# View Reviews

**Paper ID**
1242

**Paper Title**
Perturbation Theory of Stochastic Gradient Descent at Small Learning Rates

**Reviewer #1**

## Questions

**1. Rating (scale of 1-7)**
1: Clear reject

**2. Review**
I have no reason to believe this paper is technically correct or meaningful (there are various issues with large T vs small eta , that depending on how you interpret the theorems are false or vacuous). The theorems are not even stated in a way that is verifiable or precise enough to mathematically check.

Given the loose use of quanitifiers, even if the theorems were correct, I do not know how to interpret. For example, take proposition 1, how can you have an o(\eta^2) term in the large -T limit but also say the o(\eta^2) do not depend uniformly in T. This term could be poly(T) \eta^2 or exp(T)\eta^2, then the statement is meaningless. The main theorems (e.g. theorem 2) also has this issue.

A meaningful result would require uniform control over o(\eta^2) terms in T. Or at least a result that says "For \eta=\eta_0, and T> blah , this happens"

**Reviewer #2**

## Questions

**1. Rating (scale of 1-7)**
6: Strong accept

**2. Review**

Summary of results:
This paper extends existing analysis that does Taylor expansion on the step size in the limit of small step sizes for SGD. The importance of the extension lies in the fact that the authors do higher order Taylor expansions which reveal interesting dependencies of the test error on higher order moments of the gradient distribution. Higher order Taylor expansions are somewhat of a logistic nightmare: the expansion up to degree d is generally a d degree polynomial and the number of terms explodes. The authors import their toolbox from physics: they use diagrams to formalize, represent, manage and count the whole zoo of possible monomials in those polynomial representations. As those monomials are generally higher order tensors, the authors also import Einstein notation to manage all the necessary algebraic operations. The authors establish a first main theorem which establishes the true risk of a

learning problem in a combinatorial form: it can be expressed as the sum of the values (uvalues) of all valid diagrams given a dataset of size N, E epochs and a batch size B. Each diagram encodes tensor contractions and equivalences of factors participating in the same expectation (because they correspond to the same training example). This first result allows them to get higher-degree Taylor expansions for small step size and fixed number of steps T. Then the technique of “resummation” over equivalence classes of diagrams (with varying lengths of chains of Hessian products) allows the authors to achieve a stronger result which yields convergence even in some cases when the number of steps T goes to infinity. Finally, the authors use their theory to extract a number of interesting hypotheses about the interaction of generalization with noise and curvature and they go on to validate said hypotheses on carefully designed experiments.

REVIEW

Reviewing this paper has been a journey. It was one of the most frustratingly difficult papers I had to review (even though I have good expertise in optimization and ML). The main reason is the alien (for typical CS standards) notation and methodology. Initially, I was inclined to suggest rejection because not many people in CS will be able to read and understand this paper. Eventually, the notation grew on me (with considerable effort on my part) and I really appreciated the depth of the results and the abundance of wonderful insights regarding the interaction of noise and curvature (even if it just for the small step size regime). I also really appreciated the very pedagogical appendices, with many pages of slow introduction to diagrams, their properties, examples, and instructions on how to evaluate them. I am recommending acceptance and summarizing some strong points and weaknesses below.

Strong points:
1. Going beyond a Gaussian analysis for the distribution of gradients in SGD. In particular, dealing with non-isotropy and skewness seems to offer strict benefits
2. Some extensive explanation with proofs and examples in the appendix. For example, Section A.4 in the appendix is an excellent extended discussion on uvalues of diagrams.
3. I really appreciated all the great insights, including the potential for the presence of curl (!) in SGD dynamics due to the interaction of gradient covariance directions and hessian eigendirections.

Issues/weaknesses:
1. The paper is a challenge to read for a reader not acquainted with Einstein notation. You devote a bit of space in Section 1.1 to quickly explain but it’s not a sufficient intro for CS folks. I would recommend that you add a reference to some reasonable tutorial/intro to the uses of this notation with examples (or you could write one as another appendix to your paper). The reference to (Bonnabel, 2013) didn’t help me really understand what was going on. I had to read from a other sources to gather the pre-requisite understanding; you can make it easier for the CS reader.
2. There is a relevant piece of work that you should take into consideration and probably cite and discuss [1], see below. In this paper, the authors are motivated by a similar goal and arrive at somewhat relevant (though distinct results). The authors do backward error analysis. They start by explaining that when we view GD as an Euler discretization of its limiting ODE (“gradient flow”) we get an error of O(h^2). Then, they go modifying the ODE in a way that discretization still yields gradient descent, but now with an error of O(h^3). The product of this backward error analysis is a modified objective function, which now includes an explicit extra term, which can be thought of as the implicit regularization of gradient descent, now taking physical form explicitly. As far as I know the authors do not really study the role of stochasticity like you do. Still, there is a common theme of doing a higher order Taylor expansion wrt to the step size in order to understand implicit regularization, so it’s worth the discussion.

REFERENCES:

[1] Barrett, D. G., & Dherin, B. (2020). Implicit gradient regularization. arXiv preprint arXiv:2009.11162. Chicago

**Reviewer #3**

---

## Questions

**1. Rating (scale of 1-7)**
2: Below acceptance threshold

**2. Review**
This paper studies the dynamics of stochastic gradient descent (SGD) and proposes a new diagram based approach to derive bounds on the expected test loss for a model trained via SGD. The key contributions are i) proposal of a new diagram based approach to generalization which overcomes issues of summing over exponential number of terms, and ii) using the results to contrast SGD with gradient descent and stochastic differential equation based dynamics.

The paper addresses a well-motivated problem of understanding generalization error of models learnt via SGD under the small step-size regime. While the proposed insights are interesting, there are a few points worth highlighting

- Clarity of exposition: The paper in its current form is extremely hard to follow and further, even verify for correctness. The paper uses quite a few notations and assumptions, without appropriately mentioning it. For instance, the statement of Proposition 1 does not state any assumptions neither defines what a degenerate minimum is, but its proof relies on “smoothness assumption os Sec. B.1”. the tensors C and S defined in Figure 2 seem to be vector valued (as G is a vector) but the expectation is taken considering the arguments to be scalar. The paper uses non-standard notation like “we implicitly sum repeated Greek indices”, which makes the paper overall very hard to follow and verify.
- One of the main contributions, the diagram notation, is quite unclear to follow. In particular, it is not clear a priori how the introduction of this notation simplifies the understanding of the existing proofs or how it makes them more manageable. While I appreciate that the authors presented a tutorial on how to use these diagrams, it is way too dense in the writeup and the main idea is not well presented in the paper.
- There is little discussion on the proof techniques and ideas used to prove the main results presented in the paper, which makes it difficult to verify the desired claims.
- Corollary 5 seems to be a bit counterintuitive to current understanding of SGD dynamics. It seems to claim that both “flat” and “sharp” minima generalize well. In particular, there seems to be evidence that sharp minima generalize poorly as compared to flatter ones [1] but corollary 5 seems to suggest something quite different.

[1] Keskar NS, Mudigere D, Nocedal J, Smelyanskiy M, Tang PT. On large-batch training for deep learning: Generalization gap and sharp minima. arXiv preprint arXiv:1609.04836. 2016.