with figures. Sometime a single figure will take over an entire page and 100s of words of detail absent from the rest of the draft. I do not like this style, but if you're submitting to a conference with such norms, perhaps make your own decision.

QUICKLY ARRIVE AT THE PAPER'S CONTRIBUTION

As a young PhD student, and an outsider to the ML community, I was frustrated that papers were not self-contained. As a result I tried to make each paper fully understandable to an outsider. This won me some readers in the general public but (likely) cost me several early conference rejections.

Longwinded front-matter in conference papers (less applicable to journal) is **bad** for the following reasons: (1) Reviewers read 5-10 papers per conference and 50-100 papers per year in very similar areas. The basics will bore them. (2) If your contribution starts on page 5/8, you have very little excuse for having failed to do anything the reviewer asks for.

There are two issues at stake here: **knowing your audience** and **positioning** intelligently. Most of your abstract (by sentences), your intro (by paragraphs), and your paper (by pages) should articulate what **you** do.

ANTICIPATE THE READER'S QUESTIONS AND ANSWER THEM IN THE PAPER

A good reviewer will try to come up with critical questions to challenge the proposed work. **Is it possible that this method only works because X?** If the answer is "I don't know" and "no" would be damning, your paper might rightly be rejected. **If you can anticipate the question and know the answer, write it. If you do not know the answer, then run an experiment to find out.** I hope this point hits home that doing strong research and writing clearly are tightly linked.

Style

THE SCIENTIFIC "WE"

In scientific writing, narrate with the pronoun "we". This style serves a didactic purpose: "we" refers to **"you" (the reader) and "I/we" (the authors) together**. Sometimes, you may need to express an opinion. These cases should be made clear from context.

AVOID HOSTAGES TO FORTUNE

Any qualified reader, who goes through your entire draft, even if they do not share your opinions, preference for methods, or values in life, should be unable to disagree with any sentence in isolation. "Our method X outperforms Y on **most** datasets." Does it? Most out of what collection of datasets? Could your reviewer choose some dataset repository and find the statement false? Better to say "many" datasets. This is both **better defined** and much **harder to disagree** with.

A SIN OF OMISSION IS BETTER THAN A SIN OF COMMISSION

Related to the above: if you are not 100% sure about a claim, do not make it. It's hard to imagine the reviewers rejecting a paper because you omitted a one-line boast. It's easy to imagine one line inspiring a rejection.

WHEN YOU MUST EXPRESS AN OPINION, IDENTIFY IT AS SUCH

You might ask, the reviewer can disagree with my opinions, does that mean I cannot ever include an opinion in a paper? You can include an opinion, e.g., the great promise of GANs for anomaly detection, but the **factual assertion** should be **that it is your opinion**: *"in our opinion, GANs..."*.

Language

BREAK UP LONG SENTENCES

Young writers often believe, mistakenly, that long sentences reflect language skills. Great scientific writers write mostly in short sentences. **If you find yourself struggling to pack an idea in one sentence, it probably requires more than one.** Technical writing should be as clear as possible. If simplicity is possible, then make the writing simple. The contribution of your paper should be **sophisticated ideas**, not **sophisticated sentence structure**.

JETTISON INTENSIFIERS AND VACUOUS ADVERBS

Examples: **Extremely, Very, Incredibly, Completely, Barely, Essentially, Rather, Quite, Definitely, ...**

Intensifiers are bad for two reasons: (i) they undermine their own purpose: "algorithm X provides a tight approximation" sounds confident, while "algorithm X provides a **very** tight approximation" drips with insecurity, and (ii) they express opinions. Is the algorithm better? Yes. Is it **much** better? That's an opinion, thus a hostage to fortune (see above).

SUBJECTS, VERBS AND MODIFIERS, SHOULD ALL AGREE

One common mistake in writing is to attribute verbs and modifiers to the wrong subjects, e.g. **the algorithm tries to X,** or the **data is biased**. Algorithms don't **try**, just as they don't **think**. If we are speaking to desires or to intentions, then they belong to "we", the modelers, not the algorithm. This sounds like common sense but errors of disagreement plague academic writing across all disciplines. In some fields, such as interpretability and fairness (in ML), where the right definitions are not clear, sloppy writing like this can hold back the entire field.

Corollary: every action should be attributed. **Verbs with no subject can often emerge in passive constructions (where the main verb is "to be")**. For example, "LSTMs are claimed to X, Y, Z". Who is doing the claiming? This information better appear somewhere. One solution would be to append a parenthetical citation. A better solution might be clearly put the claims in the mouths of their authors.

Bibliography

CITE GENEROUSLY

The papers you ought to cite are likely written by the people who will be reviewing your paper. One common lame review will consist of an anonymous reviewer asking why you didn't cite works A, B, and C (all by the same author). If the works are **not**

relevant, then **do not cite**. If they are relevant, you have nothing to lose and much to gain by citing.

Among the good karma you'll earn: (1) **you are less likely to get a shirty review**, and (2) these are often people you want to work with later and **they may notice the citation and read your paper**.

CITE THROUGHOUT

Reviewers are lazy, and do not have photographic memories. If your work builds on others' contributions, do not confine your citations to the *related work* section – that's just to summarize your work's context in the literature. Cite throughout the text whenever you invoke methods that precede your own. This is **especially true for recent work** (last 5-10 years), which may not yet be common knowledge and thus confined to citation-dense paragraph in the *related work* section.

EXHAUST THE REFERENCES LIMIT

This is a pragmatic positioning point and applies to conference publications that limit the number of pages for references (often 1 or 2). If you omit the most related work, reviewers will nail you no matter what. But if you omit some borderline related work and they call you on it, having no room left in the references section is a good excuse. **If you are squatting on a blank bibliography page, don't expect sympathy from reviewers.**

Authors


Zachary C. Lipton

Author: Zachary C. Lipton

Zachary Chase Lipton is an assistant professor at Carnegie Mellon University. He is interested in both core machine learning methodology and applications to healthcare and dialogue systems. He is also a visiting scientist at Amazon AI, and has worked with Amazon

Core Machine Learning, Microsoft Research Redmond, & Microsoft Research Bangalore.
View all posts by Zachary C. Lipton

 Zachary C. Lipton / January 29, 2018 / Opinion / Academic Writing, Conferences, Journals, Machine Learning, Papers, PhD

21 thoughts on “Heuristics for Scientific Writing (a Machine Learning Perspective)”

 **Tom Dietterich**

January 29, 2018 at 12:54 am

I love the point about vacuous intensifiers. I especially hate papers that use “real” or “truly”, as in “real intelligence” or “truly learns abstractions”. The problem with these intensifiers is that they leave undefined “real intelligence” or “true learning”, and that is because we don’t know how to define them. My advice: don’t use these kinds of phrases.

 **Zachary C. Lipton**

January 29, 2018 at 12:58 am

Thanks Tom! I’ve admired your papers for years, so it’s assuring to see that we have some common opinions about how to write one.

 **Daniel Roy**

January 29, 2018 at 3:31 am

I’ve also grown to hate the word “rich” as an adjective. It’s a mind virus that stops people from actually explaining what is interesting/new/non-trivial (or not) about their work.

Zachary C. Lipton

June 22, 2020 at 1:17 pm

Agree! It's nearly always a bald assertion of opinion.

Greg Ver Steeg

January 29, 2018 at 4:53 am

Another vacuous word is "complex". Nobody thinks that what they are doing is simple, so this qualifier rarely specifies what is truly distinctive about your problem.

Also, I strongly recommend Pinker's "Sense of Style" as a fun but useful guide to improving writing. "Show, don't tell" is terrible advice that tells you to show without showing you how. Pinker shows you how to show.

Zachary C. Lipton

January 29, 2018 at 7:43 am

I think "show, don't tell" is sufficiently vague that it's easy to make a straw man of it. But there are some very reasonable ways to interpret that one. A good example: don't say "our model is good at …", say "trained on…, our model achieves an accuracy of …". I think that point, that we shouldn't push the conclusion on the reader but instead show them the evidence that makes the conclusion irrefutable, is sound advice.

Mundher

January 29, 2018 at 9:01 am

I am in process of writing my first PhD paper, I will keep your points in my mind. Thank you.

Bhavesh Neekhra

June 20, 2020 at 8:42 am

Such succinct advice. Thanks for this.

Eleftherios Spyromitros-Xioufis

January 29, 2018 at 9:30 am

Thanks Zack for another great post! I would only add that I personally do not like seeing opinions being expressed in research papers, especially in fields like computer science. There are better places to do that, e.g. conference panels, research blogs, etc.

Zachary C. Lipton

January 29, 2018 at 10:09 pm

Hey Lef. I take a broad view of opinions. “*very* large” expresses an opinion. I wouldn’t even allow that in the body of a technical paper. In the introduction and discussion, however, some kinds of opinions are inevitable. “Given their success on X, we believe DNNs will prove important for Y”. This expresses an opinion. And yet it is a bit scientifically relevant, it speaks to either (1) why you did the current work, or (2) what you think are good ideas to try in future work. In these cases, the best thing to do is to identify them clearly as an opinions.

David Ernst

January 29, 2018 at 9:15 pm

What’s your opinion regarding proactive mention of limitations? I have read many papers that conveniently just don’t mention the proposed algorithm’s limitation to certain types of data, or transductive nature, or prohibitive computational complexity…

That’s of course common practice is politics and business, but should it be the same in research?

Martin Becker

January 30, 2018 at 10:59 am

Thanks for the nice list!

About “KEEP YOUR ABSTRACT SHORT”: I usually like to add one more sentence about the perceived/possible impact of the work if it can be expressed concisely (and without general statements). While it lengthens the abstract a bit, it can significantly increase the reception of and interest in your work because, e.g., practitioners can glean the value for their applications.

Zachary C. Lipton

January 31, 2018 at 2:11 am

Hi Martin. Thanks for the reply. So the template I gave there is just one example. A sentence about impact is great! My main point is to keep it to the elevator pitch, not to start rambling or describing insignificant details.

Zachary C. Lipton

January 31, 2018 at 6:18 am

This hastily-prepared piece of meta-writing has now been (meta?, no just kidding)-edited. Changes only concern typos. Given some of the ad hominem rancor on Twitter throughout ML circles, I’m reminded that I forgot to point out that the only agents of interest to science papers should be other papers. “We the authors” – means “We the authors of *this paper*” not we, real, people who happened to author this paper. This is extremely important, it’s what gives us the power to decide in the future to disagree with our present arguments and to find flaws in our present methods.

Kunal Relia

January 31, 2018 at 5:21 pm

Thank you very much for listing out the heuristics, which are very useful for all the technical writers, especially for beginners like me. It is my first year of Ph.D., and these points make a perfect starting point for technical writing that leads me in the correct direction.

Miguel P Xochicale

May 11, 2018 at 7:09 pm

Many thanks for the advice. I love this one: “Great scientific writes mostly in short sentences.” As a non-native English speaker I had the misconception that making long sentences might make the paper stronger but it is just the opposite.

Zachary C. Lipton ▪

May 16, 2018 at 2:26 am

Of course man. Good luck with the paper writing!

Babak Hosseini

August 23, 2018 at 8:18 am

Thank you so much for this great post, i learned a lot of useful hints from it!

I often get the following problem when writing machine learning papers:
The reviewers usually ask me to compare my method to the methods from other domains.
For example, i work on sparse coding and dictionary learning, and they ask “why you haven’t compared it to deep learning classification?”, but, we know that dictionary-based encoding is not just for the sake of classification!

Or they ask “why you have not tried the method on big data like Imagenet?”, while i have not aimed to design a large-scale method!

I can add paragraphs in my paper talking about what other data or other methods to which my methods shouldn’t be applied/compared, but i find it so inconvenient and

odd!

Mirza Ahsan Ullah

June 22, 2020 at 8:51 am

I found this post very helping and was refereed by PhD supervisor to read. This post is as helping as your paper 'The Myths of Model interpretability', which helped me alot to understand the basics of this field.

Ariane Sasso

June 9, 2021 at 8:48 am

Wow, I loved this piece!
I think this article can help anyone improve their scientific writing (including me :)).

I have to forward this to my students as well. Sometimes they don't believe me when I say that short sentences are good. Also, they love the passive voice so much it is hard to convince them otherwise. I think non-native speakers believe this is fancy writing (I know I used to haha).

Thanks for the great text!

Zachary C. Lipton

July 1, 2021 at 9:44 pm

Thanks Ariane. Yes, everything in moderation. Sometimes the subject behind a verb is not really important and the passive voice is appropriate. But too often active statements are made indirect and passive to no end, even when the subject is already taking up real estate inside the sentence.